HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 10, 2025
https://doi.org/10.30658/hmc.10.5

Transformations and Revelations: The Communicative Constitution of Trustworthiness and Trust Through AI Development Practices

Tessa Bruijne[1], **Anouk Mols**[2] **and Jason Pridmore**[3]

1 Department of Digital Governance & Regulation, TNO Vector, Netherlands
2 Research Group Curriculum Development in Primary and Secondary Education, University of Applied Sciences Utrecht, Netherlands
3 Department of Media & Communication, Erasmus University Rotterdam, Netherlands

Abstract

This study explores AI development practices to understand how trustworthiness is built into AI systems, and how this generates trust in AI. Through a multi-sited ethnography-based methodology, we analyze observations, interviews, and documentation from AI developers working on trustworthy AI. Our analysis shows two key practices: transformation and revelation. Through transformational AI development practices trustworthiness is (re)constituted, though more or lesser degrees. Through revelation practices, AI developers communicatively engage with others to generate trust. This focus on developers adds to the user-centric perspective and shows the role nontechnical development practices have in shaping trust and trustworthy AI before it is implemented. Policy guidelines lack clarity on nontechnical aspects, so we argue that further attention on communication can benefit AI practice and policy.

Keywords: artificial intelligence, trustworthy, sociomateriality, ethnography, communicative constitution of organising

Funding Institution: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

CONTACT Tessa Bruijne • tessa.bruijne@tno.nl • Department of Digital Governance & Regulation • TNO Vector • Netherlands

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

Introduction

The growing popularity of artificial intelligence (AI) has proliferated despite recent cases of AI-inflicted harm. While blame is often put on technological aspects, harms are not the result of either technological deficiency or human misuse. Rather, they are inherently sociotechnical. Recent examples of AI harms, understood as "the adverse lived experiences resulting from a system's deployment and operation in the world" (Shelby et al., 2023, p. 723), include Google Gemini producing historically inaccurate images and controversial text-based responses (Pequeño IV, 2024), and Australia's automated decision-making (ADM) system Robodebt reducing payments to welfare recipients, causing them increased debt, depression, shame, and even some to commit suicide (van Krieken, 2024). The output of generative AI (genAI) systems is often accepted as accurate and trustworthy by users (Kleinman, 2024), particularly when this is repeated often enough. While (incorrect) outputs of genAI concern a technical limitation, the interaction with human heuristics makes harms become more readily apparent. In ADM systems, harms emerge from a combination of limited technology review processes and lack of direct human oversight. AI experts increasingly highlight the need for better analyses for understanding and anticipating (harmful) consequences of AI systems (Shelby et al., 2023). Despite potential for harms, AI implementation continues while organizations behind AI systems risk losing public trust (Oomen et al., 2024).

With notable events of AI harm, trust in AI and trustworthy AI have garnered increased attention in research and in practice. The most recent comprehensive contribution is the report on Ethics Guidelines for Trustworthy AI by the High Level Expert Group on AI (AI HLEG) working on behalf of the European Commission (EC) (AI HLEG, 2019; Stix, 2022). Trustworthy AI is intended to avoid AI harms and even improve people's lives but the literature shows that trustworthiness is no guarantee for trust (Durán & Pozzi, 2025; Reinhardt, 2023). Therefore, this study aims to understand how trustworthiness is built into AI systems and how trust is facilitated, before AI systems can cause harm to society or individuals. Thus, the central research question of this study is: *How does trust and trustworthiness materialize through communication in/through AI development practices?*

We conducted an ethnographic-based study to understand how trustworthiness emerges from and materializes in AI development practices. Our study recognizes that all AI developments are materially (re)constituted and organized through communication practices. For our analysis, we borrow from the communicative constitution of organising (CCO) perspective. Specifically, we make use of Cooren's (2020) conceptualization of materiality as a "necessary property of existence" (p. 2) in communication. This study contributes to debates within Human-Machine Communication (HMC) research wherein human- or machine-centric focus are critical concerns (Guzman & Lewis, 2020). Specifically, this study shifts the focus from interactions between users and (implemented) systems to interactions between AI developers and AI systems. Our results show that trustworthiness materializes as systems are being developed, and that trust is facilitated through revelations.

Theoretical Background

The Shift Toward Trustworthy AI

A focus on social components like trustworthiness is relatively recent for AI. The history of AI development can be roughly divided into three waves (Xu, 2019). The first two waves provided the first AI applications and improvements of AI as a technology, the latter was enabled through large datasets (Plasek, 2016). From here on, challenges emerged around data collection, processing, and storage, and issues arose concerning bias, equality, trust, transparency, and sustainability. The third (present) wave of AI development displays an increased focus on responding to societal concerns rather than (model) optimization, requiring increased awareness of human needs and ethical design (Georgieva et al., 2022; Xu, 2019). Such concerns align with debates about how AI can be understood in communication research, particularly with a focus on the agency of AI (Etzrodt et al., 2024; Guzman & Lewis, 2020). The more we understand AI as having agency, the more we need to ensure the development of beneficial forms of AI. Unfortunately, many conceptualizations of beneficial AI have emerged and disappeared without creating adequate standardization for how this should occur in practice, limiting positive potentials for AI development (Blanchard et al., 2024; Stix, 2022).

