Unpacking Trust Dynamics in the LLM Supply Chain

An Empirical Exploration to Foster Trustworthy LLM Production & Use

Balayn, Agathe; Yurrita, Mireia; Rancourt, Fanny; Casati, Fabio; Gadiraju, Ujwal

DOI
10.1145/3706598.3713787

Publication date
2025

Document Version
Final published version

Published in
CHI 2025 - Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems

Unpacking Trust Dynamics in the LLM Supply Chain: An Empirical Exploration to Foster Trustworthy LLM Production & Use

Agathe Balayn
ServiceNow
Amsterdam, Netherlands
Trento University
Trento, Italy
Delft University of Technology
Delft, Netherlands
agatheb.research@gmail.com

Mireia Yurrita
Delft University of Technology
Delft, Netherlands
m.yurritasemperena@tudelft.nl

Fanny Rancourt
ServiceNow
Montreal, Quebec, Canada
fanny.rancourt@servicenow.com

Fabio Casati
University of Trento
Trento, Italy
ServiceNow
Zurich, Switzerland
fabio.casati@gmail.com

Ujwal Gadiraju
Web Information Systems
Delft University of Technology
Delft, Netherlands
u.k.gadiraju@tudelft.nl

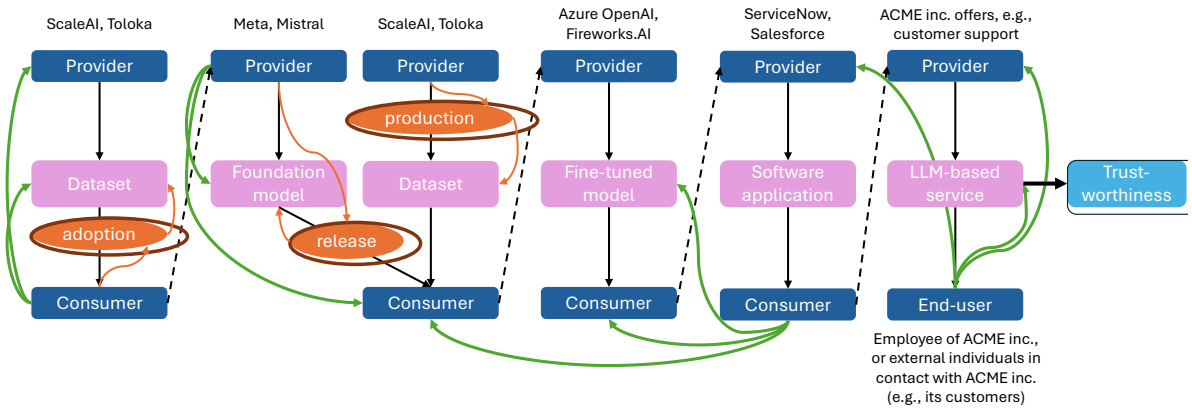

Figure 1: Illustrative example of an LLM supply chain inspired by the descriptions from our study participants. Respectively in pink and in blue, we show the technical artifacts and actors (with examples) involved in the supply chain. In orange, we provide examples of decisions these actors might have to make concerning the LLM production at different junctions of the LLM supply chain. Note that along the supply chain, each actor adopts both provider and consumer roles (dashed arrows) since their production work relies on the consumption of existing technical artifacts. Besides, in practice, one actor might be involved as the provider of multiple technical artifacts (e.g., one actor might produce both the fine-tuned LLM model and the software application around it, and might even provide this software application as a service to its customers). The current representation is simplified, as some actors downstream of the supply chain might also be involved upstream (e.g., ACME inc. might give some of its data to the fine-tuned model provider in order to fine-tune this model further). The green arrows represent examples of trust relations that traverse the LLM supply chain (the tail of the arrow refers to the trustor and the arrow's pointer to the trustee) and might ultimately affect the trustworthiness of the LLM-based service.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713787

Abstract

Research on trust in AI is limited to several trustors (e.g., end-users) and trustees (especially AI systems), and empirical explorations remain in laboratory settings, overlooking factors that impact trust relations in the real world. Here, we broaden the scope of research by accounting for the supply chains that AI systems are part of. To

this end, we present insights from an in-situ, empirical, study of LLM supply chains. We conducted interviews with 71 practitioners, and analyzed their (collaborative) practices using the lens of trust drawing from literature in organizational psychology. Our work reveals complex trust dynamics at the junctions of the chains, with interactions between diverse technical artifacts, individuals, or organizations. These junctions might constitute terrain for uncalibrated reliance when trustors lack supply chain knowledge or power dynamics are at play. Our findings bear implications for AI researchers and policymakers to promote AI governance that fosters calibrated trust.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; *Empirical studies in collaborative and social computing*; • **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → *User characteristics*; **Management of computing and information systems**.

Keywords

trust in AI, large language models, collaborations, AI supply chain, calibrated trust

ACM Reference Format:
Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. 2025. Unpacking Trust Dynamics in the LLM Supply Chain: An Empirical Exploration to Foster Trustworthy LLM Production & Use. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3706598.3713787

1 Introduction

Trust is essential for collaboration [131]. Given the rise of human-AI collaboration modalities in the use of AI systems (e.g., AI outputs as recommendations to the user of the AI system [2, 110]), HCI researchers have started exploring trust relations between humans and AI systems [73, 126, 133], and the conditions for developing calibrated trust in deployed AI systems [8, 11, 121, 145]. To this end, they have limited the trustee (i.e., the actor that is trusted) to the AI system, and the trustor (i.e., the trusting actor of such an AI system) to the user of the AI system who is conducting a task with the help of the system [66, 78, 134], the decision-subject in the task at hand [2, 132], or the general public [2, 68].

The collaboration between humans and AI is however not limited to their interactions during task execution. The development and deployment of AI systems also result from collaborative work between actors in the AI supply chain [29, 34]: the collaborations at each of its junctions might then be governed or influenced by trust dynamics (see example in Figure 1). For instance, an AI system such as a Large Language Model (LLM) requires an organization to develop a foundation model, the same or another organization to fine-tune the model and deploy it in an application, and a consumer organization to adopt this application. There, before an end-user even starts working with the LLM-based service to carry out certain tasks, the consumer organization probably first develops trust in the foundation or fine-tuned model (or even in the organization developing these models [132]) to then adopt the LLM-based service. Decades of organizational psychology research [5, 7, 53, 72, 76, 92, 102, 111, 120, 122] and the few in-situ studies of trust in AI have already hinted at the existence of such trust relations within (AI) lifecycles inside the supply chain [18, 132] and of organizational factors impacting these relations [66]. By playing a key role in the use of AI systems but also in their adoption, deployment, and development, trust might ultimately affect the resulting trustworthiness of these AI systems, and how responsible their production is.

In this work, we argue that the study of trust between humans and AI systems should be expanded beyond end-user interactions and consider the complex AI supply chains. By building on valuable prior research, we explore the trust dynamics that may power or hinder the AI supply chain, and draw the landscape of trust-related notions (e.g., vulnerabilities and expectations, miscalibration, reliance) which might impact the trustworthiness of the resulting AI system and its production. We characterize these notions that are typically not considered in the study of trust in human-AI collaborations, to inform future research opportunities for building trustworthy AI systems. To this end, we conducted a qualitative empirical study. We interviewed 71 actors of the LLM supply chain (over 3600 minutes of recording), including providers, deployers, and consumers of such LLMs. We enquired about their work practices and collaborations with colleagues and other organizations. We focused on LLMs due to their timely relevance and concomitant supply chain governance challenges [55]: with the rise of LLMs, many small and medium-sized companies depend on foundation models provided by larger organizations and fine-tuned to their specific needs by intermediate organizations. We analyzed the interview transcripts through a trust lens, aided by prior work on trust in organizational psychology.

Our results illustrate the importance, diversity, and complexity of trust relations beyond known trustors, trustees, and trust activities, both within and across organizations. Among others, these include generalized trust in technologies and trust in individuals and organizations complementing trust in specific AI systems. Our findings also point out new factors that impact trust relations and can be vectors for uncalibrated trust, such as the dependencies between trust relations involving different trustees or trust activities, vulnerabilities revolving around organizational reputation, or historical relations. Finally, our insights show that the LLM supply chain junctions are not always sustained by trust but by reliance, since trust might not be practically and conceptually achievable for all trustors.

These results bear implications for AI trust researchers and policymakers who intend to regulate LLMs and foster their trustworthy production and adoption. They encourage interrogating the appropriate object of study along the AI supply chain, whether it should be trust or resulting behaviors such as reliance, blind or calibrated trust, and trust in AI systems, trust in AI technologies, trust in technical artifacts, or trust in individuals and organizations. They also prompt revisions of current transparency and literacy tools and other modes of governance that communicate trustworthiness cues, to account for the new trustees, trust factors, and causes of miscalibrated trust that emerge from the AI supply chain. Finally, our results shed light on power relations that impact these trust dynamics, and the ways they can be addressed to build trustworthy

AI systems that fulfill different stakeholders expectations. We hope that this paper acts as a call to broaden the scope of policy and research work, and to particularly address pressing issues (e.g., uncertainty, accountability horizon, power relations hindering trust) in complex trust dynamics along AI supply chains.

2 Background & Related Work

2.1 The Fundamentals Of Trust

What is trust? Trust is an attitude relevant in many domains. Next to trust between humans and automation technologies [76], there are trust relations between various trustors and trustees, such as trust between individuals and especially coworkers in a supply chain [120], trust between individuals and organizations or institutions, or trust between organizations. Researchers investigate trust through diverse lenses, be it organizational, sociological, interpersonal, psychological, and neurological perspectives [76].

In this work, we primarily rely on insights stemming from organizational psychology [86], the research domain interested in the attitudes and behaviors of employees in organizations. This domain perfectly coincides with our research context, which deals with the collaborations between individuals and/or organizations along the AI supply chain. Researchers in organizational psychology [116] characterize trust relations with a **trustor**, a **trustee**, and a trust **activity**. For instance, the user of an LLM-based service might trust the service to summarize texts that do not require expertise (activity 1), but might refrain from relying on it when they need to summarize scientific papers (activity 2). Despite trust not having an agreed-upon definition, a majority of work agrees to further characterize the trust relation by the **positive expectation** the trustor has for the trustee, and the trustor's **vulnerability** and uncertainty vis-a-vis the trustee [76, 133].

Researchers argue that trust is useful in handling complex situations as it contributes to completing activities that require delegation [76] –trust might lead to the trustor's reliance and compliance behaviors on the trusted party. Ideally, trust should be **appropriate or calibrated** [87], i.e., there is alignment between the perceived and actual trustworthiness of the trustee, or in other words, the trustors solely trust trustworthy trustees [64]. Instead, inappropriate trust might result in inappropriate reliance behaviors such as misuse, disuse, or abuse of a technology [64].

In the context of AI, studied trustors are either a user [66, 78, 134] of an AI system, an individual who is the subject of an AI output (termed "decision-subject") [132], or an external observer (the public) [2, 68]. The trustee is the AI system, conceptualized as one entity that outputs predictions about input samples [57] or as a general technology [68]. Past research is particularly relevant for understanding the conditions that lead to the *adoption* of AI-based software systems [5, 66], and appropriate reliance on such systems. However, AI systems, especially those based on LLMs, are much more complex artifacts [29]: they result from supply chains composed of different actors collaborating or relying on each other's work to develop and deploy the AI systems. Trust or lack thereof might affect reliance behaviors among these actors, and the nature of those behaviors ultimately shapes the behavior that the AI systems manifest downstream. Hence, we broaden the scope of research and inspect trust across the actors in the LLM supply chain.

Which factors impact trust among individuals or towards technologies? Researchers have identified several factors that impact trust [80, 86]. We will investigate whether the same trust factors apply along the AI supply chain.

