❌ An Empirical Exploration of Trust Dynamics in LLM Supply Chains ♋

AGATHE BALAYN, ServiceNow, Trento University, Delft University of Technology, The Netherlands

MIREIA YURRITA, Delft University of Technology, The Netherlands

FANNY RANCOURT, ServiceNow, Canada

FABIO CASATI, ServiceNow, Switzerland

UJWAL GADIRAJU, Delft University of Technology, The Netherlands

With the widespread proliferation of AI systems, *trust in AI* is an important and timely topic to navigate. Researchers so far have largely employed a myopic view of this relationship. In particular, a limited number of relevant trustors (e.g., end-users) and trustees (i.e., AI systems) have been considered, and empirical explorations have remained in laboratory settings, potentially overlooking factors that impact human-AI relationships in the real world. In this paper, we argue for broadening the scope of studies addressing '*trust in AI*' by accounting for the complex and dynamic supply chains that AI systems result from. AI supply chains entail various technical artifacts that diverse individuals, organizations, and stakeholders interact with, in a variety of ways. We present insights from an in-situ, empirical study of LLM supply chains. Our work reveals additional types of trustors and trustees and new factors impacting their trust relationships. These relationships were found to be central to the development and adoption of LLMs, but they can also be the terrain for uncalibrated trust and reliance on untrustworthy LLMs. Based on these findings, we discuss the implications for research on 'trust in AI'. We highlight new research opportunities and challenges concerning the appropriate study of inter-actor relationships across the supply chain and the development of calibrated trust and meaningful reliance behaviors. We also question the meaning of building trust in the LLM supply chain.

Additional Key Words and Phrases: AI supply chain, trust, human-AI collaboration, empirical study, LLM

1 INTRODUCTION

Trust is a fundamental necessity for collaboration [51]. For that reason, HCI researchers have started investigating trust relationships between humans and AI [30, 50, 52]. Such research has primarily explored the conditions for the development of appropriate trust in AI systems [5, 48, 56], for society to leverage their potential (e.g., as decision-support systems [7]) while avoiding their harmful impacts when inappropriately developed and used [6]. To this end, it has limited the trustee (i.e., the actor that is trusted) to the AI system (that makes predictions). In turn, the trustor, i.e., the trusting actor of such an AI system, has been limited to a human entity, whether it be the user of the AI system [25, 32, 53], the decision-subject who might have to decide whether to contest the decisions taken [2] or the general public (trust in AI technologies) [2, 27].

However, the collaboration between humans and AI is not limited to the time they interact to conduct tasks collaboratively. The development and deployment of AI systems are also a result of collaborative work between many

actors in the AI supply chain [13, 15]. For instance, an AI system such as an LLM requires individuals in an organization to develop a foundation model, the same or another organization to fine-tune such models and deploy them in an application, a consumer organization should then be willing to adopt this application, and then the end-users can finally interact, engage, or collaborate with the LLM-powered application to carry out tasks. Trust dynamics along the AI supply chain can, therefore, govern or influence the nature of collaboration at each juncture. Coming back to our example, one can for instance imagine that the consumer organization first develops trust in the foundation model or fine-tuned model to decide to adopt the LLM-powered application.

In this paper, we argue that trust between humans and AI should not be limited to the study of the final interaction, and should also account for the complexities of the supply chain that AI systems result from. Besides decades of organizational psychology research [3, 4, 20, 29, 31, 38, 41, 44, 47, 47, 49], the few in-situ studies of trust in AI have already hinted at the existence of many trust relations within AI lifecycles inside the supply chain [10] and many organizational factors impacting these relations [25]. It is hence evident that trust plays a key role not only in the adoption and use of AI systems but also in other synergies within the AI supply chain.

Ultimately, trust dynamics can influence the production of AI systems and their adoption, impacting the trustworthiness of AI systems employed in society, how appropriate user interactions with AI systems are, how responsible their production is, and therefore the direction in which the AI industry is headed. That is why we build on valuable prior research on trust in AI and and aim to broaden the scope of understanding trust relations that influence downstream collaborations and outcomes in the AI supply chain.

In this paper, we explore and identify *the actors beyond the end user and the AI system that serve as trustors and trustees in the AI supply chain, and the factors that impact trust relationships relevant to these additional actors.* To this end, we report preliminary findings from a study investigating trust relationships within LLM supply chains. We focus on LLM supply chains due to their complexity [21] and timely relevance. Our results illustrate the importance, diversity, and complexity of trust relations beyond known trustors and trustees, both within and across organizations. In the future, going beyond a myopic understanding of AI trust will require accounting for new factors impacting trust relations (e.g., organizational reputation and historical relations), and developing new metrics for understanding trust and trust calibration. We hope that this paper acts as a call for AI trust researchers to broaden the scope of their work to address pressing issues in unaddressed but highly relevant trust relations along AI supply chains.