The current framing of "beneficial AI" is present in discussions surrounding the EC's introduction of Trustworthy AI (AI HLEG, 2019; Stix, 2022). The EC outlines guidelines for lawful, ethical, and robust AI along seven core requirements: (1) Human agency and oversight, (2) Technical robustness and safety, (3) Privacy and data governance, (4) Transparency, (5) Diversity, nondiscrimination, and fairness, (6) Societal and environmental well-being, and (7) Accountability (AI HLEG, 2019). These requirements are deemed necessary for creating trustworthy AI and, by extension, trust in AI (Reinhardt, 2023). Stix (2022) argues that the concept Trustworthy AI provides clearer boundaries for practice and less interpretative freedom than previous iterations. Yet, the guidelines still seem to offer little practical direction on how to make AI trustworthy, a common criticism of ethical guidelines (Blanchard et al., 2024; Georgieva et al., 2022). As such, the question remains how AI development practices result in trustworthy systems.

Trustworthiness of and Trust in AI

The HLEG guidelines suggest that their directions will result in trustworthy AI systems. Yet, the guidelines are contradictory in their conceptualization of trust and trustworthiness (Reinhardt, 2023). Trustworthiness cannot be seen as an automatic guarantee for being trusted. Instead, trust is an attitude while trustworthiness is a set of properties that would give reasons for being worthy of trust (Durán & Pozzi, 2025; Reinhardt, 2023). The complexity of AI beyond human intuition and comprehension requires us to relinquish some hold over processes where AI is implemented. Durán and Pozzi (2025) argue that "trust becomes relevant precisely in situations lacking full control" (p. 16). This suggests that other factors come into play that make an AI system from being trustworthy to being trusted. Thus, we argue for the necessity of better understanding of how AI systems are built with trustworthiness values in mind because this concerns communicative actions (Orlikowski, 2007).

How trustworthy AI is developed in practice remains elusive, besides noting the need for ethical practices, for example, around data (Plasek, 2016). An emphasis on trustworthiness and generating trust requires the reorganization of AI development practices that goes beyond the technical performance of models. Managing requirements for trustworthiness requires a weighing of different experiences, demands, technological potentials and limitations, ethical considerations, accountability, and privacy-preserving processes. Key to how AI development practices work toward implementing requirements for trustworthiness and aim toward trust are communication processes.

Communication and the Communicative Constitution of Trustworthiness

AI development is digital work completed at various points and performed in the context of organizations and society at large (Orlikowski & Scott, 2016). Developers bring together technical components but also interact with each other, their communities, and the AI systems. Thus, trustworthy AI development is a practice reliant on sociomateriality, as "there is no social that is not also material, and no material that is not also social" (Orlikowski, 2007, p. 1437). From an HMC perspective, AI emerges through and from social practices that are materially (re)constituted and organized through interactive development practices that are inherently communicative. Practices are performances of routine behaviors intertwined with competence and meaning communicated by human and machine agents (Reckwitz, 2002; Shove et al., 2012). In AI development, practices encompass variations of combinations of materials (digital tools, software, and human bodies), competencies (like programming and data science), and meanings (e.g., motivations to solve challenges with technology). Increasingly, the move from performance to social qualities like trustworthiness shifts the meaning of AI development practices. This affects AI development performances, how materials come into play, required competences, and how this is communicated.

To create a deeper understanding of AI development practices and how they result in the emergence of trustworthiness, we focus on communication processes. We argue that communication is foundational for the organization of trustworthy developments. Drawing on Communicative Constitution of Organisations (CCO), communication brings into being organizations and organizing practices (Schoeneborn et al., 2019). Trustworthiness is not an inherent value of a system, technology, or organization, but instead, emerges from and structures communication. Through communication practices in AI development, trustworthiness materializes: it emerges and makes itself present to others (Cooren, 2020), affecting developers in turn. Materialization is an ongoing process, constantly emerging from, changing, and structuring practices. Additionally, CCO presupposes a relational view on communication (Cooren, 2020; Schoeneborn et al., 2019). This suggests that variations in relations affect how and to what degree trustworthiness materializes or how trustworthiness passes "from one matter to another" (Cooren, 2020, p. 2). Materialization occurs in different communicative acts of AI developers (Cooren, 2020), including thoughts, ideas, or organizational communication behaviors (such as speeches as less tangible forms and written documents, policy, and code as more tangible forms).

Method

Study Design

To study AI development practices, we followed ethnographic research principles of multi-sited ethnography (Hallett & Barber, 2014; Marcus, 1995). The study incorporates both physical and digital sites of an EU-funded project about trustworthy AI development. Based on relevant literature, we established topics of interest (user requirements, metrics, trustworthy AI, explainability, transparency, accountability, privacy, resilience, fairness, legislation and regulation, organizational context and collaboration). These topics guided engagement with physical sites (during organization visits and in-person project meetings) and digital sites (meetings via Zoom, Teams or Google Meet, and digital documentation). We employed traditional ethnographic methods, such as observations, interviews, and document analysis.