Trust first depends on the **trustee's trustworthiness characteristics**, whose impact is mediated by the affordances that expose them to the trustor, such as transparency frameworks and documentation [80]. The ABI framework [86] defines ability, benevolence, and integrity as the primary trustee's trustworthiness characteristics. "*Ability* is that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain", "*benevolence* is the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive", while "the relationship between *integrity* and trust involves the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable" [86]. Specifically for AI, researchers have identified factors of trustworthiness that are intrinsic or extrinsic to AI systems and related to the ABI framework [13, 64]. These include the quality of the AI outputs [5, 39, 64] and ethical considerations about the AI system (e.g., discrimination) [98] and the ways they are handled by the system producer [5] in terms of ability; explanations provided for each AI output [64] in terms of integrity; transparency about the AI system [13, 39] and designed user-interaction [5, 64] in terms of benevolence. Available information about the organization developing the AI system, e.g., demonstrated expertise, guidelines for integrity, and social responsibility, can also impact trust [13].

Trust also depends on the **trustor's inherent and acquired characteristics** (e.g., gender, age, income, and employment situation [98]) and their attitudes (e.g., disposition for trust) [5, 76]. Finally, the **context of the trust activity** mediates trust by impacting the trustor's systematic and heuristic processing of the trustee's affordances [80]. In the context of an AI system conducting a prediction activity, the activity characteristics are important, such as whether it is technical or requires social intelligence [49], whether it has high stakes [4], how complex it is [63], and the application it serves [98].

2.2 Trust & Supply Chains

Researchers in organizational psychology explicitly discuss trust within organizations [7, 53, 92, 102, 120] or supply chains [72, 111, 122]. They have shown that trust is developed through an ongoing relationship between individuals within the supply chain, and it constitutes a competitive advantage for organizations in the supply chain and a catalyzer for collaborations and hence performance [72, 111]. There, different types of trust exist, such as trust in the integrity, competence, or predictability of the trustee, as well as calculative trust (economic calculation for assessing the benefits and costs that can be derived from creating and sustaining a relationship) [48, 92, 120]. Such trust is impacted by different factors, such as bargaining power, contracts, relationship duration, exchanges of information and confidentiality, the reputation of the trustee, and commitment [48, 72, 122]. The organizational context, i.e., interactions between co-workers that inform about the trustee,

and indirect information related to the trustee such as based on institutional trust and meta-trust, mediate trust [76].

In the context of AI, researchers have investigated AI lifecycles [93, 101] and AI actors involved in trustworthy AI questions [36, 52, 62, 103, 118]. There are now explicit references to AI as a supply chain and investigating the implications of this notion for accountability [29, 139], explainability [37, 118], and political economy [33, 34]. However, these ideas of lifecycles and supply chains are still rarely acknowledged in studies of AI trust. A few works show that trust in AI systems is impacted by characteristics beyond the AI system, such as the identity of the organizations around the AI system [13, 132], the certainty on who is accountable for it, the interlocutor in case of complaints, or the existence of AI ethics governance processes, and the lack thereof [98]. Furthermore, to the best of our knowledge, Toreini et al. [125] are among the only ones that present AI as a pipeline with various stages involving various activities and technical components, and that discuss trust along these different dimensions. Besides, Benk et al. [15] and Jacovi et al. [64] proposed to reframe the investigation of trust in AI, to trust in the socio-technical system surrounding the AI system, especially accounting for the various human stakeholders interacting with AI. Yet, none of these works have empirically and comprehensively studied such a supply chain in relation to trust — a gap we address.

2.3 Trust In Policy Spaces

To the best of our knowledge, trust dynamics along AI supply chains lack not only visibility in HCI research but also explicit account in policies and regulations discussing trustworthy AI. For instance, the proposal of the European Union for regulating AI (AI Act [30]) mentions trust as an objective [75] without delving into the diverse needs for trust or reliance and the conditions for calibrated trust.[1] Instead, it relies on the idea that regulations revolving around the ability of the systems and consequently their trustworthiness will enhance trust in them, neglecting the other trust actors and factors that the literature has been hinting at. Similarly, many industrial AI governance initiatives seem to conflate trust and trustworthiness within their guidelines for trustworthy AI [109]. More broadly, several researchers have already pointed out that policy documents do not reflect the complexities stemming from the fact that AI algorithms are integral parts of supply chains [9, 27, 29, 30, 40]. Hence, by investigating the intricacies of trust dynamics in AI supply chains, we contribute a first set of empirical insights to characterize AI supply chains' realities and to inform future policies. Beyond such contribution, our work offers a meta-contribution: it illustrates how HCI research can inform future policies [142] by empirically unpacking the regulated site [29].

3 Method

3.1 Approach Of The Study

There is a well-acknowledged need for more qualitative methods to advance our understanding of *trust in AI*, and for more empirical *in-situ* investigations of the AI supply chain and the prevalent trust dynamics that characterize it [18, 29, 66, 132]. Qualitative investigations are particularly useful to explore broad and understudied areas such as AI supply chains, and suitable to elicit new trust factors due to the rich insights they can result in [115]. Hence, we chose to conduct semi-structured interviews.[2] We adopted an iterative approach to conduct and analyze these interviews. With knowledge of prior work on human-AI collaboration and awareness of the characterization of trust in other fields, we started with a broad exploratory question in mind –"*What are the trust dynamics prevalent within the AI supply chain and how are they relevant to trustworthy AI objectives and concerns?*" Then, while conducting the interviews, analyzing them, and iteratively identifying themes and codes, we refined our objects of inquiry and the interview questions based on the importance of our early findings (e.g., how surprising they are, how acknowledged in the literature they are, how relevant and impactful toward trustworthy AI they are).

3.2 Interview Setup

We adopted an exploratory approach to these semi-structured interviews to investigate potential trustors and trustees in the LLM supply chains, and interviewed a broad range of supply chain actors. Before the interviews, we asked them to fill in a questionnaire, surveying them about their positions in their organization, their knowledge and prior experience with AI and LLMs, their understanding of trustworthy AI topics concerning LLMs (e.g., fairness, explainability), and their preferences in terms of responsible AI values. This provided us with useful information to prepare for the interviews, for instance to understand the relevant activities of the participants and to identify the vocabulary to employ to avoid misunderstandings. During the interviews, to avoid biasing our interview participants, we refrained from directly prompting them about trust in LLMs. Instead, we questioned them about their practices and challenges vis-a-vis the LLM and broadly vis-a-vis the daily tasks they have to execute. We asked about their awareness of the other actors in the supply chain and their relations with these actors. We also discussed their opinions about LLMs and their benefits and risks, since their AI mental model could shape their perceptions of AI trustworthiness and their overall trust in the supply chain [2]. When mentioned by the participants, we further prompted them about trust. We conducted two pilot interviews and iterated on the questions in the questionnaire to make sure that the information collected would lead us to obtain insights relevant to trust. The interviews were reviewed and approved by our institutional ethics committee. The interviews lasted between thirty minutes and one hour per participant.

3.3 Study Participants

Recruitment of the participants. We interviewed 71 participants across 12 private organizations. We recruited several participants working within the same organization to better understand and compare reported trust relations. All participants contributed to developing or deploying LLMs or LLM-based services, or used such services regularly either for their professional work or personal use.

[1] "This proposal aims to implement the second objective for the development of an ecosystem of trust by proposing a legal framework for trustworthy AI. [..] Rules for AI [..] should therefore be human-centric, so that people can trust that the technology is used in a way that is safe and compliant with the law, including the respect of fundamental rights." [30]

[2] See the Appendix for reflections on our positionality, examples of interview questions, and more details about our process for analyzing data.

Through a mixture of snowball and convenience sampling, starting via our professional network including practitioners responsible for AI adoption and AI development decisions, we recruited actors who played different roles along the supply chain and vis-a-vis LLMs (cf. Table 1). Given our exploratory lens and since we did not know which actors of the LLM supply chain were relevant to (dis)trust dynamics that impact AI trustworthiness, we ensured simultaneously breadth and depth in our participant recruitment process. We not only made sure to recruit participants from different organizations and with different positions in these organizations, but we also adopted a purposeful approach to recruitment when we identified particularly relevant roles. For instance, since we realized that UX researchers were often in conflict with engineering teams and discussed distrust towards them, we made an effort to recruit more UX researchers. Similarly, because we immediately identified misalignment between product owners from provider and consumer organizations, we recruited more product owners. While conducting the interviews, we reached saturation and a fixed set of themes and codes with fewer than our 71 participants, but reached this number to ensure coverage in terms of actors' positions in the supply chain. Participants working for provider or consumer organizations were recruited voluntarily based on their intrinsic motivation. End-users of LLMs received monetary compensation for participating in our study (50 US dollars for one-hour long interviews).

Descriptive attributes of the recruited participants. Our participants' tasks covered a plurality of activities concerning the production or consumption of LLMs. These activities included engineering (AI developers, AI researchers), user experience (user research, content design), business (product managers, business analysts, customer service), and governance (legal and risk teams, government relations). Their practical tasks involved, e.g., implementing, testing, and overseeing LLM-based services, or strategizing around the requirements of such services and broader organization (e.g., functional requirements, definition of risks, etc.). Their responsibilities spanned from decision-making (executives, process owners) to product managers, designers, and developers. Finally, we also interviewed end-users of the LLM-based services developed by our participants. They used the services for tasks outside the production of an LLM, such as to facilitate their job as customer support agents.

The supply chains in which our recruited participants were involved were interconnected, sharing the organization that developed the foundation model or the one that fine-tuned it. Our participants' organizations that participate to producing LLM-based services spanned various specializations (e.g., data annotation, LLM fine-tuning). Those consuming the services spanned various sectors (e.g., banking, insurance). The LLMs involved allowed users to conduct tasks such as text summarization, text improvement, document search from internal knowledge bases, or retrieval-augmented text generation. These LLMs were intended for professional use, such as for facilitating tasks in the workplace, e.g., text generation for marketing campaigns, or customer-service work.

Table 1: Statistics about our interview participants. In each row, the numbers in parentheses indicate the number of participants within each category (participants can only belong to one category per row). We use multiple numbers within the same parentheses when multiple, distinct, organizations fall within the same category, so as to show the distribution of participants who work in each of these organizations.

Dimension	Values and counts
Location in the supply chain	LLM provider (34, 14), data and annotation provider (2), consumer:banking (3, 2), consumer:insurance (1), consumer:caregiving-infrastructure (2), consumer:pharmaceutical (2), consumer:consulting (1, 1), consumer:hardware (2, 5), independent end-users (2)
Primary role	AI governance (5), legal / risks (10), product manager (9), UX research (9), UX design (3), business-oriented (8), AI researcher (8), AI engineer (7), software engineer (5), customer service agent (2)
Gender	Woman (28), man (43)
Work location	Canada (16), US (33), India (8), Italy (1), Switzerland (4), Netherlands (7), Belgium (1)
Experience with AI	3- years (22), 4-5 years (17), 6 years (15), 7 years (10), 8+ years (6)
Educational background	Computer science and AI (20), computer science (15), UI, UX and psychology (13), legal and governance (12), business (9), no study mentioned (2)

3.4 Data Analysis

We turned the interview recordings into anonymized transcripts and analyzed them using reflexive thematic analysis [17], focusing on latent meanings, adopting a deductive and inductive approach. This method is particularly relevant for our research since it enables a thorough exploration of the collected data to identify and interpret patterns therein. We proceeded with three stages of analysis. Note that these stages were iterative: due to the large number of interviews, we conducted these stages sequentially for each interview transcript. First, we coded transcripts with two criteria in mind to familiarize ourselves with the transcripts. On the one hand, we open-coded the transcripts with surprising, important, meaningful insights, also relying on memos that we wrote during and after each interview. On the other hand, we coded the interview transcripts systematically with initial marker codes to identify aspects of trust discussed in prior literature (see Section 2). For instance, to annotate trust factors, we used the ability, benevolence, integrity model of trust (ABI model) [86] to guide our investigation of trust factors in the AI supply chain. The model conveniently separates trust from trustworthiness, considers trustworthiness as one of the antecedents to trust, and characterizes the main trustee-related factors which might consequently impact trust relations in an organizational context. Additionally, prior literature allowed us to broaden our investigation, not only centering our focus on AI systems and their users, but also reminding ourselves of the potential relevance of other actors, both in terms of trustors and trustees, be it individuals, organizations, or technical artifacts. In the second

stage of analysis, we investigated how our prior descriptive codes associate together using iterations of digital affinity diagramming. Organizational psychology literature partially drove our analysis here, helping us articulate the identified patterns such as the types of trust relations. In the third stage of analysis, *axial coding*, we developed the larger themes that constitute our paper. While revising these themes, existing literature in AI and in organizational psychology enabled us to assess the novel or confirmatory character of our findings, and to identify where to draw more attention. Table 2 presents an overview of the final themes, sub-themes, and examples of finer-grain codes associated to these themes.