2 BRIEF OVERVIEW OF THE BACKGROUND AND METHODOLOGY

Background. Despite trust not having an agreed-upon definition, a majority of work agrees to characterize a relation of trust by the **positive expectation** the trustor has for the trustee, and the trustor's **vulnerability** and uncertainty vis-a-vis the trustee in the context at hand [31]. According to prior literature, trust relations are not only characterized by the **trustor** and the **trustee**, but also by the trust **activity** [45]. For instance, the user of an LLM might trust it to (activity 1) summarize texts that do not require expertise, but might refrain from relying on it when they need to (activity 2) summarize scientific papers. The literature further distinguishes between trust that is or not **appropriate** and that does or not lead to a reliance and compliance behavior of the trustor on the trustee [36]. Finally, researchers within and outside AI have already identified several **factors** that impact trust [33, 35]. Existing works have explained that trust depends on a) the trustee's characteristics (particularly the ABI framework consisting of ability, benevolence and integrity [35]) mediated by the affordances that expose these characteristics to the trustor [33], and on b) the trustor's characteristics (e.g., disposition for trust). c) The context of the trust activity further mediates the trust relation, e.g., impacting the trustor's systematic and heuristic processing of the trustor's affordances [33].

In the context of AI, trust has been investigated between a user and an AI system conceptualized as one entity that makes individual predictions on input samples [23]. However, AI systems, especially LLMs, are much more complex artifacts [13]. We therefore intend to broaden the scope and look into trust across actors in the LLM supply chain.

Methodology. We conceptualize trust in LLM supply chains by combining prior work in organizational psychology, trust in AI, and the empirical insights we generated from our qualitative study. We are convinced of the necessity to bring more qualitative methods to the trust-in-AI research and especially more empirical *in-situ* investigations of the LLM supply chain and the trust relationships that shape it. The few real-world, empirical, research projects about trust [10, 25] and their insights have hinted at such importance in the past. Hence, we conducted semi-structured interviews with 71 practitioners across 10 organizations that develop, deploy, or use LLMs. Through a mixture of snowball and convenience sampling, we recruited practitioners who play different roles across the supply chain and vis-a-vis LLMs, such as UX researchers and data scientists working for the development of LLMs, legal and risk teams for their assessment, product managers of consumer organization for their adoption, and end-users (cf. Figure 1). We questioned them about their practices and challenges vis-a-vis the LLM, insisting on their relations with other practitioners in the supply chain, and discussing their opinions about LLMs and trustworthy AI. These interviews (over 3600 minutes of recording) were reviewed and approved by the ethics committee of one of our institutions. We then analyzed the interview transcripts deductively and inductively. We especially relied on aspects of trust discussed in prior literature above to inform our analysis and assess the novel or confirmatory character of our findings. Note that the analysis of the transcripts is still in progress. In the future, we may identify more trust relations and trust factors, and nuances to those discussed in this paper, that would allow us to better disentangle the role trust plays in the production and adoption of LLMs.

3 PRELIMINARY FINDINGS & DISCUSSION

3.1 Trustors & Trustees: Trust Relationships Are Extremely Diverse Across the Supply Chain

3.1.1 Many Trustors and Trustees Shape the Supply Chain. We find that, while the trustor is always a human entity, the trustee can be either a human entity or a technical artifact related to an LLM.

Human trustors and trustees can be organizational or individual entities. We find that the human entities in trust relationships can be of two different natures: individual or organizational entities. For instance, the individual end-user of an LLM deployed by organization A might trust the LLM they use, or they might trust organization A to build a responsible LLM, or they could even indirectly trust the LLM because one of their friends might have used it extensively and argued for its usefulness profusely. Similarly, an organization B that decides to buy the LLM services of organization A typically develops a relation of trust with the LLM or organization A before their purchase decision.