Participants

Our participants are from different organizations, such as public research organizations, small and mid-size enterprises (SMEs), large enterprises, and (technical) universities based in the EU, often active in the specific context of research and development (R&D). These organizations worked on various AI technologies, such as privacy in networks and infrastructure, federated learning, cybersecurity, edge technologies, and Internet of Things. Participants have a background in computer science or similar and often combine research with AI development. We conducted interviews with 22 participants (see Table 1), who were recruited through an EU-funded project in which we participated as a partner. This project on the development of trustworthy AI spanned from 2021 to 2024, we conducted our fieldwork in 2021 and 2022. Additionally, two of this project's external advisors participated to reflect on how the work is embedded in the broader international landscape of AI development.

Procedure

We observed meetings in various setups, which were recorded and transcribed. When recordings were not possible, we made detailed notes for the analysis. During the process, we often discussed and compared researcher notes for clarity and completeness. For the document analysis, we studied meeting minutes and deliverable reports. We gathered final versions of reports (now publicly available). An overview of included documents is listed in Table 2.

The interviews were done individually, in pairs, or small groups, either digitally or in person. They lasted between 10 minutes (during breaks at project meetings) to 70 minutes, and were recorded and transcribed. The interview topics are included in Table 3.

TABLE 1 Participant Overview With Pseudonyms

#	Pseudonym	Position	Organization
1	Rick	Mid-level AI developer/researcher	Public research organization
2	Jacob	Senior-level researcher	Public research organization
3	James	Mid-level AI developer	Small to medium enterprise (SME)
4	Robert	Mid-level AI developer	SME
5	Ferdinand	Senior-level AI developer	SME
6	Lucy	Senior-level researcher	SME
7	Anton	Senior-level AI developer	SME
8	Theo	Senior-level AI developer/researcher	Large enterprise
9	Gary	Mid-level AI developer/researcher	Large enterprise
10	Andre	Mid-level researcher	Technical university
11	Joel	Senior-level researcher	Technical university
12	Victor	Senior-level AI developer/researcher	SME
13	Damien	Mid-level researcher	Technical university
14	Jeffrey	Senior-level researcher	Technical university
15	Alex	Senior-level AI developer	SME
16	Harold	Senior-level AI developer	Large enterprise
17	William	Mid-level AI developer/researcher	Technical university
18	Lars	Software developer AI course	SME
19	Nina	Co-lead AI course	SME
20	Sonia	Co-lead AI course	SME
21	Edward	Project advisor	University
22	Keith	Project advisor	University

TABLE 2 Overview of Project Documents Used as Data

#	Document	Type and Main Focus
1	Doc1	Project deliverable about requirement analysis
2	Doc2	Project deliverable about security threats
3	Doc3	Project deliverable about accountability and resilience features
4	Doc4	Project deliverable about accountability, resilience, and privacy metrics
5	Doc5	Project deliverable about bias and data quality
6	Doc6	Presentation slides about project output
7	Doc7–Doc24	Small-group project meeting minutes
8	Doc25–30	Project-wide meeting minutes

TABLE 3 Overview of Themes and Topics for the Interviews

Theme or Topic	Question Examples
Context of the use case and role in the project	What is the context of the use case? Who are the intended clients, users, audiences?
User requirements	How do user requirements play a role? How are user requirements determined?
Metrics	How are metrics defined? How do metrics play a role?
Trustworthy AI and related concepts (e.g., explainability, transparency, accountability, privacy, resilience, fairness)	How does the chosen concept play a role in your project work? What are trade-offs in relation to these concepts? How can users determine trustworthiness of AI?
Legislation and regulation	Which legislation or regulations apply to your work? (GDPR, AI Act, other)
Project context and collaboration	How do you: Share knowledge? Facilitate collaboration? Identify and manage varied approaches to work? Manage individual, team, and project goals?

Ethics and Data Availability

This study received approval from the Ethics Review Board (ETH2122-0675). Participants provided informed consent. All identifying information was removed from transcripts or replaced with pseudonyms. The data are not publicly available due to privacy considerations. Inquiries and reasonable requests can be made to the corresponding author.

Data Analysis

We employed an automated transcription service compliant with the General Data Protection Regulation (GDPR). The transcripts were pseudonymized and corrected where necessary. We coded the pseudonymized transcripts, documents, and observer notes in the web version of ATLAS.ti (http://web.atlasti.com), allowing for online collaboration without risking duplicates or data loss. Our analysis followed principles of reflexive thematic analysis (Braun & Clarke, 2006, 2021), in which we paid particular attention to the experiences, practices, opinions, interactions, and processes visible in the data, rather than applying a lens of existing theories (Braun & Clarke, 2006). Key to this research was understanding the role of communication and communicative practices in the production of trustworthy AI. Through an iterative and reflexive process of open coding and theme clustering, we remained open to nuances, contradictions, and contrasts existing in the dataset.

During the analysis of transcripts, documents, and observer notes, we labeled parts of the texts with open codes. For example, Jacob and Rick discussed specific domain expertise needed to analyze a ML model. We labeled their discussion with the open code "domain expertise." We added descriptions to all the codes and double-coded part of the data to ensure validity. Afterward, we clustered open codes into subthemes (see Table 4) whereby we regularly cross-checked subthemes with the research data to ensure alignment. Such reflexivity is essential (Braun & Clarke, 2021), as methodological flexibility can detract from a consistent and systematic approach (Braun & Clarke, 2006). Finally, we clustered subthemes into two overarching themes, namely Transformation and Revelation.