4 Results

Prelude: Overview of the LLM Supply Chain

Since the supply chains we observed within our study are complex,[3] we clarify in Table 3 their components before delving into our findings. In the rest of this paper, we will refer to "upstream" and "downstream" entities to designate entities that are relatively higher or lower in the supply chain. For instance the consumer of a foundation model who builds an LLM using this model is downstream the provider of this model and upstream the consumer of their LLM. Interested readers can refer to the supplementary material for an overview of the AI supply chain complexities.

4.1 Theme 1: Nature Of Trust Components Along The LLM Supply Chain For Trust Directed Toward LLM-Related Artifacts

4.1.1 Intricate trustors and trustees. We find that trust in the LLM-based services produced by the supply chain is one of the primary drivers of the LLM supply chain, as it impacts both production and consumption junctions. Yet, we also find high multiplicity in who the exact trustor and trustee in such trust relations is. In practice, multiple trust relations between a supply chain actor and one of its technical artifacts hindered or fostered the development and consumption of the LLM.

Multiplicity of the trustor. Trust relations were mentioned between **(1) the individual end-user** and the LLM-based service. For instance, a product manager in a deployer organization (P39) responsible for assessing the value of adopting new technical systems, discussed whether to adopt an LLM-based service. The aim of such service is to help the organization's employees in querying information related to human resources. P39 thus discussed trust of the internal end-users towards the service: *"We always want to dig into trust, and what our users need to trust the experience with the system. Otherwise, they could simply not use it, and overload our help desk with phone calls".*

In addition, we found trust relations between the **(2) consumer organizations** and the LLM-based service or its individual technical components such as the fine-tuned model or the training dataset. Indeed, actors relying on a technical artifact from an upstream provider organization often hindered the adoption of the artifact due to distrust, mentioning potential artifact issues like the low performance and biases of the LLM model. For instance, a solution success consultant (P35) working within an LLM-provider organization, discussed how some consumer organizations are interested but reluctant to use their services to handle human-resources-related tasks. This is due to distrust in the capabilities of LLMs: *"There was a hesitancy to use the LLM. There was a lack of trust. [customers typically say] 'I want to see it proven before I bring it into my organization.' [..] They see the value, but they still question elements such as 'is the user [of the potential service] getting marginalized because they asked about their benefits eligibility in a non-standard way? Does it bring discrimination within the organization?"* Note that two thirds of the consumer organizations we interviewed expressed such distrust. Those who work within highly-regulated sectors such as banking, insurance, and pharmaceutical wondered about privacy of the end-users and errors in the service outputs.

Finally, we also found that **(3) the artifact providers** themselves needed to trust their artifact to release it to consumers. Three of our participants within provider organizations and four of the participants within consumer organizations discussed specific roles within their organizations that hold great power on deployment decisions, and make these decisions only when they trust the technology. For instance, an AI engineer (P50) in a provider organization who contributes to building the wrapper of the LLM model, discussed the trust other stakeholders should build in the outputs of their engineering work for these outputs to become useful and deployed: *"First, you need to show the quality to your product manager and get buy-in from them. Then, we have an AI governance board that needs to vet any application that is built on top of an AI component."*

Multiplicity of the trustee. While the need to trust the appropriate functioning of a technical artifact is not necessarily a new idea (at least with regard to the end-users of an AI system), we found that the exact expectations across the LLM supply chain junctions were complex, ambiguous, and subjective. The aspects of an LLM-related trustee that trustors discussed varied in terms of nature and scope. The trustee might be the **LLM-based service**, which includes the LLM model and the in-use components around it such as the filtering workflows around the LLM outputs, the software system around the LLM model, and even potentially the hardware infrastructure that powers this LLM. This was especially the case for consumers and end-users. The trustee can also be specifically the **LLM model** itself, or its **build-up components**, such as the training dataset or the data annotations used to build the LLM model. This was especially discussed by providers of such artifacts and consumers who built other artifacts on top thereof. Finally, six of our participants working for consumer organizations discussed simultaneously trusting the LLM-based service for matters of accuracy, and its built-up components. This was because the quality of these components should potentially indicate the quality of the final service and because they cared for how trustworthy and responsible the production of the service was, e.g., in terms of dataset privacy.

We also identified the presence of **generalized trust**, i.e., trustor's trust in a trustee without a specified activity, especially for technical researchers, UX developers, and consumer organizations. They trusted broad concepts, such as trust in technology or in the potential of LLMs (e.g., for productivity increase, for social good, etc.), or in the science behind LLMs. Such trust sustained development

[3] Figure 1 shows a concrete example of an LLM supply chain that our participants were involved in.

Table 2: Themes developed from the analysis of the interviews. Each theme is illustrated by its primary sub-themes and a few intermediate codes that were used during the analysis and the synthesis of the study results.

Sub-themes	Codes
Theme 1: Nature Of Trust Components Along The LLM Supply Chain For Trust Directed Toward LLM-Related Artifacts	
•Intricate trustors and trustees	Multiple trustors. Different aspects of LLM systems for which participants discuss trust: deployed system, underlying LLM algorithm, development steps. Generalized trust toward AI or scientific progress.
•Diverse trust expectations and related trust factors	Expectations and vulnerabilities differ across trustees and trustors. Trustee's factors aligned with the ABI model and these expectations: ability of technical artifacts (e.g., accuracy, latency); benevolence, integrity of the artifact.
•Trust (mis-)calibration	Misinterpretations of technical artifacts: algorithmic literacy, complexity of systems (e.g., dynamicity, lack of information). Distrust from AI uncertainties.
Theme 2: Characterization Of Interpersonal And Interorganizational Trust Along The AI Supply Chain	
•Trust between diverse actors	Inter-personal trust (equal-level colleagues, or trust up or down hierarchies), organizational trust (towards upstream or downstream, or in-between teams), generalized trust in global institutions and organizations. Trust with regard to AI and non-AI related topics; trust bi-directionality.
•Trust factors aligned with the ABI model for individual and organizational actors	Ability (e.g., academic reputation, expertise, transparency), benevolence (e.g., responsible AI policies, perceived motivation for social good, merging of technical artifact and surrounding organization), integrity (e.g., governance structures, ethical behavior, ethical development mechanism) of each type of entity above. Trust cues (e.g., communication and publication releases)
•(Mis-)calibration / distrust	Apprehension toward individuals and organizations preventing integrity and benevolence, incentive to foster calibrated trust.
Theme 3: Interplay Between Multiple Trust Relations, Trust Attitudes And Reliance Behaviors At Supply Chain Junctions	
•Interplay between trust relations across trustees	Required trust in combined trustees, substitute trustees, trustee's influence on trustors, past trust.
•Non-linearity between trust and reliance	Miscalibration –wrong perception/belief of third-party's trust relation, misinterpreted transitivity of trust, lack of relevance for historical trust relations. Impossible trust building: misaligned expectations –disparate weight attribution, disparate perception of ABI properties–, different conceptions of AI challenges. Reliance without trust: competing interests, absence of alternative.

Table 3: The primary entities that constitute an LLM supply chain, and that are relevant to the study of the trust dynamics that power such a chain.

Entity	Explanation
Technical artifact	Technical components used to build an LLM-based service such as an LLM model, a training dataset, or data annotations; or components that constitute this service when in-use, such as the fine-tuned LLM model, the workflows used to monitor or filter its outputs, etc.
Individual	Individuals within the supply chain who occupy different roles within the organizations that employ them, be it technical roles, user-research ones, customer-oriented ones, etc.
Organization	Organizations that employ the individuals, and participate in producing and providing or consuming a technical artifact (this can be the LLM-based service itself or one of the technical components).
Junction	Sites of the supply chain where primary decisions about the technical artifacts are made. E.g., decisions to produce the artifact, release it, access it and adopt it, or even decisions to contest its usage.

efforts and adoption at the production and consumption junctions. For instance, an AI engineer (P52) who has observed the transition of his provider organization from developing traditional machine learning applications to LLM-based services, talked about consumer trust: *"We have classic AI that gives [consumers] information about things. They trust that. So they already have a level of trust with us and our AI technologies that carries over to our new generative AI features."*

4.1.2 Diverse trust expectations and factors impacting such trust relations. Because of the many trustors and trustees, we found a multitude of trust-related positive expectations and vulnerabilities. Positive expectations and vulnerabilities engaged the social conscience of the trustors: they revolved around the **technical artifacts constituting the LLM-based system and the direct impacts such artifacts can have on end-users**. At the system level, expectations revolved around the capability of LLM-based services to create value for their consumers through their outputs, i.e., providing informative and trustworthy answers to queries; while limiting the risks of these outputs to impact them negatively, e.g., by using offensive language or being discriminatory. For individual technical components, positive expectations were primarily about these components being effective and efficient for their intended task. For instance, sixteen AI engineers and researchers and the data steward all talked about using a rather diverse dataset to train less biased LLMs, or relying on a filtering workflow that is fast enough for the user of the LLM-based service not to wait to receive

an answer from the LLM. Participants whose roles involved AI governance discussed the components being ethically produced, such as collecting dataset annotations while appropriately treating the data annotators, or collecting data samples that do not contain any copyrighted data; and the components not having any other negative impact such as environmental impact [71, 117]. Additionally, seven participants mentioned **individual risks and vulnerabilities**, and worried about losing their job if they were blamed for any issue related to the artifact.

Expectations and vulnerabilities could also relate to the financial, operational, legal, compliance, and reputational **benefits and risks of the provider and consumer organizations** in the supply chain. Six employees within provider organizations discussed benefits for their organizations: the capabilities of the provider's service could support them in retaining competitiveness vis-a-vis the LLM-related offers other organizations might make. Three and four product managers in provider and consumer organizations respectively, also referred to the capability of the LLM-based service to speed up the consumers' work and support them in cutting costs by increasing their productivity. Finally, sixteen interview participants referred to (shared) reputational risks between the provider and consumer organizations in case of unexpected or uncontrolled problematic outputs, with one participant citing the OpenAI lawsuit [51] as an example of such risk. In that regard, an individual end-user of an LLM-based service (P70) who uses it to carry out tasks for their own business, e.g., refining emails for their clients or revising posts for their website, discussed the financial and legal risks that flawed outputs provided by the service could cause them: *"If it's wrong and I blame AI, that's gonna make me look unprofessional in front of my clients. So I try not to overuse it. I'm probably going to get sued or in trouble from my clients. If I'm using AI as help, many people don't trust it already. I'd have to take responsibility for that as the big companies are probably gonna blame me [for the hypothetically incorrect usages of the service by P70]."*

The trust factors that trustors typically paid attention to directly related to the trust expectations and vulnerabilities above. These factors were also aligned with those that prior works have investigated and that communicate the trustworthiness of an AI system [80], and adapted to the specificity of LLM models (in comparison to more traditional machine learning models). For instance, the accuracy of the LLM-based service and its rate of offensive outputs, its internal inference mechanisms and the data sources on which it relies, and the ethicality of the ways in which the technical components were built, were discussed with regard to the ability, process integrity, and intended benevolence of the service. Note that the way ability, benevolence, and integrity were instantiated differed depending on the trustor and trustee considered, as we explained that trustors had different expectations for different scopes and nature of technical artifacts. For instance, while certain AI engineers primarily discussed the latency for an AI system to produce an output, certain product managers in consumer organizations discussed whether the LLM outputs could be offensive to their end-users. In Section 4.3.1, we will show how additional factors that were not related to any technical artifact further got entangled in the assessment of a system's trustworthiness.