Individual and organizational entities play diverse roles vis-a-vis the LLM, entailing different trust activities. These entities all have one primary role vis-a-vis the LLM.
Producer: An entity might participate in the development of any technical component that is used downstream to produce and maintain the final LLM (e.g., one might build the foundation model on which the LLM would be fine-tuned for a specific application). Producers often need to trust upstream components (e.g., the foundation model) to produce downstream ones (e.g., the fine-tuned model). Their primary vulnerability might be the reputational risk, in case a technical component or the LLM would be discovered to behave in a harmful manner or to rely on harmful premises –as this has happened in the past, e.g., testified by the OpenAI lawsuit [18]. Note that it is often not the same organizational

and individual entities working on the different components of the LLM. Policy documents [1, 34] make the useful distinction between the "developer" and the "deployer" of the LLM, where the developer creates a fine-tuned LLM and the deployer makes this LLM ready for use, e.g., integrating it within a software stack, adding filters in the output of the LLM to circumvent potential offensiveness. We find even more fine-grained categories of roles at both the individual and organizational levels.

Consumer: This can either be an organization deciding to use the services of an LLM and providing access to these services to its employees or external users, or it can be an independent user. The consumer employs the LLM to conduct a certain task for their personal or professional work, and typically requires some trust to adopt and continuously use the LLM with, e.g., the vulnerability that the LLM could provide them with offensive outputs.

Indirect stakeholders: Other human entities, beyond those explicitly participating in the LLM supply chain, might be impacted by the LLM or might simply have an opinion about the LLM and its impact on society. This could be considered a trust relation if the entity would consider themselves or others in society at risk due to the development or use of the LLM [28, 46].

Trusting LLMs is more than trusting a stand-alone, shrinkwrapped, technical artifact. As hinted above, an LLM is developed along multiple phases and relies on several technical components. We find that the trustee in trust relations might revolve around the final LLM or its separate components, either the components of the LLM system in use or the components that have served in building these in-use components. In use, an LLM-powered system is composed of a fine-tuned model for the application at hand, but also user interfaces for its end-users, workflows for post-filtering the outputs of the LLM (e.g., for toxicity), logging data and monitoring potential issues, infrastructures to support the computations and data needs, etc. Note that an LLM system might even be composed of several fine-tuned models chained together [39]. Building each of these in-use components requires various build-up components, such as pre-training datasets, the foundation model on which the model was fine-tuned, a fine-tuning dataset, training scripts, data processing scripts, more infrastructures to handle data and computations, etc. Naturally, the granularity of these components is not set, and sub-components make up these in-use-components and build-up-components, e.g., a training dataset is made of data samples and annotations, a filtering workflow might be made of several filters for gender bias, race bias, toxicity, etc. *P33 "[Deployer organization], as a trusted platform, is easier to get started with and to deploy use cases on than building a new system from scratch, and having to work through your own security and firewalls and load balancers, and all the complexity behind standing up a new web application."*

3.1.2 Many Trust Relations Connect These Entities Across the LLM Supply Chain. Based on all these entities in the LLM supply chain, we identify a diversity of trust relations that are marked by different trust activities. For each trust relation, we systematically disentangle the trustee(s), trustor(s) and trust activity(ies).

Downstream to Upstream, Entities Trust Different Technical Artifacts. Entities relying on an LLM or one of its components, i.e., entities that are relatively downstream in the supply chain, often posed unexpected obstacles upstream due to potential issues (e.g., low performance and bias of technical artifacts), respectively for the adoption of the LLM, or for the adoption of its components and consequently for the development of the LLM itself. *P35 [consumer organization] "There was a hesitancy to use it. There was a lack of trust. "I want to see it proven out before I bring it into my organization." Now there's a stronger appetite to bring it in because they see the efficiency gains."* Hence, there, trust relations are not only between (1) the individual end-user and the LLM, but also between the (2) consumer organization and the LLM, as well as trust (3) between the deployers and the fine-tuned model (and potential additional in-use components provided
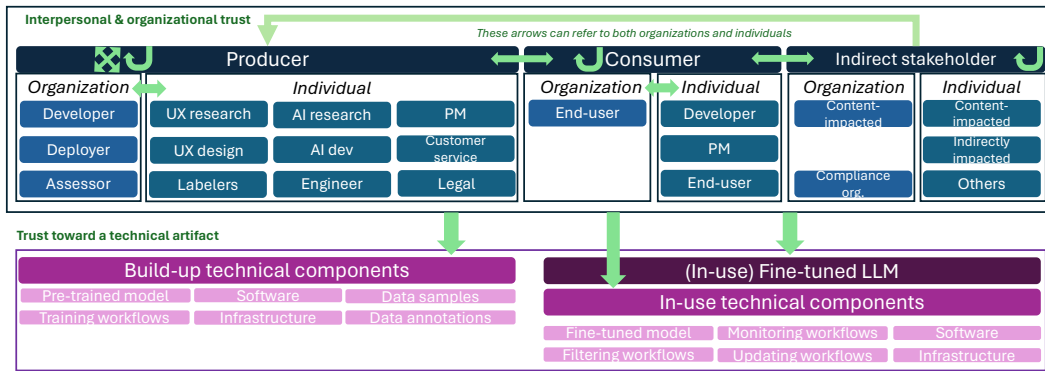
Fig. 1. The primary entities (boxes) and trust relations (arrows) that shape the supply chain. The ✕ cross and ↻ loop arrows respectively represent trust relations across different entities in the same box (e.g., developer and deployer organization), and trust relations across entities of the same type (e.g., two different developer organizations). Remember that the supply chain is not only made of trust relations, but also other relations across entities (e.g., one developer organization might develop a pre-trained model and another one might adopt this model to fine-tune the LLM.)