Results and Discussion

The findings show how trustworthiness of AI materializes in and through development practices, highlighting that AI is not something that exists out there but is materially (re)constituted and organized in communication processes. Trustworthy AI development practices are predicated on communication activities that can be categorized under two themes: *Transformation* and *Revelation*. Throughout the discussion, we employ Cooren's (2020) materialization as transferring a quality from one thing to another.

Activities clustered under Transformation show how trustworthiness materializes through various development practices, transferring trustworthiness from one thing (e.g., data, requirements) to another (the AI system), though to different degrees. The activities in this cluster correspond with the development of the technology and demonstrate how developers use communication to create and enable trustworthy AI systems through sociomaterial practices (Cooren, 2020; Orlikowski, 2007). The activities under Revelation demonstrate how AI developers make trustworthiness visible to others with the aim to create or facilitate trust. Revelation takes place in communication with others, outside of the development relationship, and provides AI developers with opportunities to reflect on development processes. Table 4 outlines the two overarching themes and the subthemes that emerged in this study.

TABLE 4 Overview of Themes and Subthemes That Emerged From the Analysis

Transformations for Trustworthiness	Revelations for Trust
Incorporating **requirements and metrics** for trustworthiness into AI systems	Being **transparent** about choices made as part of AI development practices
Apply ethical **data practices** for developing AI systems	Implementing **explainability** measures to show inner workings of AI systems
Remaking **pre-existing tools and legacy systems** into trustworthy AI systems	Being responsive in/through **communication and dissemination** about AI projects
Bringing in new and share **knowledge** for integration into AI systems	Making the **reputations** of (partner) organizations visible
Anticipating new and implementing existing **legislation** into AI systems	

Transformation

The first theme, Transformation, covers practices that take place during the design, development, and testing of AI systems. Each transformation facilitates but also complicates the materialization and passing on of trustworthiness. While the challenges presented here are not fully unique to trustworthy AI, their impact is compounded by the additional requirements for development practices. We discuss the five key activities presented in Table 4 in order.

Requirements and Metrics

Before building AI systems, developers determine a set of requirements and metrics that need to be achieved in the final product. As listed in Doc1, participants created an initial set of 81 requirements for trustworthy AI relating to topics such as privacy, security, and transparency—not dissimilar to the AI HLEG guidelines (2019). The set is based on domain expertise of participants, their work contexts, and literature research. Damien, a mid-level researcher at a technical university, acknowledges the large number of items, but also states:

> "We just want a very clear set of principles and guidelines that anyone can follow. And just because your product doesn't meet all the principles doesn't mean it's not accountable or explainable."

While highlighting the complexity of managing requirements, Damien suggests it is possible to make good AI systems, deserving of trust, without meeting all trustworthiness requirements. Based on this research, requirements can be seen as a transformation of the aims for an AI system that are transformed into principles others might use to analyze the AI system or use to build new systems. Iterations of transformation can introduce or perpetuate errors and affect the passing of trustworthiness from one stage to another. If transformations of requirements are incomplete, ineffective, or contradictory, this may impact trustworthiness.

Data Practices

A second key component of AI development relates to practices concerning data, since ethical data practices are at the core of trustworthiness (Plasek, 2016). Data is both at the heart of and supportive of transformations. Depending on what kind of data is relevant, data can represent people, network traffic, malicious activity, and many more. However, data is considered more than what it might represent. From Doc1: "The biggest challenge is to collect a considerable enough amount of data that is sufficiently representative for the problem to solve." This statement shows that data supports the transformation of the AI system as a solution to a particular problem. However, when discussing data, training practices, and AI development, our participants tend to bring up complications related to characteristics of data and trade-offs. The notion of trade-offs (e.g., privacy versus security or explainability versus performance) was regularly addressed as something that all participants have

faced previously. Although seen as self-evident and helpful, the use of trade-offs calls into question how problems in developing trustworthy AI might be approached. A trade-off might suggest a perception of contradictions between expectations, where such a view might not be constructive or necessary. In turn, this might affect the transfer of trustworthiness requirements, or the degree to which such requirements might materialize.

Legacy Systems and Pre-Existing Tools

Another area of transformations concerns the integration of new AI systems into established environments or using pre-existing tools for new developments. Transforming what is currently present into a functional and trustworthy AI system is complex according to project deliverables written by the participants (Doc1–Doc5). From Doc1:

> "The development process of AI models differs from those of traditional software systems. [. . .] Given the heterogeneous nature of each task, it is a substantial challenge to orchestrate the execution of these in a fluent, interoperable, and coherent way. This raises concerns from a system architecture perspective since the architecture must support such pipeline stage integration."

Alongside this, we observed complications with developing new tools and implementing them in current environments or using pre-existing tools, specifically in terms of suitability of the current environments. The transformation of older non-AI legacy systems into trustworthy AI systems is not self-evident, primarily because of limitations regarding interoperability. It is difficult to pass trustworthiness on to existing system architectures or to have it emerge to a larger degree in these new configurations.