4.1.3 Trust miscalibration as a result of the complexity of technical artifacts. Next to the trustor's lack of knowledge about the functioning of AI algorithms that is typically pointed out as a cause for miscalibrated trust, we found that the complexity of the trustee and its definitions created more opportunities for miscalibrated trust. Seven participants within provider organizations who described interactions with consumers, as well as two UX developers, two UX researchers, and two solution success consultants who discussed their own experiences reported two main causes for miscalibrated trust. The lack of understanding about the different components that constitute an LLM-based system and their interdependence were the main causes of misinterpretation of trustworthiness cues building a wrong understanding of a system's ability. For instance, one participant (P8) in a provider organization mentioned a consumer organization who wrongly distrusted the LLM-based service because it did not grasp that different components of LLM-based services impacted each other, and that issues in the outputs of the LLM could be prevented using a filtering workflow. Similarly, an AI researcher (P12) contrasted how certain product managers within provider organizations and consumers were not aware of the dynamicity of the LLM supply chain (refer to Edwards [40] for more information about AI dynamicity), and therefore did not re-calibrate their trust when technical artifacts were updated. *"People were asking how we manage the models, the engineering [of the surrounding software], and the* version control. *It's the basis of the trust. If you can't know what you put in production, then the users won't know either, and they won't trust what you do, or trust you while they shouldn't."* Interestingly, three AI developers who were aware of the dynamicity of the supply chain discussed that they continuously distrust the LLM-based service as the updates could negatively impact its ability at any time.

4.2 Theme 2: Characterization Of Interpersonal And Interorganizational Trust Along The AI Supply Chain

4.2.1 Intra- and inter-organizational trust prevalent at all junctions. Next to trust toward technical artifacts, we found that trust towards individuals and organizations was prevalent along the supply chain. First, **generic inter- and intra- organizational trust** that is not specific to the production of the LLM-based service was necessary for organizations, teams, and individuals to collaborate, contributing to the production and consumption of LLM-based services. Employees trusted their organization, e.g., that it will respect agreed-upon contracts, that it will not retaliate when issues are flagged by whistleblowers. In turn, organizations trusted their employees, e.g., not to disclose trade secrets about the LLM components, not to divulgate sensitive information about the capabilities and limitations of the LLM. Employees also trusted each other's work within and across organizations. For instance, three managers mentioned trusting their teams to develop a good LLM model, and team members typically trusted each other. A solution consultant (P23) in the service provider organization explained how interpersonal trust reinforced the customer-base of the provider and interorganizational trust: *"It's important that I always answer the customer truthfully about the LLM, then our customers trust me*

and also [the provider organization]." Organizations further trusted each other, e.g., not to re-use their data for training purposes.

Second, trust relations existed within or across organizations, where **trust activities, expectations, and vulnerabilities revolved around the production of high-performance and potentially trustworthy AI systems** or technical components. Trustors trusted others to develop an appropriate technical artifact, or to conduct activities to produce this artifact responsibly. This could be trust in a rather general trust activity. Eight interview participants from provider organizations, and three from consumer organizations discussed trusting their respective organizations or specific research, data steward, legal, and compliance teams to account for the trustworthiness of LLM-based services. These participants also trusted them to develop responsible AI practices, be it the responsible development, deployment, and release of trustworthy LLMs, and conducting due diligence. For instance, a product manager in a provider organization stated: *"We want trust. And it's not only for the customers, it's also for us internally, to keep working at [provider organization] and develop the LLM model."* Expectations also revolved specifically around making specific decisions in a trustworthy manner. For instance, three product managers discussed trusting other employees for choosing an appropriate LLM to deploy or for collecting truthful information about the LLMs to make such a choice, while one executive discussed trusting their employees enough to delegate the definition of responsible AI principles and their operationalization. Seven employees of consumer organizations explicitly trusted providers to contribute technical components with well-understood and communicated limitations.

Third, our interviews showed that trust can be **bidirectional**. While we primarily discussed downstream actors trusting the artifact providers, upstream providers also trusted downstream consumers to use the providers' artifact responsibly and avoid any harm and reputational risk. Providers expected consumers not to use the service outside the pre-defined scope of applications, and not to game the LLMs, e.g., via adversarial attacks to recover private information. They expected them to put safeguards in place or to provide training for their end-users to avoid misuse. A customer service agent (P23) from a provider organization who supports consumer organizations in implementing the LLM-based service discussed how they have to trust their customers who are involved in early access programs where the service might be more faulty: *"Every customer should in theory start by explaining to its testers all the different limits [of the service], or we should do it. But we don't have time to do this within the testing sessions. So, at some point we trust the customer to do that. That's the main trade-off for early access programs: letting customer use the service without checking that everyone is aware of all limitations."* In turn, the consumer organizations trusted their individual end-users not to misuse or over-use the system in professional contexts, e.g., avoiding over-reliance and eliminating negative biases that could have been generated, to use it ethically, e.g., not infringing artists' rights, and to properly handle problematic outputs of the system. For instance, one deployer organization mentioned trusting its internal end-users to internally report on any toxic or offensive output from the LLM-based system and not to cause a public outcry by discussing the issue on social media.

Finally, **generalized trust in the capabilities of researchers and organizations** to develop the technology sustained junctions. For instance, one quality engineer at a provider organization told us about their belief and trust that the internal research teams could teach them about trustworthy LLMs, and was especially counting on their expertise to discover and communicate *all* potential harms of the systems for the quality engineering team to further assess. Furthermore, employees of provider and consumer organizations expressed the need for international public institutions to consider potential job displacement issues that LLM-based services could cause. There, they often displayed trust in the future and institutional trust in the establishment of such a hypothetical organization to regulate this problem.

4.2.2 Trust factors adapted to the nature of the trustee. Trust factors relevant to individual and organizational trustees were different than those related to technical artifacts

*The ABI properties of **organizations**.* Organizational trustees put forward cues related to their ability, integrity, and benevolence, particularly towards their customers, i.e., for consumption junctions, and put efforts into developing their actual trustworthiness in that regard. For instance, a solution success consultant (P35) at a provider organization, discussed their organization releasing information to their consumers, which allowed the organization to display its ability: *"If we had not been able to explain our model, that would have greatly hurt our ability to convince [consumer organization] to use our LLM. [..] Releasing the white paper and showing how we went about developing this model, and what were the different considerations, it really helped articulate some of the benefits [of our new LLM-based service] to our customers."* The hiring of researchers, publications of research papers, and cross-organization collaborations towards the development of LLM models [25] further provided a sense of ability to internal and external trustors. Benevolence and integrity were illustrated by signed contracts and treaties [135], as well as responsible AI principles [35], emerging governance structures, and continuous support provided to another organization (e.g., the provider's solution consultants helping the consumers debug the service). For instance, six participants mentioned that with the rise of LLMs, their organization slowly established review boards and working groups, ethics and sustainability training, data stewardship workflows, and whistle-blowing processes towards handling the risks of LLMs. A customer support agent (P23) discussed their own trust in their employer (provider organization): *"There are so many trustworthy AI efforts, that helps me to have trust in our company. [..] We now have trustworthy AI guidelines. And sometimes, we might hold back on deploying the product because we might have uncovered some concerns. It might feel frustrating short-term, but those conscious decisions helps me to have trust in our company, it shows it's taking the problems seriously."* Marketing campaigns and other release of communication pieces served as cues to communicate these efforts to certain trustors, particularly consumer organizations.

*The ABI properties of **individual** trustees.* Trustors of all junctions discussed trust towards employees, referring to their expertise in building LLMs, their integrity in building the LLMs, and general benevolence. Ability was recognized by the publication of academic

papers in prestigious venues, corporate recognition, and the responsible AI-related communication efforts and actions organized by a few individuals, such as presentations at well-known industrial seminars for user-experience research. For instance, a UX researcher (P25) at a provider organization discussed their admiration for an AI governance employee *"I often partner with [designer] who is an AI strategist. She opens my eyes [to responsible AI]. She gave an amazing presentation about our internal responsible AI efforts at the last [industry conference on AI]."* In terms of benevolence, participants discussed their own motivation and the dedication their colleagues expressed towards social good and developing ethical systems. Discussions about integrity only revolved around the end-users of LLMs. Organizations employing these users typically trusted users' integrity in terms of not querying LLMs with prompts that would copy the styles of living artists or infringe any other copyright as a content designer in a provider organization (P34) discussed: *"We use MidJourney responsibly. Leadership is pushing us and trusting us to use Firefly a little more because Firefly is from Adobe, and it's been trained on copyrighted images, so it's a little more bulletproof if someone asks questions."*

4.2.3 Need for trust as a cause of miscalibrated (dis)trust. We found that it is not only the trustor who needs to trust the trustee and rely on them, but the trustee sometimes needs the trustor's trust too. The need of one upstream organization to gain the trust of its downstream consumers was expressed by one of its AI governance officers (P8) *"We have to convince our customers that we understand and control generative AI so that they don't go to [competitor organization]. Once we have governance frameworks in place, we can sell trust as a value since its a value that consumers appreciate."* This bidirectional need for trust sometimes played a role in the initial trustor's beliefs in the trustee's integrity. For instance, three participants working within provider and consumer organizations expressed distrust towards their organizations despite these organizations' displayed efforts towards trustworthy AI, arguing about their lack of benevolence, their profit motives, and their inability to serve the good of their end-users or employees. Two other participants also discussed that such a need for trust and other organizational constraints such as the protection of trade secrets could trigger the upstream trustees to foster uncalibrated trust by displaying untruthful or misguiding trustworthiness cues. Yet, the need for trust sometimes became a motivation for trustworthiness. The two provider organizations we interviewed recognized that trustworthiness could become their competitive advantage to develop customer's trust instead of focusing on LLM accuracy, and consequently put more effort into developing trustworthy LLMs. One AI engineer (P15) in one of these organizations illustrated this: *"If you throw solutions which lose customers, perhaps because they can become offensive to end-users, we lose everything. So we are mindful: we want to be responsible as we want trust."*

Finally, we found that trust toward downstream actors could also be miscalibrated within the production junctions because of wrong beliefs of the trustors. The unclear allocation of responsibilities or lack of awareness thereof among the actors producing technical artifacts along the LLM supply chain led actors to wrongly believe that others had conducted certain trustworthy AI tasks, which directly affected the design choices for the next technical artifacts in the chain and hence the trustworthiness of the resulting LLM-based service. For instance, a product manager in one of the provider organizations (P18) acknowledged this problem: *"There's some inherent trust that it is someone else's problem down the line. But it's not always true, so the problem gets kicked down the line, and is never addressed."*

4.3 Theme 3: Interplay Between Multiple Trust Relations, Trust Attitudes And Reliance Behaviors At Supply Chain Junctions

4.3.1 Interplay between trust relations as a new trust and miscalibration factor. We showed in Section 4.2 how interpersonal and interorganizational relations can be direct vectors of reliance in the AI supply chain. We now discuss how they can also constitute an additional trust factor that indirectly affects (miscalibrated) trust towards a technical artifact, which in turn impacts the trustworthiness of LLM-based services according to Section 4.1.

Additional pre-requisites or context for trust in technical artifacts. Twelve participants mentioned trusting a technical artifact only if they can trust its technical properties *and* the organizations involved with it. For instance, a solution success consultant (P27) in a provider organization described one such example of composite relation that they heard employees of consumer organizations discussing: *"Because the employees [of a consumer organization] trust the [provider organization], they trust its implementation of AI. And they also trust that [consumer organization] will responsibly use the AI system. [..] This way, they trust the AI system, its deployment, and also use it.* For an AI developer who worked in a consumer organization in the past (P7), their trust was built around the ability of the service and the trustworthiness of the workflows the provider organization puts in place around the service *"Trust is first of all about the output of the AI. [..] And then, it's about accountability: if something happens and I know who is in charge to handle the issue, then I can feel better in using the AI."*

Softer than a prerequisite, we also found that the belief a trustor had about an adjacent trust relation influenced their trust in the technical artifact. For instance, five product managers in consumer organizations which rely on a specific LLM-based service mentioned their awareness of another consumer organization trusting this service or the service provider, or using any LLM-based service, and because of that, they were comforted in their choice of trusting the service at hand. These beliefs were sometimes chained. For instance, one of these product managers mentioned trusting their (consumer) organization to conduct due diligence about the LLM model, and believing that their organization trusted the LLM provider to build a trustworthy service.

In this context, because LLM supply chains were long with more than five organizations and at least twenty individuals involved, the possibility for the trustors to develop a well-informed trust in all required trustees was very thin —a challenge coined the accountability horizon [29, 139]—, reinforcing the potential for miscalibrated trust.

Substitutes for trust in technical artifacts. Interpersonal or interorganizational trust relations sometimes directly replaced the ABI properties of the technical artifact and led to trusting it, particularly

the benevolence of the LLM-based service was conflated with the benevolence of the provider organizations towards the consumer (similarly to [66, 80, 132]). By believing in the trustworthiness of an individual or organizational trustee involved in the production or the technical artifact, trustors automatically trusted this technical artifact without explicitly reflecting on its ability, benevolence, or integrity. For instance, the solution consultant for one of the provider organization (P27) discussed blind trust that consumers have in providers thanks to their continuous relationship *"Trust in the company itself is a big part of our successful adoption rate [instead of trust in our AI systems]. A different company with a different history might not receive that same level of trust."* And an individual end-user (P59) mentioned personally knowing another end-user who liked the service and decided to trust the service: *"It was word of mouth. I'm a forward-starter trying to get ahead of the curve and figure out how to make my business run more efficiently. Family and friends were like 'you should try AI', and I was nervous and skeptical. But it's become a valuable tool."*

Prior trust relations toward another technical artifact or its provider also sometimes functioned as substitutes for current trust relations. For instance, a solution engineer (P40) working at a consumer organization to set up systems provided by upstream organizations pointed to the substitution of current trust relations with past relations dating from when LLMs were not yet used by the provider: *"If we didn't have a great relationship [with you, the upstream provider] for a long time using your software, then you coming out with something as game-changing as your version of LLMs, I wouldn't feel comfortable with it. That's not an easy decision. So the trust that we have built is very important."*

Among such trust substitutes, unreliable ones could constitute a source of trust miscalibration. For instance, a product manager (P48) at a deployer organization reflected on the blind trust they have in the provider organization and consequently in the systems it produces: *"That's [long lasting relation] one of the things that establishes [provider organization] as an authority. Engaging us in things like this [work sessions organized by the provider to understand the needs of the deployer] helps establish [provider] as an authority. Maybe it's a bad thing, it's kind of that segregation of duties: we've built this trust, and therefore we're going to trust you. Maybe it should be a 'trust but verify' situation."*

4.3.2 The non-linear relation between trust and reliance. Beyond causes of *miscalibrated* (dis)trust (Sections 4.1.3, 4.2.3), our analysis of the supply chain complexity also shows the existence of factors that render trust at certain junctions *impossible*, and that instead foster continuous distrust. The subjectivity of trust expectations and vulnerabilities and of the perception of trust cues, led to disagreement among the many actors of the supply chain, and such disagreement hindered their trust. For instance, an AI developer in a provider organization (P40) pointed to disparate priorities for organizational benevolence and integrity, with certain employees and end-users wishing for transparency about the LLM internally and vis-a-vis the customers, while other teams within the organization pushed against transparency to preserve trade secrets: *"Everyone's gonna want something different. It is gonna come to trust. And if I can't trust it, I'm not going to use it."* These misalignments sometimes created explicit tensions and distrust, for instance one UX developer described from their personal experience that the work of their UX research team revolved around the development of safe interactions between the consumer and the LLM, while engineering teams were more detached from such considerations, and did not perceive the implementation of these interactions as urgent. Furthermore, impossibility to trust also came from fundamentally different conceptions of AI production. The tension between business pressures for deploying a service and the technical limitations for making this service more trustworthy within the business constraints made certain trustors reluctant to trust. For instance, six participants acknowledged other employees' benevolence but discussed the practical impossibility of developing meaningful governance processes considering the race toward LLM development. One participant discussed the role of data stewards and questioned whether they could comprehensively fulfill their duty facing the rapidity with which they had to assess datasets.

Finally, we found that despite the impossibility of trust at certain junctions, reliance on a technical artifact, individual, or organization could still be involved. In these cases, reliance stemmed from competing interests and absence of alternative for the trustor. For instance, for individual employees, thirteen employees at provider organizations shared their distrust in their employer because of the lack of ethics within LLM-based services, yet they still worked for their employer to keep their job. In a sense, the lack of trust in a trustee to fulfill a certain expectation was substituted by trust towards a different trustee and expectation. For instance, at the level of organizations, six employees mentioned that the service provider might release an LLM-based service because they trust the potential profit resulting from releasing it more than its trustworthiness. An AI governance officer in such a provider organization (P8) illustrated this: *"To really trust it, I would like my organization to do more about AI trustworthiness than just meeting the legal bar. But I understand they also have to ship things fast to convince customers to stay with us and our generative AI offering: it's survival for the company."*

5 Discussion, Implications, and Future Directions

Ultimately, our research aims to understand how stakeholders can be supported in AI supply chains that can lead to trustworthy AI. Thus, we investigated how different actors might build (calibrated) trust in the technical artifacts contributing to the AI system, and how such trust shapes various decisions. Our findings corroborate that studying trust all along the AI supply chain is meaningful and urgent, and provide guidance for fostering calibrated trust in the future. They enrich the insights from prior works on trust in human-AI collaboration by expanding to trust concerning other actors (see Table 4).

5.1 Supporting Trust Calibration In Technical Artifacts By Considering The Entire LLM Supply Chain

5.1.1 Embracing the complexity of the LLM supply chain. Our findings present three key insights to foster actors' calibrated trust in the technical artifacts that constitute an LLM-based service.

Topic	Prior work	Key insights from our work	Detailed results from our study
Types of trustors, trustees, and vulnerabilities	Importance of calibrated trust between an end-user or decision-subject and the AI system [44, 66].	Diverse types of trust relations power the supply chain and impact how trustworthy the AI system is.	•Trust in technical artifacts and their intricate boundaries •Trust in individual and organizational actors that either directly relates to AI or not •Generalized trust in technology, research and progress, regulatory organizations
Factors impacting trust	Ability and process integrity of the AI system, intention benevolence of its developers [80].	A multitude of factors relevant to artifacts, individuals, or organizations can lead to (dis)trust.	•Characteristics of technical artifacts •ABI characteristics of individuals and organizations (e.g., communication and publication releases) •Interdependence between trust relations
Trust as attitude or behavior	Trust shapes (over-)reliance [87].	The supply chain junctions present a complex interplay between trust attitudes and reliance behaviors.	•Trust mis-calibration due to numerous possibilities for misinterpretation along the supply chain •Impossible trust from accountability horizon and organizational concerns •Reliance at certain junctions without trust because of organizational and personal incentives and constraints

Table 4: Summary of the insights from our study, structured following the main topics discussed in existing literature on trust in human-AI collaboration (see Section 2).

- Many supply chain junctions were surfaced where trust is involved. This included the production or consumption of technical artifacts, which engage a multitude of trustors. To the best of our knowledge, these junctions had not been highlighted in prior works focused primarily on individual trustors at the end of the AI supply chain and AI algorithms as trustees [73, 126, 133]. However, that the LLM supply chain is sustained due to a diversity of trust relations—where trustors can be organizations, teams, and individuals—is aligned with prior research on trust outside the context of AI, where trustors and trustees were shown to be varied [76, 120].
- Our findings also point out to the complexity of the trustee, and particularly to the multitude of mental models trustors have of "the AI system", the different artifacts of this system in which they might place their trust, and the different expectations and vulnerabilities they can associate to these artifacts. This complements the few prior works that present the trusted AI system as a combination of the core algorithm and additional components (such as user-interfaces) [5, 125].
- Our findings shed light on the diversity and complexity of the factors that impact the extent of trust and trust calibration in these technical artifacts. Beyond trustors' potentially low AI literacy [42] leading them to misinterpret trustworthiness cues [10, 20, 26, 80], our findings particularly point out the historical, generalized, interpersonal and inter-organizational trust relations that impact each other in various ways (e.g., influence or substitution), which has found limited discussion [5] in the context of AI, despite being debated in organizational psychology [120].

These insights play an important role in characterizing the trust dynamics that traverse the LLM supply chains, and in disentangling how they might impact the resulting LLM systems. By acknowledging their complexity, we make a valuable contribution towards shaping future HCI, policy and organizational efforts towards more trustworthy AI [29].

5.1.2 Fostering calibrated trust in technical artifacts. Our insights invite the reconsideration of current approaches for fostering calibrated trust towards technical artifacts. Trustors should be facilitated access to relevant and meaningful trustworthiness cues we identified, while avoiding miscalibration as a result of misinterpretations of these cues.

To share trustworthiness cues, prior research on AI transparency and documentation [46, 89] could be revisited to incorporate the breadth of trustworthiness factors relevant to the different trust relations that impact trust towards the technical artifact, instead of solely focusing on the ability of the LLM. This can include information about the ability, benevolence, and integrity of every technical artifact [13, 80], and fostering social transparency [41] by including information about the individuals and organizations involved in each junction of the chain. Personalizing such documentation to the different trustors (acknowledging their differences, e.g., in terms of trust expectations) could help tackle typical challenges of cognitive overload, privacy infringements, and the potential to reveal trade-secrets.

Since trustors are prone to misinterpret trust relations and trust cues due to the complexity of the supply chain, there is a need to help them reflect [12, 84, 108] on their own knowledge of the supply chain and on the way trust relations impact their judgments of the technical artifact. Directing research efforts to help practitioners develop more accurate mental models of the AI supply chain, its materiality, its capabilities, and its limitations, could also be beneficial to avoid the misinterpretations, particularly as AI imagineries [61, 91] of actors within the AI supply chain remain under-studied in contrast to imagineries of the public [56, 82, 112]. Leveraging and adapting AI literacy frameworks [81, 94] and training programs on trustworthy AI [1, 21, 24, 114] to account for the properties of the AI supply chain currently left out (e.g., the dynamicity, the involvement of many stakeholders), and adapting strategies for appropriate trust established by organizational psychologists [7] (such as *sense-making*, i.e., a collective learning process, and *transference*, i.e., exploiting the transferability of trust between actors) can be a way forward.

5.1.3 Revising the scope and methods of trust research. These insights suggest reconsidering the design of studies on human-AI trust, and especially invite us both to expand the scope of studies and to revise current research methods. Future studies should investigate the junctions of the supply chain that are upstream the end-users of the final AI system and the factors that foster (miscalibrated) trust and reliance, by further delving into the individual trustors we identified. We envision a mixed-methods approach to be necessary. Empirical qualitative methods can enable us to remain grounded in the realities of the study participants as illustrated by the value of our findings. Quantitative studies could build over the results of the formative qualitative studies to disentangle the significance of potential factors identified. Such studies would require realistically accounting for the entangled factors and trust relations that impact trust towards a technical artifact (e.g., incorporating considerations about the organizations and individuals involved in the supply chain), and unambiguously framing the conceptual boundaries of the artifact of interest within the study, moving beyond existing practices.

5.2 Supporting Calibrated Trust In Technical Artifacts By Explicitly Accounting For Interpersonal And Organizational Trust

5.2.1 Acknowledging the prevalence of diverse trust relations along the AI supply chain. Our study has shown that LLM supply chains are sustained not only by trust towards technical artifacts, but also by diverse trust relations where trustors and trustees can be organizations, teams, individuals, or institutions. To the best of our knowledge, our study is the first to shed light on these interpersonal, intra- and inter-organizational trust relations and dynamics all *along* the LLM supply chain. Only a few empirical studies have previously hinted at the importance of organizational context in the appreciation of an AI system for end-users' trust, and this context was limited to the integrity of the individual developers and involved organization [18, 41, 66, 132]. Our findings extend these studies by showing that interpersonal or interorganizational trust might substitute or at least influence *any trustor*'s trust. These relations can equally be prone to miscalibration and can constitute sources of miscalibrated distrust. In order to comprehensively cater to existing trust relations along the LLM supply chain, our results point out the factors impacting such trust relations. These factors are aligned with those discussed in prior research —including a trustee's ability, benevolence, and integrity [86], how these are communicated through cues [80], and contextual factors [48, 72, 100, 122, 132].