Knowledge Sharing and Collaboration

Our findings further indicate that collaborations with others require developers to work on different transformations but also facilitate transformations. One of the first observations from the project collaborations shows that the project partners invested time and effort in developing a similar vocabulary through collaborating on project documentation (from Doc1–5). Due to the diverse expertise and experience in collaborations, the participants had different understandings of concepts such as transparency, reliability, resilience, or put the emphasis differently. Creating a shared understanding of these terms proved a good starting point for setting up effective and efficient collaboration, enabling the transfer of trustworthiness from their individual understandings to the project work, and to a stronger degree.

The participants also discussed how collaborations provided them with the opportunity to expand their knowledge and skills in more areas of trustworthiness than planned previously. One of the participants spoke directly to their experiences of collaborating and becoming acquainted with other dimensions than their original focus. Theo states:

> "Our use case was very focused on privacy and now we're thinking of including some sustainability aspects."

The collaborative project combined with an open attitude toward grasping opportunities enabled the degree to which requirements of trustworthy AI could be incorporated. While it was not always feasible to incorporate all requirements, Theo's experience is exemplar of other participants. Through our observations, confirmed by our researcher notes, we witnessed how partners discussed at multiple times how they collaborate with specific partners to incorporate the others' experience and knowledge to further develop their AI system for components of trustworthy AI. Thus, a transformation in individual knowledge and creating a commonly understood language facilitates the materialization of trustworthiness and increases the degree to which trustworthiness can materialize.

Legislation and Policy

The participants also spoke about the role of legal frameworks in the development of trustworthy AI. Part of development practice is to transform legislature into actionable content. This emerges in part through determining the requirements for AI systems, but it is also more complex than that. While legislation might provide structures for AI systems, like the guidelines may provide directions for achieving trustworthiness, our participants seem to perceive limited impact on their work, at least in early stages of R&D. Participants vocalized that the distance to the market of their work relieves some of the need for attention to transforming regulations. James, mid-level AI developer at an SME, states:

> "We don't have to actually be an expert on this. We don't have a dedicated person who is in charge of this kind of thing."

The distance between AI development and research and the market-readiness reduces the applicability or at least the pressure on implementing certain legal prescriptions, thus limiting the passing on of trustworthiness from the regulations into AI systems.

These perspectives from the interviews seem to suggest that legal frameworks and policies have limited impact on research and development of AI. However, the document analysis and observations suggested otherwise. For example, the GDPR is discussed regularly and included in the set of requirements. And the AI Act, while still under discussion at the time of data collection, is considered in Doc1, and regularly observed in project discussions. When regulation is implemented and actionable, development practices are transformed to comply with legislation. The most effective mechanisms of legislature were concrete directions as well as consequences such as fines, showing that the (effort made for the) passing on of trustworthiness can be facilitated by external factors. But it seems that these legal frameworks are (partially) rendered invisible in the developers' everyday work once they are in effect.

Our participants also discussed the limitations related to legal frameworks and policies and how these can affect the extent to which trustworthiness might materialize, especially when ethics are considered more broadly. Edward, a project advisor affiliated with a university, believes that:

> "it's almost so well laid out by the European Commission that it inherently just seems procedural. It's just forms we have to fill out versus being really part of the entire process."

The reality of legislation or policy may not meet the reality of implementation, and some developers may only meet the guidelines to the extent that can be legally required from them. Thus, to effectively support the materialization of trustworthiness of AI in R&D contexts, legal frameworks and policies need concrete descriptions of how they might be applied in practice and what consequences might arise if they are applied incorrectly. Unfortunately, many nontechnical practices are rarely addressed in regulatory or industry documentation, though these results show the value and need to better understand such aspects of AI development. Stronger engagement with development communities, collaborations, and understanding of communication can support a move away from or beyond the dichotomy of AI development as a technical challenge and AI harm as a societal challenge. However, guidelines such as those from the EC do introduce a normativity that legislators, developers, and other stakeholders need to be aware of and reflect on (Hagendorff, 2020) given the potential trade-offs they may introduce (Blanchard et al., 2024).

Revelation

The second overarching theme in the research data is Revelation. Activities within revelation, summarized in Table 4, center around enabling trust and how this is communicated more directly. Interactions between AI developers and stakeholders outside of the AI development process matter greatly to these efforts. Our participants' experiences suggest that the revelation of or transparency about both the inner workings of AI or ML models, and processes around the development, are used to foster trust. This connects to the literature on trust, and how transparency seems to play a significant role (Durán & Pozzi, 2025), though not without (fair) criticism.

Transparency About the AI Development Process

One of the activities clustered under revelation concerns transparency about every aspect of the AI life cycle to enable and support trust. For our participants, transparency allows for discussions and reflections of data collection or generation, curation, and use practices. Rick, mid-level AI developer/researcher at a public research organization, voices the aim:

> " . . . to make the whole process transparent. So, indicate which model was used and which data was used and how the development steps and improvement steps look like. That could be helpful for developers that will integrate our model."

However, transparency can also be limited by contextual factors as they might lead to certain risks or trade-offs, for example for security. Anton, senior-level AI developer at an SME, mentions in an interview that transparency can increase vulnerability of AI systems. This corresponds to the experience that transparency is often "seen in absolute terms" (Durán & Pozzi, 2025, p. 16). However, Alex, senior-level AI developer at an SME, highlights that this does not need to be the case. In his words: "you can still have a transparent, well-documented, and reliably built interface."