5.2.2 Establishing support structures for fostering organizational trust. Our findings corroborate prior work by Jiang and Luo [65] showing that trust within organizations is highly concomitant with the existence of organizational efforts that promote integrity, benevolence, and ability. Many participants in our study discussed trust in organizations based on the efforts for AI trustworthiness they were aware of. This is consistent with existing organizational efforts for AI trustworthiness (e.g., AI principles [59, 90, 138], toolkits [14, 77], alliances[4], and pacts[5]). Since these efforts become vectors for trust, we should explicitly question how to materialize them to ensure their efficiency in building well-informed trust, and how to make them transparent to the employees. To foster employees' trust, organizations can develop safe procedures for reporting on AI systems [60] as some interview participants showed wariness towards the intention behind internal governance efforts (aligned with debates around *ethics washing* [130]) and towards their own security when making use of existing procedures. E.g., they wondered about reporting confidentiality and potential retaliation against whistleblowers. To foster appropriate trust from external employees and organizations, organizations should also explicitly reflect on both the ways in which they communicate their trustworthiness and how they conduct due diligence to assess others' trustworthiness. Researchers could further incentivize organizations in building appropriate trust cues and ensuring calibrated trust among their employees by incorporating considerations around trust dynamics into frameworks that map the maturity levels of organizations in terms of responsible AI [107].

To deepen our understanding of such relations, researchers could investigate whether the inter-organizational trust development factors affecting non-AI supply chains [19, 38, 45, 65, 69, 104, 120, 131, 143, 146] are relevant in the context of AI, e.g., the voluntary position of vulnerability, organizational atmosphere, and cultural background or geographic location of the trustor and trustee [38, 143]. Knowledge of these factors could provide more clues to develop well-calibrated trust in the future.

5.2.3 Accounting for the necessity of diverse trust relations at certain junctions of the supply chain. Many trust relations substitute a trustor's knowledge about the trustworthiness of the LLM-based service. LLM supply chains are the culmination of the trend that makes software supply chains increasingly complex [29, 54], in terms of the number of actors and technical artifacts involved [29, 139] and of the number of iterative development workflows [23, 105]. This complexity worsens the accountability horizon [29] and hinders access to knowledge of the AI supply chain. To avoid miscalibration and acknowledge the necessity of interpersonal and interorganizational relations to circumvent practical challenges in assessing the trustworthiness of a technical artifact, we argue that it is pragmatic to differentiate the types of trust attitudes to foster across junctions and actors. For instance, fostering informed and reflexive trust among trustors and trustees who are close in the LLM supply chain, and "blind" but calibrated trust among more distant actors would ensure accounting for the complexities of the AI supply chain we identified. This way, a consumer organization using the LLM-powered application of a deployer organization would develop well-informed and calibrated trust in terms of the LLM trustworthiness, but "blind" trust toward the upstream organizations that might have collected training datasets, by trusting that the deployer organization has a well-calibrated trust in these upstream organizations.

By replicating prior studies at the different junctions of the LLM supply chain, researchers could identify the trust relations that currently play a larger role in trustors' decisions at the junctions. For

[4] https://thealliance.ai/
[5] https://digital-strategy.ec.europa.eu/en/policies/ai-pact

instance, Qi et al. [104] pointed the role of interpersonal trust over inter-organizational trust in knowledge exchanges across organizations, but conditioned it on the nature of the collaboration needed (e.g., a 'simple' transaction cost, or collaborative work); and Bruneel et al. [19] found that the level of inter-organizational trust decreases with technological complexity due to the need for protecting one's technological knowledge vis-a-vis the trustee. Whether this is the case within AI supply chains is not yet known, yet it would inform us on which trust relations to investigate in priority.

5.3 Reflecting On Calibrated Trust Attitudes And Reconsidering Reliance Behaviors

5.3.1 Acknowledging miscalibrated trust and distrust. We found that supply chain junctions might be traversed by miscalibrated trust, stemming from trustors' misinterpretations exacerbated by the supply chain complexities. Such complexities might also induce a lack of trust, for instance due to disagreements among the actors of the supply chain about AI values [67, 88]. Furthermore, LLM-based services are shrouded in uncertainties due to their novelty and fast-paced innovations in the field [58, 95, 140, 141]. Thus, knowledge used traditionally as a vehicle for software trustworthiness (e.g., audits for software quality, and agreement on software requirements) is not available for LLM supply chains [3, 50]. Such uncertainties are an additional cause of miscalibration or distrust. Despite distrust, we still found reliance due to competing incentives, which emphasizes that trust does not necessarily correlate with reliance behaviors [124]. This is particularly important as AI trust literature has used reliance as a proxy for trust [87], while organizational psychologists distinguish between trust as an attitude and reliance as a behavior [76].

5.3.2 Tackling the impossibility of trust. These insights invite us to explore the factors beyond trust that impact the production and adoption of LLMs via reliance on various actors and artifacts, and encourage methodological caution to explicitly identify when trust and reliance can be used interchangeably. In doing so, particular attention should be brought to the object of trust and how it affects behaviors, as we found that trust varies depending on the activity and expectation at hand. E.g., employees trusted their organization to develop good LLMs, while distrusting it about job displacement. These insights also suggest adapting existing documentation tools to integrate the mechanisms behind reliance decisions to facilitate accountability in the supply chain.

As for circumventing the risks that reliance under miscalibration and distrust bring, organizations could focus on (i) promoting transparency and enhanced accountability for contentious decisions [29, 106], and (ii) ensuring the upholding of clear distributed responsibilities [99, 139]. Prior work shows that trust can be complemented or substituted by forms of governance such as contracts [127]. To this end, shifting the expectations a trustor has of a trustee could be effective. For instance, if it is not possible to comprehensively test for the integrity of an LLM, processes could be in place to continuously test (and correct) the LLM -based service after deployment. The responsibility for this could be allocated to the LLM producer. The consumer would then not necessarily trust the LLM's performance or the capability of its producer to produce a "good" system but instead the producer's responsibility after deployment.

5.3.3 Revising obligations to counterbalance power dynamics hindering calibrated trust and reliance. Recognizing that AI system development is evolving as a supply chain, researchers [27, 29, 70] have called for investigating the dynamics of these AI supply chains, their power relations, and their impact on accountability. Outside the technological context, organizational psychology and information systems researchers have shown the complex interplay between interpersonal and organizational trust and power [6, 45, 74, 131]. For instance, trustors need to place trust in a trustee especially when they do not have *control* over the trustee's activity [45]; in turn, the *power of the trustee* (and trustor's power in cases of bi-directional trust) is reinforced with the trustor's trust [6]; yet the power of a trustee might also hinder trust in them [131]. While our participants did not directly talk about power, our findings strongly resonate with these ideas and hint at some ways in which existing sources of power impact trust dynamics along the AI supply chain, ultimately impacting trustworthy AI.

Some trust relations towards upstream actors and trust cues put forward by upstream organizations (e.g., scientific publications or public responsible AI guidelines) have been related to the power these organizations have to impact downstream stakeholders of the AI supply chain. The economic and political power of certain upstream organizations enables them to ultimately impact the cultural image of AI and of themselves [82, 97, 129], for instance, by influencing the direction of technological development [83, 96, 119, 137], the definition of AI transparency [136, 144], or the establishment of other governance processes [22, 70]. Their power also enables them to display potentially untruthful trustworthiness cues we identified: for instance it is now well-understood that transparency displays can act as a cue of integrity while the information being displayed is carefully selected to provide a skewed impression of the system's ability [16, 26, 31]. Note that in the cases we identified where an upstream actor was the trustor toward a downstream trustee, this trustor again exerted power over the trustee by allocating responsibility onto trustees. This was for instance the case when producer or consumer organizations respectively expected and trusted the consumer organization or the end-users not to over-rely on the LLM. This made apparent the regulatory power these organizations have, accompanied by regulatory power of institutions that participants hinted at in the shape of generalized trust in institutions. Second, the situations where trust was absent but reliance present because of the absence of alternatives can be associated with both market power—the AI arms race inciting organizations to deploy or adopt AI systems without trusting them [47]—, and the power of the actor that is relied on—economic and infrastructural power [83, 128] or the political control they can exercise, e.g., on the acceptable biases in the LLM [28, 79].

These insights re-emphasize the complexity of addressing miscalibrated trust, impossibility of trust, and over-reliance in the supply chain. Doing so would require interrogating how trust and broader power dynamics affect whose voices are heard within the chain, normatively deciding those to be prioritized [123], and accounting for and potentially counter-balancing the power sources. We encourage researchers to explicitly study these dynamics when catering to AI trustworthiness. We also suggest policymakers to build additional incentives for upstream trustees [68], for instance

by enforcing audits of the organizations and of their trust cues [32, 43].

5.4 Caveats, Limitations, and Methodological Considerations

Our study represents one of the first efforts to explore trust along the AI supply chain. Future work should explore the extent to which the trust factors that we surfaced are strong predictors for (mis)calibrated (dis)trust and (over) reliance. It is worth noting that we only focused on a few LLM supply chains related to a small set of sectors—only private institutions in highly-regulated sectors. We did not account for all stakeholders in these supply chains (e.g., organizations' accountants, contract managers, and developers of foundation models are missing). Despite our exploratory recruitment process, every supply chain is different and it is possible that additional actors could be relevant. We acknowledge the struggles stemming from conducting empirical work within organizations of the AI supply chain, and the challenges in publishing all interview data [113]. Finally, our choice of conceptual boundary for AI-based services (i.e., the supply chain) merits further exploration. This choice affected our insights — the multitude of factors we identified is explained at least partially by the topical shift we operated, going from a technology-centric view on AI [20, 57] to seeing AI as part of a system. Instead, an AI system could be considered to be a fixed or dynamic ensemble of technical artifacts that drive an AI service, or the object of study could be a supply chain centered around a fixed provider or consumer organization. Since these alternatives all bear implications in the study of trust and in the future tools, policies, and regulations that would ensue, we acknowledge our choice, and hope that future studies do so too.

6 Conclusion

By regulating the capability of AI systems and consequently mitigating their potential for harm, policymakers have focused on enhancing the trustworthiness of AI systems (e.g., the AI Act [85]) and hence increasing public trust. Simultaneously, HCI researchers have investigated how to foster AI adoption and calibrated trust among end-users or decision-subjects. In this work, we questioned this assumed relation between trust in AI, the trustworthiness (ability) of AI systems, and their adoption. Inspired by prior works discussing AI supply chains [29, 139] and using LLMs as a use-case, we opened up the scope of research on trust. We confirmed that trust not only plays a role in adoption decisions but also powers junctions of the supply chain that lead to developing the AI system itself. We found that trust is developed via many interactions between the ability, benevolence, and integrity properties of artifacts, individuals, and organizations in the supply chain. Finally, we revealed the ambivalence of trust along the supply chain— sometimes necessary but prone to miscalibration, and sometimes unachievable due to disagreements or a lack of knowledge, or some other times dispensable because of power dynamics.

Our findings call for HCI researchers and policymakers to account for the complexities of AI supply chains, to continuously probe and evaluate the attitudes and behaviors along the chains' junctions, and to normatively disambiguate the types of trust and reliance we want to foster. These findings also invite researchers to revise tools and workflows to facilitate responsible decision-making and reporting by the supply chain actors, and to support them in reflecting on the challenges brought by supply chain notions of trust. Finally, policymakers and organizational governance boards are encouraged to interrogate the dynamics they sustain in the AI supply chains and to develop structures that would drive more controlled adaptations to AI innovations.

References

[1] Andrea Aler Tubella, Marçal Mora-Cantallops, and Juan Carlos Nieves. 2024. How to teach responsible AI in Higher Education: challenges and opportunities. *Ethics and Information Technology* 26, 1 (2024), 3.

[2] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & society* 35 (2020), 611–623.

[3] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. 2014. Provable bounds for learning some deep representations. In *International conference on machine learning*. PMLR, 584–592.

[4] Maryam Ashoori and Justin D Weisz. 2019. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint arXiv:1912.02675* (2019).

[5] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2022. A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction* (2022), 1–16.

[6] Reinhard Bachmann. 2001. Trust, power and control in trans-organizational relations. *Organization studies* 22, 2 (2001), 337–365.

[7] Reinhard Bachmann, Nicole Gillespie, and Richard Priem. 2015. Repairing trust in organizations and institutions: Toward a conceptual framework. *Organization Studies* 36, 9 (2015), 1123–1142.

[8] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7, 1 (2023), 52–62.

[9] Agathe Balayn and Seda Gürses. 2021. Beyond Debiasing: Regulating AI and its inequalities. *EDRi report* (2021).

[10] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–17.

[11] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[12] Margaret Bearman and Rola Ajjawi. 2023. Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology* 54, 5 (2023), 1160–1173.

[13] Patrick Bedué and Albrecht Fritzsche. 2022. Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management* 35, 2 (2022), 530–549.

[14] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[15] Michaela Benk, Suzanne Tolmeijer, Florian von Wangenheim, and Andrea Ferrario. 2022. The value of measuring trust in AI-a socio-technical system perspective. *arXiv preprint arXiv:2204.13480* (2022).

[16] Clare Birchall. 2011. Introduction to 'Secrecy and Transparency' The Politics of Opacity and Openness. *Theory, Culture & Society* 28, 7-8 (2011), 7–25.

[17] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[18] Jacob T Browne, Saskia Bakker, Bin Yu, Peter Lloyd, and Somaya Ben Allouch. 2022. Trust in clinical AI: Expanding the unit of analysis. In *1st International Conference on Hybrid Human-Artificial Intelligence: HHAI2022.*

[19] Johan Bruneel, André Spithoven, and Bart Clarysse. 2017. Interorganizational trust and technology complexity: Evidence for new technology-based firms. *Journal of Small Business Management* 55 (2017), 256–274.

[20] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[21] Roger Burkhardt, Nicolas Hohn, and Chris Wigley. 2019. Leading your organization to responsible AI. *McKinsey Analytics* (2019), 1–8.

[22] Corinne Cath. 2018. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180080.

[23] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? arXiv:2307.09009 [cs.CL]

[24] Zhisheng Chen. 2024. Responsible AI in Organizational Training: Applications, Implications, and Recommendations for Future Development. *Human Resource Development Review* (2024), 15344843241273316.

[25] Yoana Cholteeva. 2023. IBM and Meta launch AI alliance with over 50 collaborators. https://erp.today/ibm-and-meta-launch-ai-alliance-with-over-50-collaborators/

[26] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems.. In *IUI workshops*, Vol. 2327.

[27] Jennifer Cobbe. 2024. Governance and interdependence in data-driven supply chains. *Global Governance by Data: Infrastructures of Algorithmic Rule* (2024).

[28] Jennifer Cobbe and Jatinder Singh. 2021. Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges. *Computer Law & Security Review* 42 (2021), 105573.

[29] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1186–1197.

[30] EU COM. 2021. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Proposal for a regulation of the European parliament and of the council.

[31] Eric Corbett and Emily Denton. 2023. Interrogating the T in FAccT. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1624–1634.

[32] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.

[33] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

[34] Kate Crawford and Vladan Joler. 2018. Anatomy of an AI System. *Anatomy of an AI System* (2018).

[35] Paul B de Laat. 2021. Companies committed to responsible AI: From principles towards implementation and regulation? *Philosophy & technology* 34 (2021), 1135–1193.

[36] Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who are the Stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 227–236.

[37] Shipi Dhanorkar, Christine T Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. 1591–1602.

[38] Mark Dodgson. 1993. Learning, trust, and technological collaboration. *Human relations* 46, 1 (1993), 77–95.

[39] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 297–307.

[40] Lilian Edwards. 2022. Regulating AI in Europe: four problems and four solutions. *Ada Lovelace Institute* 15 (2022), 2022.

[41] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[42] Upol Ehsan and Mark O Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5, 6 (2024).

[43] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence* 3, 7 (2021), 566–571.

[44] Andrea Ferrario and Michele Loi. 2022. How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1457–1466.

[45] Michael J Gallivan and Gordon Depledge. 2003. Trust, control and the role of interorganizational systems in electronic partnerships. *Information Systems Journal* 13, 2 (2003), 159–190.

[46] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[47] Edward Moore Geist. 2016. It's already too late to stop the AI arms race—We must manage it instead. *Bulletin of the Atomic Scientists* 72, 5 (2016), 318–321.

[48] Anupam Ghosh and Jane Fedorowicz. 2008. The role of trust in supply chain governance. *Business Process Management Journal* 14, 4 (2008), 453–470.

[49] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[50] Ian Goodfellow and Nicolas Papernot. 2017. The challenge of verification and testing of machine learning. *Cleverhans-blog* (2017).

[51] Michael Grynbaum and Ryan Mac. 2023. The Times sues OpenAI and Microsoft over A.I. use of copyrighted work. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

[52] H Güngör. 2020. Creating value with artificial intelligence: A multi-stakeholder perspective. *Journal of Creating Value* 6, 1 (2020), 72–85.

[53] Kholekile L Gwebu, Jing Wang, and Marvin D Troutt. 2007. A conceptual framework for understanding trust building and maintenance in virtual organizations. *Journal of Information Technology Theory and Application (JITTA)* 9, 1 (2007), 5.

[54] Seda Gürses and Joris van Hoboken. 2018. *Privacy after the Agile Turn*. Cambridge University Press, 579–601.

[55] Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1112–1123.

[56] Sne Scott Hansen. 2022. Public AI imaginaries: How the debate on artificial intelligence was covered in Danish newspapers and magazines 1956–2021. *Nordicom Review* 43, 1 (2022), 56–78.

[57] Gaole He and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making. In *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22)*.

[58] Hans-Martin Heyn, Eric Knauss, Amna Pir Muhammad, Olof Eriksson, Jennifer Linder, Padmini Subbiah, Shameer Kumar Pradhan, and Sagar Tungal. 2021. Requirement engineering challenges for ai-intense systems development. In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*. IEEE, 89–96.

[59] Merve Hickok. 2021. Lessons learned from AI ethics principles for future actions. *AI and Ethics* 1, 1 (2021), 41–47.

[60] Alexander Hicks. 2022. Transparency, compliance, and contestability when code is (n't) law. In *Proceedings of the 2022 New Security Paradigms Workshop*. 130–142.

[61] Michael Hockenhull and Marisa Leavitt Cohn. 2021. Hot air and corporate sociotechnical imaginaries: Performing and translating digital futures in the Danish tech scene. *New Media & Society* 23, 2 (2021), 302–321.

[62] Robert R Hoffman, Gary Klein, Shane T Mueller, Mohammadreza Jalaeian, and Connor Tate. 2021. The Stakeholder Playbook for Explaining AI Systems. *PsyArXiv Preprints* (2021).

[63] Alex Ingrams, Wesley Kaufmann, and Daan Jacobs. 2022. In AI we trust? Citizen perceptions of AI in government decision making. *Policy & Internet* 14, 2 (2022), 390–409.

[64] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[65] Hua Jiang and Yi Luo. 2018. Crafting employee trust: from authenticity, transparency to engagement. *Journal of Communication Management* 22, 2 (2018), 138–160.

[66] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 77–88.

[67] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453* (2023).

[68] Bran Knowles and John T Richards. 2021. The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 262–271.

[69] Frens Kroeger. 2012. Trusting organizations: The institutionalization of trust in interorganizational relationships. *Organization* 19, 6 (2012), 743–763.

[70] Steffen Krüger and Christopher Wilson. 2023. The problem with trust: on the discursive commodification of trust in AI. *AI & SOCIETY* 38, 4 (2023), 1753–1761.

[71] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. POTs: protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 177–188.

[72] Ik-Whan G Kwon and Taewon Suh. 2004. Factors affecting the level of trust and commitment in supply chain relationships. *Journal of supply chain management* 40, 1 (2004), 4–14.

[73] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[74] Luke A Langlinais, Heath A Howard, and Jeffery D Houghton. 2022. Trust me: Interpersonal communication dominance as a tool for influencing interpersonal trust between coworkers. *International Journal of Business Communication*

(2022), 23294884221080933.
[75] Johann Laux. 2023. Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI & SOCIETY* (2023), 1–14.
[76] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[77] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.
[78] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (2023), 107539.
[79] Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
[80] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1257–1268.
[81] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.
[82] Inga Luchs, Clemens Apprich, and Marcel Broersma. 2023. Learning machine learning: On the political economy of big tech's online AI courses. *Big Data & Society* 10, 1 (2023), 20539517231153806.
[83] Dieuwertje Luitse. 2024. Platform power in AI: The evolution of cloud infrastructures in the political economy of artificial intelligence. *Internet Policy Review* 13, 2 (2024), 1–44.
[84] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.
[85] Tambiama Madiega. 2021. Artificial intelligence act. *European Parliament: European Parliamentary Research Service* (2021).
[86] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.
[87] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. 2023. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction. *arXiv preprint arXiv:2311.06305* (2023).
[88] Siddharth Mehrotra, Catholijn M Jonker, and Myrthe L Tielman. 2021. More similar values, more trust?-the effect of value similarity on trust in human-agent interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 777–783.
[89] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
[90] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature machine intelligence* 1, 11 (2019), 501–507.
[91] Jakub Mlynar, Farzaneh Bahrami, André Ourednik, Nico Mutzner, Himanshu Verma, and Hamed Alavi. 2022. AI beyond deus ex machina–Reimagining intelligence in future cities with urban experts. In *Proceedings of the 2022 Chi conference on human factors in computing systems*. 1–13.
[92] Debmalya Mukherjee, Robert W Renn, Ben L Kedia, and Deepraj Mukherjee. 2012. Development of interorganizational trust in virtual organizations: An integrative framework. *European Business Review* 24, 3 (2012), 255–271.
[93] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
[94] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2 (2021), 100041.
[95] Maria Nordström. 2022. AI under great uncertainty: implications and decision strategies for public policy. *AI & society* 37, 4 (2022), 1703–1714.
[96] Helga Nowotny. 2021. *In AI we trust: Power, illusion and control of predictive algorithms*. John Wiley & Sons.
[97] Rodrigo Ochigame. 2019. The invention of 'ethical AI': How big tech manipulates academia to avoid regulation. *Economies of virtue* 49 (2019).
[98] Nessrine Omrani, Giorgia Rivieccio, Ugo Fiore, Francesco Schiavone, and Sergio Garcia Agreda. 2022. To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change* 181 (2022), 121763.
[99] Will Orr and Jenny L Davis. 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society* 23, 5 (2020), 719–735.
[100] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.
[101] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.
[102] Michael Pirson. 2009. Facing the trust gap-measuring and managing stakeholder trust. *DOING WELL AND GOOD: THE HUMAN FACE OF THE NEW CAPITALISM, Julian Friedland, ed., Information Age Publishing* (2009).
[103] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (2018).
[104] Cong Qi and Patrick YK Chau. 2013. Investigating the roles of interpersonal and interorganizational trust in IT outsourcing success. *Information Technology & People* 26, 2 (2013), 120–145.
[105] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv:2310.03693 [cs.CL]
[106] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.
[107] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
[108] Yoram Reich. 2017. The principle of reflexive practice. *Design Science* 3 (2017), e4.
[109] Karoline Reinhardt. 2023. Trust and trustworthiness in AI ethics. *AI and Ethics* 3, 3 (2023), 735–744.
[110] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 223–233.
[111] Mohammad Asif Salam. 2017. The mediating role of supply chain collaboration on the relationship between technology, trust and operational performance: An empirical investigation. *Benchmarking: An International Journal* 24, 2 (2017), 298–317.
[112] Laura Sartori and Giulia Bocca. 2023. Minding the gap (s): public perceptions of AI and socio-technical imaginaries. *AI & society* 38, 2 (2023), 443–458.
[113] Morgan Klaus Scheuerman. 2024. In the Walled Garden: Challenges and Opportunities for Research on the Practices of the AI Tech Industry. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 456–466.
[114] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2020. Principles to practices for responsible AI: closing the gap. *arXiv preprint arXiv:2006.04707* (2020).
[115] Carolyn B. Seaman. 1999. Qualitative methods in empirical studies of software engineering. *IEEE Transactions on software engineering* 25, 4 (1999), 557–572.
[116] Paul Smart, Brian Pickering, Michael Boniface, and Wendy Hall. 2021. Risk Models of National Identity Systems: A Conceptual Model of Trust and Trustworthiness. (2021).
[117] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv:2306.05949* (2023).
[118] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[119] Aurelia Tamò-Larrieux, Clement Guitton, Simon Mayer, and Christoph Lutz. 2024. Regulating for trust: Can law establish trust in artificial intelligence? *Regulation & Governance* 18, 3 (2024), 780–801.
[120] Hwee Hoon Tan and Augustine KH Lim. 2009. Trust in coworkers and trust in organizations. *the Journal of Psychology* 143, 1 (2009), 45–66.
[121] Ting Fang Tan, Arun James Thirunavukarasu, J Peter Campbell, Pearse A Keane, Louis R Pasquale, Michael D Abramoff, Jayashree Kalpathy-Cramer, Flora Lum, Judy E Kim, Sally L Baxter, et al. 2023. Generative artificial intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. *Ophthalmology Science* 3, 4 (2023), 100394.
[122] Gaurav Tejpal, RK Garg, and Anish Sachdeva. 2013. Trust among supply chain partners: a review. *Measuring Business Excellence* 17, 1 (2013), 51–71.
[123] Petros Terzis. 2023. Law and the political economy of AI production. *International Journal of Law and Information Technology* 31, 4 (2023), 302–330.
[124] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on*