Based on Rick's and Alex's statements, developers do not need to be transparent in an absolute sense. Instead, they can choose to be strategic in what they make transparent and

still be trustworthy and trusted. However, developers need to take care here due to the potential for trade-offs that can simultaneously reduce trustworthiness.

Explainability

Another core practice in the data is revelation through explainability. Through this activity, our participants aim to shed light on what is typically seen as "black-boxed" by making visible what the ML model does and/or why certain outcomes are reached. However, complications exist in making this visible for trust, and other complications may arise as well.

First, AI developers can choose many means for explanation depending on the needs of intended audiences. Explanations can be presented in numbers, icons, graphs, or even interactive modes. Participants explain that flexibility exists in developing explanations and how they can be visually represented. The choices developers make depend on who they conceive of as the main users. This can be a facilitator or a complicator of creating or facilitating trust. For example, Keith, project advisor affiliated with a university, talks about efforts that can support people with different levels of expertise:

> "They [random forests] say what they're doing directly, but even for somebody who's not trained in statistics, you can find plenty of metaphors or analogies to give that will give an intuition about what it's doing as well."

Similarly, Rick and Jacob discuss the need for domain expertise:

> "If you have tabular data or complicated data, like for example, the multi match use case, then explanations are totally different than if you want to explain the behavior of an image classifier. To analyze the machine learning model from multi machine and data, you need to be a domain expert already."

Keith, Rick, and Jacob explain how different models require different capacity levels to be able to understand the provided explanation. If the wrong type of explanation is provided, this may affect the trustworthiness of an AI system. Moreover, it is not always possible to make visible what is happening inside the black box of AI, especially for ML models. Rick explains:

> "So, in our use case, the standard model to use is a convolutional neural network, which is super complex and non-interpretable in all directions. So, yeah, that's why we need this post-hoc explanation, to even understand what it's doing."

Yet, even if post-hoc explanation mechanisms are used, it remains questionable whether this type of explanations reflect the actual inner workings of ML, resulting in what Durán and Pozzi (2025) refer to as transparency regress. Transparency regress occurs when the interpretative predictor cannot be explained, resulting in a vicious cycle of uncertainty. Our participants noted that it may indeed be impossible to fully visualize or use explainability mechanisms how AI systems come to output. Yet, they also consider that current mechanisms have their value in fostering trust as explainability measures do not reveal

the inner workings of the technology, but also the efforts of developers to present as trustworthy.

Communication and Dissemination Practices

Communication and dissemination practices are often framed as secondary to AI development. Yet, our participants shared insights that suggest that both are extremely valuable to development processes and raising trustworthiness. Communication outlets, such as a public website and social media presence helped to reach wider and often nonexpert audiences, resulting in valuable connections outside of established collaborations. Joel, senior-level AI researcher at a technical university, says:

> "[publications and general dissemination activities] gradually build up trust if we have a good exposure to the channel that they [our target audiences] access."

While outreach has a more indirect effect, its relevance should not be understated. Connections to other professionals allow AI developers to gather in-depth feedback and inspiration. Academic publications and conference participation are other examples. Harold, senior-level AI developer at a large enterprise, explains a recent development that he values:

> "You see that many conferences are now fostering this artifacts presentation. You don't just send your paper, but you also send the code and the data, so that people can rerun your code with that data and see what's going on."

Harold indicates how investments in professional relations can support perceptions of trustworthiness, especially when their audience can interact with the AI systems and provide feedback. This reflects the relational perspective on communication as discussed previously, and how trust might be facilitated and strengthened through reciprocity.

Reputation

The reputation of individual researchers, organizations, or funding agencies backing the project, were seen as another contributor to building trust. Here, the role of the EU's reputation as a funder of the participants' work was mentioned. Joel explains:

> "For EU companies, typically the results from the EU-funded projects add already one layer of trust because it's out of competition with many projects to get funding from the EU. And then, on top of that, they will also look at what kind of partners, or the quality of the partners in the consortium."

According to our participants, funding sources and collaborations are sources for establishing trust. The trustworthiness of AI-based tools or AI systems is tied to trust placed in the reputations of developers and funders, which shows that AI systems can be trusted by proxy (Durán & Pozzi, 2025). In sum, revelation practices show that trustworthiness cannot be merely derived from technical qualities but also require attention for social

processes of trust. It requires developers to perform in a trustworthy manner and to upkeep a positive reputation.

Conclusion

This study's aim is to understand how trustworthiness materializes through AI development practices and how this is facilitated by communication. Our interest in the (re)constitution of an organization of trustworthy AI development steered us toward a CCO perspective (Cooren, 2020; Schoeneborn et al., 2019). Applying materialization (Cooren, 2020) as a frame helps us understand how trustworthy AI emerges in practice. We contribute to a broader understanding of how a focus on either technical or nontechnical solutions is inadequate for solving AI-enabled societal challenges.

The two themes that emerged from our data were Transformation and Revelation. Transformation deals with the materialization of trustworthiness in AI systems, and shows that the degree to which trustworthiness materializes, is limited or varying. Not one requirement is immune to limitations that transformations bring with them. Perhaps this suggests that trustworthiness is an umbrella term (Reinhardt, 2023) by practical necessity. The second theme, Revelation, shows that AI development practices are not limited to activities transferring trustworthiness from one thing to another as additional activities support trust. According to Durán and Pozzi (2025) and Reinhardt (2023), trustworthiness does not guarantee trust as other factors come into play. Our results show AI developers implicitly accounting for this and managing activities that reveal trustworthiness to others in order to garner trust.