user modeling, adaptation and personalization. 77–87.
[125] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 272–283.
[126] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. 2022. Trust in human-AI interaction: Scoping out models, measures, and methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.
[127] Andrew H Van de Ven and Peter Smith Ring. 2006. Relying on trust in cooperative inter-organizational relationships. *Handbook of trust research* (2006), 144–164.
[128] Fernando van der Vlist, Anne Helmond, and Fabian Ferrari. 2024. Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society* 11, 1 (2024), 20539517241232630.
[129] J Van Dijck. 2018. The platform society: Public values in a connective world.
[130] Gijs van Maanen. 2022. AI ethics, ethics washing, and the need to politicize data ethics. *Digital Society* 1, 2 (2022), 9.
[131] Siv Vangen and Chris Huxham. 2003. Nurturing collaborative relations: Building trust in interorganizational collaboration. *The Journal of applied behavioral science* 39, 1 (2003), 5–31.
[132] Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. 2024. Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
[133] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.
[134] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. 2022. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 763–777.
[135] Kent Walker. 2023. Our commitment to advancing bold and responsible AI, together. https://blog.google/outreach-initiatives/public-policy/our-commitment-to-advancing-bold-and-responsible-ai-together/
[136] Hao Wang. 2022. Transparency as manipulation? Uncovering the disciplinary power of algorithmic transparency. *Philosophy & Technology* 35, 3 (2022), 69.
[137] Meredith Whittaker. 2021. The steep cost of capture. *Interactions* 28, 6 (2021), 50–55.
[138] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 195–200.
[139] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society* 10, 1 (2023), 20539517231177620.
[140] Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. 2023. Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. *International Journal of Human–Computer Interaction* 39, 3 (2023), 494–518.
[141] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
[142] Qian Yang, Richmond Y Wong, Steven Jackson, Sabine Junginger, Margaret D Hagan, Thomas Gilbert, and John Zimmerman. 2024. The Future of HCI-Policy Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.
[143] Akbar Zaheer and Jared D Harris. 2005. Interorganizational trust. *Handbook of Strategic Alliances, Oded Shenkar and Jeffrey J. Reuer, eds* (2005), 169–197.
[144] Monika Zalnieriute. 2021. " Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism. *Critical Analysis L.* 8 (2021), 139.
[145] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).
[146] Weiguo Zhong, Chenting Su, Jisheng Peng, and Zhilin Yang. 2017. Trust in interorganizational relationships: A meta-analytic integration. *Journal of management* 43, 4 (2017), 1050–1075.

A Positionality Statement

We acknowledge our positionality and its potential impact on our study setup and the analysis of the interview transcripts. We, the authors of this manuscript, identify with different genders and hail from different continents. Some of us work in universities while others work in industry. We all have training in computer science and practical experience with machine learning and data science projects. Some of us also have a background in human-computer interaction, have conducted several empirical studies with machine learning practitioners across the world, and bear a strong interest in critical machine learning literature and trust in AI literature. All of us acknowledge the harmful impacts that machine learning, and particularly LLMs, can have, both within their production processes and in interaction with their users. As a result, we are motivated by our desire to thoroughly inspect LLM supply chains as socio-technical practices in order to identify human sources of such harmful impacts and potential solutions. Accounting for our positionality, we did our best to accurately report and fairly account for all opinions of the study participants.

Besides, some of the authors were acquainted with some of the study participants before the study, either because these participants worked in the same organization or related organizations to the ones of these authors. We made sure however to mitigate any potential dynamic that could affect what the participants would explain during the interviews (e.g., by ensuring them that the information would not impact any of the involved organizations).

Due to constraints from some organizations where we recruited our participants (e.g., difficulty in accepting financial compensation for a study where a partner organization is involved), we decided to homogenize our retribution process. For all employees of organizations that contribute to developing, deploying, or consuming LLMs, we did not provide any financial compensation –these employees mentioned participated voluntarily with the intrinsic motivation of reflecting on their own practices during the interview, and learning about others'. Instead, those whose primary position revolves around using an LLM to conduct a task (e.g., customer service agents) were compensated based on the duration of the interview and a rate higher than the minimum rate in their country. Compensating differently the participants is a shortcoming of our study as we cannot know with certainty how it impacted our participants' responses. This does not however seem to bear a significant impact on the responses (e.g., on average, the durations of the interviews with non-/compensated participants were the same) as the compensated participants were not related to any of the interviewers or their organizations.

B Examples of questions used during the interviews

While answering the following questions, our participants either explicitly mentioned trust or trust factors (in which case we prompted them further about it), or they discussed relations between various actors of the LLM supply chains (with potential reliance behaviors and power dynamics) that we also explored further.

Understanding the interview participant's work.

- Could you briefly talk about your role and responsibilities in your current organization?
- To what extent do you think your work is essential for the society at large? Does it impact how meaningful you feel your work is?
- What are the challenges you face when working with AI? What are their causes?

- Are you aware of any effort to prevent such issues? What do you think about them?
- What would you like to see organizations do about AI? -To curb these issues? To mitigate any other AI issue?
- How much feedback do you receive about the impact of your work? How does the feedback you receive about the impact of your work influence the knowledge you have of its results?

Mapping the actors in the supply chain and the perceptions of the interview participants (e.g., related to responsible AI work).

- Who do you work with, and what is the nature of this work?
- Who do you think are the individuals inside or outside your organization that impact AI systems at your organization? It could be a direct or indirect impact.
- How does it work in your organization for an AI system to be accepted, i.e., deployed or adopted? Is there any responsible AI consideration? Are there tests? Or other review processes?
- Can you describe the organizational actors that are related to responsible AI, their role and work, and their impact?
- What challenges do all these actors face when working on or along responsible AI efforts?
- Which actors do you think have the responsibility to work on responsible AI across organizations? In what sense? And for which reasons?
- Do you think any other individual could or should be involved? Why and who?
- Do you think organizations are concerned with responsible AI? Why? Should they be concerned with responsible AI? Why? What should organizations do about responsible AI? What should be their roadmap?
- If you have experience with other organizations, do you think things are similar in the other organizations? Why or why not? And in what sense?
- Do your thoughts on responsible AI match those of other individuals at your company and/or other organizations you work with? What's the organizational culture?

Delving into consumption junctions.

- Do you use AI-related tools in your daily work? Which ones? Why these ones? How / who decided on them?
- What are the main benefits of AI? How useful is it in your work?
- Can you talk about bad experiences you had with AI?
- Are you aware of any effort to prevent such issues? What do you think about them?
- What would you like to see organizations do about AI? -Would you like to get any kind of training or onboarding? any information that you would like to access about AI?

Delving deeper into responsible AI activities.

- Organizational-level considerations:
  - Do you see any tension between responsible AI and the work of any organization? Have you ever encountered or identified trade-offs among responsible AI efforts, or between such efforts and other objectives?
  - What, if any, are some of the challenges you see with responsible AI within your company or the clients you work with?
  - What do you think you and your role can or should bring in comparison to others in the organization?
- Participant-level considerations:
  - How difficult is the responsible AI work, and why?
  - What are the main hurdles you encounter in your daily work related to responsible AI?
  - Can you describe challenging situations you've faced with responsible AI? What happened? Why did it happen? Additional prompts: lack of expertise, support from the organization, time, budget, conflicting interests.
  - Challenges brought by responsible AI efforts: Have there been any efforts towards responsible AI, that have turned into hurdles for your own work? What happened? (e.g., time-consuming review processes, lack of information to fill in documentation)
  - Do you have any wishes to help you in relation to RAI and possibly these challenges?

Trust-specific questions. [Adapted to the role of the participant in the LLM supply chain.]

- Do you trust the organization building the AI?
- Do you trust the organization that is using your AI system?
- What do you think of the final end-users?
- Why do you trust them (or not)? In what sense do you trust them?
- Who else do you trust or distrust among all these actors you described?
- What kind of due diligence should one do when building or using AI?
- What would the organization / actor need to do for you to trust them more?

C Details about our Data Analysis Process

Initial marker codes used during the familiarization phase can be found in see Table 5. These included trustors, trustees, trust activities, vulnerabilities, and positive expectations, as well as factors impacting trust. In this phase, aligned with organizational psychology literature that distinguishes attitudes (such as trust) from behaviors (such as reliance), we also annotated any relevant quote. This later resulted in theme 3 about the complex interplay between trust and reliance, as it was not always evident that trust was actually at play at the supply chain junctions.

In the second stage of analysis, we started by re-working and enriching the initial codes. We looked back at these codes, merged and reconciled similar ones, and dug deeper into composite codes and broke them down further. We also coded relations between the basic marker codes as trust is fundamentally relational, e.g., the absence or presence of trust between certain entities at different junctions. Using all our prior descriptive codes, we investigated how they associate together, for example by considering each type of trustor, trustee, or junction, and analyzing potentially relevant patterns. This enabled us to identify surprising patterns within our initial results, e.g., the bi-directionality of trust relations or the need for trust. Organizational psychology literature also partially drove our analysis. For instance, when we noticed that multiple trustees referred to abstract concepts, we associated this observation to the concept of 'generalized trust' that crystallized this sub-category of code in our analysis.

Type of code	Codes
Actors	related to individuals (e.g., product manager, UX researcher, manager or less senior in the hierarchy), related to organizations (provider, consumer, end-user, annotator, within or in-between organizations)
Other entity types	technical artifact (LLM algorithm versus LLM system), abstract concepts
Junctions of the supply chain	development decisions, deployment decisions, adoption decisions, usage decisions
Expectations	trustworthy output, useful output, responsible development, ability to do one's own work
Vulnerabilities	social, environmental impact, organizational risks, individual risk

Table 5: Preliminary marker codes inspired from the literature on trust and human-AI collaboration.

In the third stage of analysis, we developed the larger themes. We first found the nature of trustees to be the most relevant dimension due to the points of interest we identified from comparing similar trustees, such as the similar expectations and vulnerabilities within trustees of the same nature, and the challenges that emerged for specific trustees, e.g., the lack of agreement on what is an AI trustee. This drove the development of themes 1 and 2. Yet, we also found merit in comparing trust factors, attitudes and behaviors, across types of trust relations as they might sometimes be intertwined, impacting each other, or not necessarily even distinguishable. For instance, certain trustors talk about trusting an AI system, but they rapidly substitute this initial trustee with trust in other individuals or organizations. These observations pushed for the creation of the third theme. This is why the final themes start with the traditional focus on AI systems as trustees, then expand to other trust relations we identified that are typically not discussed, and finally shows how these various relations relate to each other.

Note that the three stages were not exactly sequential. Because of the large number of interviews, we conducted these stages sequentially for each interview transcript, but did not wait to have conducted all interviews to do so. Instead, we did so simultaneously with the progress on the interviews, e.g., we conducted five interviews and then these three stages, then conducted more interviews and the stages for these new interviews, etc. We also looked at the batches of interview transcripts altogether not to miss any interesting patterns in-between related participants or in-between codes. For instance, only few participants mentioned certain aspects of generalized trust, that we would not have identified as predominant and relevant without comparing our codes across various batches of interviews. All in all, this process was iterative, as it led us to repeat the three stages several times based on the new interview data acquired.

The first author conducted all the interviews and went through all transcripts and all stages. The third and fourth authors participated in interviews with ten participants, wrote memos that were used in the first stage, and the third author additionally conducted the three stages for the first five transcripts and discussed with the first author to reconcile their codes and themes. The second, fourth, and fifth authors were later involved several times when conducting the second and third stages to discuss latent codes, put them in parallel with existing literature, and shape the final themes.