This study thus provides insight into the complexity behind a more material understanding of trustworthiness and trust in AI systems. It is possible that trustworthiness will never fully materialize in one AI system. The complexities of AI systems and the variations between disciplines or fields of application make it nearly impossible to determine and prescribe strong guidelines for their development. However, our results also show how trustworthiness and trust may effectively support each other when taking a communication perspective. This is also where potential risks for over-trusting lies (Reinhardt, 2023), if critical reflections on what is transformed and to what degree, and what is revealed, are absent. It is exactly here that the potential for AI harms remains (Shelby et al., 2023), further adding to the notion that blind or over-trusting is problematic (Reinhardt, 2023). We suggest that trustworthiness and trust cannot feasibly exist without appropriate oversight or evaluation.

Understandings of transformation and revelation run counter to any prioritization of technical practices to create more trustworthy AI. In fact, our findings support the notion that technical and social practices are entangled in such a way that separation is impossible. This study shows that there is a need for both breadth and depth within HMC research in AI development to include a development perspective that crosses the boundaries of functional and relational (and perhaps even metaphysical) communication aspects (Guzman & Lewis, 2020).

We argue that more attention should be brought to the social and implicit components of AI development. Many nontechnical practices are only limitedly or implicitly addressed

in regulatory and industry documentation, often with a certain normativity (Hagendorff, 2020), whereas our study shows the value and need of better understanding social aspects of AI development. Stronger engagement with developer communities, collaborations, and communication, supports a focus beyond the dichotomy of AI development as a technical challenge and AI harm as a societal challenge.

Research on trust and AI tends to approach the topic from a user perspective or focus on abstract and experimental settings or very applied situations where AI systems are already active and affecting society (Oomen et al., 2024). Debates within HMC include a focus on interdisciplinary potentials and problems in relation to more traditional methodologies and concerns about how for instance AI takes on an agential role (Banks & Graaf, 2020). Yet when focusing on AI harms, causes are normally traced back to ML models, learning algorithms, and training data (Plasek, 2016), all part of AI development. By focusing on how AI development materializes in practice, this study shows the importance of development practices in shaping AI before it becomes normalized into communicative practices. That is, communicative processes are central to how trust and trustworthiness are claimed and proclaimed.

Limitations

This study was subject to limitations due to the scope of the participants' project. The participants focused on explainability and transparency, which explains how these concepts dominated the data, even though trustworthy AI is broader as outlined by the AI HLEG (2019). Moreover, as social science partners in a technology-oriented project, we faced challenges and found opportunities (e.g., in creating a common vocabulary and understanding). Our immersion helped us find mutual ground and gain understanding of each other's language, enabling us to shed light on practices that were accepted and perhaps taken for granted previously.

A methodological limitation resulted from the COVID pandemic. We planned to work alongside use case partners for longer periods at their workplaces. However, since the pandemic, work-from-home became the norm, complicating the original plan. We instead had to implement alternative data collection methods and moved our work to online sites, which, while useful, reduced our opportunity for observing intra-organizational practices.

Future Research

Despite the limitations for the materialization of trustworthiness and trust, it does not mean that efforts toward trustworthiness are without merit. Rather, reflecting on the incompleteness of trustworthiness and the practices involved in its materialization allows for improving and strengthening future AI development and research. This presupposes cross-disciplinary collaboration, which is still limited in practice. It would require continued funding and investments to drive collaborative research and development of trustworthy AI.

Given the critical focus on development practices, future research should examine how AI development can evoke and maintain trustworthiness and trust while being highly dependent on relational contexts in organizational or social settings. Communication

scholars could directly engage with the AI development community to gain insights on how to develop AI for trustworthiness and its effects on perceptions of trust.

Author Biographies

Tessa Bruijne holds a PhD in Media & Communication, in which she combines insights from organizational communication and Science and Technology Studies to study the role of communication in developing trustworthy AI. After completing her PhD, Tessa transitioned to TNO, an applied research organization. Here, she conducts qualitative and mixed methods research and participates in collaborative projects on responsible governance and implementation of digital technologies, such as AI.

https://orcid.org/0000-0002-2872-4006

Anouk Moulds holds a PhD in Media & Communication focusing on everyday experiences of privacy, surveillance, and AI in families, work environments, and neighborhoods. With her background in Media Studies and sociology of culture, media, and arts, her expertise lies in qualitative, mixed-methods, and participatory research. Currently, her work focuses on education around (social) media literacy, digital literacy, and digital resilience of children and young people.

https://orcid.org/0000-0003-0355-9849

Jason Pridmore is a Professor of Human Centric AI in Society with a focus on Emerging Technologies and Social Change. His research interests focus primarily on practices of digital science communication, digital identification, the use of new/social media, and consumer data as surveillance practices, and digital (cyber) security issues. He leads and participates in research focused on privacy, data ethics, mobile devices, policing practices, citizenship, branding, and quantified self-movements. Jason currently participates in an advisory capacity for a range of European Union Research projects and Dutch funded projects on new technologies, privacy, and security issues.

https://orcid.org/0000-0001-9159-8623

References

AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Banks, J., & Graaf, M. de. (2020). Toward an agent-agnostic transmission model: Synthesizing anthropocentric and technocentric paradigms in communication. *Human-Machine Communication*, *1*(1). https://doi.org/10.30658/hmc.1.2

Blanchard, A., Thomas, C., & Taddeo, M. (2024). Ethical governance of artificial intelligence for defence: Normative tradeoffs for principle to practice guidance. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-01866-7

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2021). To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health*, *13*(2), 201–216. https://doi.org/10.1080/2159676X.2019.1704846

Cooren, F. (2020). Beyond entanglement: (Socio-) materiality and organization studies. *Organization Theory*, *1*(3), 1–24. https://doi.org/10.1177/2631787720954444

Durán, J. M., & Pozzi, G. (2025). Trust and trustworthiness in AI. *Philosophy & Technology*, *38*(1), 1–31. https://doi.org/10.1007/s13347-025-00843-2

Etzrodt, K., Kim, J., Goot, M. van der, Prahl, A., Choi, M., Craig, M., Dehnert, M., Engesser, S., Frehmann, K., Grande, L., Liu, J., Liu, D., Mooshammer, S., Rambukkana, N., Rogge, A., Sikström, P., Son, R., Wilkenfeld, N., Xu, K., … Edwards, C. (2024). What HMC teaches us about authenticity. *Human-Machine Communication*, *8*(1). https://doi.org/10.30658/hmc.8.11

Georgieva, I., Lazo, C., Timan, T., & van Veenstra, A. F. (2022). From AI ethics principles to data science practice: A reflection and a gap analysis based on recent frameworks and practical experience. *AI and Ethics*, *2*, 697–711. https://doi.org/10.1007/s43681-021-00127-3

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human-Machine Communication research agenda. *New Media & Society*, *22*(1), 70–86. https://doi.org/10.1177/1461444819858691

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hallett, R. E., & Barber, K. (2014). Ethnographic research in a cyber era. *Journal of Contemporary Ethnography*, *43*(3), 306–330. https://doi.org/10.1177/0891241613497749

Kleinman, Z. (2024, February 28). Why Google's 'woke' AI problem won't be an easy fix. *BBC*. https://www.bbc.com/news/technology-68412620

Marcus, G. E. (1995). Ethnography in/of the world system: The emergence of multi-sited ethnography. *Annual Review of Anthropology*, *24*, 95–117.

Oomen, T., Gonçalves, J., & Mols, A. (2024). Rage against the AI? Understanding contextuality of algorithm aversion and appreciation. *International Journal of Communication*, *18*, 609–633.

Orlikowski, W. J. (2007). Sociomaterial practices: Exploring technology at work. *Organization Studies*, *28*(9), 1435–1448. https://doi.org/10.1177/0170840607081138

Orlikowski, W. J., & Scott, S. V. (2016). Digital work: A research agenda. In B. Czarniawska (Ed.), *A research agenda for management and organization studies*. Edward Elgar Publishing. https://doi.org/10.4337/9781784717025.00014

Pequeño IV, A. (2024, February 26). Google's Gemini controversy explained: AI model criticized by Musk and others over alleged bias. *Forbes*. https://www.forbes.com/sites/antoniopequenoiv/2024/02/26/googles-gemini-controversy-explained-ai-model-criticized-by-musk-and-others-over-alleged-bias/

Plasek, A. (2016). On the cruelty of really writing a history of machine learning. *IEEE Annals of the History of Computing*, *38*, 6–8. https://doi.org/10.1109/MAHC.2016.43

Reckwitz, A. (2002). Toward a theory of social practices: A development in culturalist theorizing. *European Journal of Social Theory*, *5*(2), 243–263. https://doi.org/10.1177/13684310222225432

Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. *AI and Ethics*, *3*(3), 735–744. https://doi.org/10.1007/s43681-022-00200-5

Schoeneborn, D., Kuhn, T. R., & Kärreman, D. (2019). The communicative constitution of organization, organizing, and organizationality. *Organization Studies*, *40*(4), 475–496. https://doi.org/10.1177/0170840618782284

Shelby, R., Rismani, S., Henne, K., Moon, Aj., Rostamzadeh, N., Nicholas, P., Yilla, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). *Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction* (arXiv:2210.05791). arXiv. https://doi.org/10.48550/arXiv.2210.05791

Shove, E., Pantzar, M., & Watson, M. (2012). *The dynamics of social practice: Everyday life and how it changes*. Sage.

Stix, C. (2022). Artificial intelligence by any other name: A brief history of the conceptualization of "trustworthy artificial intelligence." *Discover Artificial Intelligence*, *2*(26), 1–13. https://doi.org/10.1007/s44163-022-00041-5

van de Poel, I. (2020). Embedding values in Artificial Intelligence (AI) systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

van Krieken, R. (2024). The organization of ignorance: The Australian 'robodebt' affair, bureaucracy, law and politics. *Critical Sociology*, *50*(7–8), 1379–1398. https://doi.org/10.1177/08969205241245257

Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions*, *26*(4), 42–46. https://doi.org/10.1145/3328485

HUMAN-MACHINE
COMMUNICATION