with the fine-tuned model), and trust (4) between the developers and the build-up components they are given. In all cases, the positive expectation is for the technical components to work properly. The vulnerability consists in the repercussions that using a flawed AI system can have, e.g., while it might be reputational risk for an organization due to problematic outputs of the AI system or unethical ways to build it, it would be about the impact of problematic outputs (e.g., inaccurate or offensive) for the end users. *P70 LLM user "If it's wrong and I blame AI, that's gonna make me look unprofessional. So I try not to overuse it, it's a delicate balance. I'm probably going to get sued, or I'm gonna be the one in trouble. If I'm using AI as help, a lot of people don't trust that already, they want the Encyclopedia Britannica. I'd have to take responsibility for that. The big companies are probably gonna blame me."*

Upstream Entities Trust Downstream Entities to Responsibly Use Technical Artifacts. Our interviews also show that trust is bidirectional when talking about the usage of an LLM artifact. It is not only that the user of an artifact trusts the artifact or its developer (discussed above). It is also that the developer of the artifact trusts its consumer to use it appropriately (e.g., not outside the pre-defined scope of applications) to avoid any harm, at the risk of reputational issues. They also have to trust the consumers that they will not game the LLMs, e.g., via adversarial attacks to recover private information, and that they potentially put safeguards in place themselves. They might also trust the consumer organization to create trainings for its end-users to avoid misuse, e.g., not to over-rely on LLMs. In turn, the consumer organization trusts its individual end-users not to misuse the system and to properly handle problematic outputs of the system. For instance, one organization mentioned that when the end-users are internal, they trust their end-users to internally report on any toxic or offensive output and not to cause a public outcry and discuss the issue on social media. In summary, these upstream-to-downstream trust relations impact the responsible use of LLMs.

3.1.3 Discussion. Our results confirm our intuition: the LLM supply chain depends on a diversity of trust relations. These trust relations involve many more trustors and trustees than current research on trust in AI envisions. The objects of these relations (i.e., the activity conducted by the trustee) also revealed to be more diverse than studied until now. This is aligned with prior research on trust in the context of automation (prior to AI) [31], where one has investigated trust between individuals ("partner trust", "interpersonal trust"), between individuals and organizations (or

even institutions), or between several organizations. Trust relations are often prerequisites to drive the production and adoption of LLMs. They are also vectors that support or hinder the establishment of trustworthy LLMs and their responsible development. Closest to our findings is the study conducted by Browne et al. [10] on trust in clinical AI contexts. This study shows that clinicians (trustors) trust AI systems (trustees) based on different activities all along the AI lifecycle (e.g., how AI training data were collected, which team developed the AI), and such trust is necessary for the clinicians to adopt and use the AI system. Yet, this study does not adopt a supply chain perspective. By accounting for the dynamics of the supply chain, our results further highlight the importance of inter-organizational trust, and identify trustors and trustees across the supply chain whose collaborative work impacts LLM production and adoption.

3.2 Trustee-Related Factors: Trust is Fostered By Diverse Trust Cues About the Trustee

Our study revealed that the same types of factors as identified in prior works impact trust in the newly-surfaced relations. Therefore, through the lens of the ABI framework [35], for each entity (i.e., organizations, individuals, LLMs), we discuss the cues that might express their ability, benevolence, and integrity.

*3.2.1   The ABI Properties of **Organizations** Target Trustors Differently.* Organizations, i.e., trustees, put forward cues related to their ability, integrity, and benevolence, particularly towards their customers, and put efforts into improving their actual trustworthiness in that regard. For instance, the hiring of researchers and publications of research papers, and cross-organization collaborations towards the development of LLMs [11] provide a sense of ability, signed contracts and treaties [54], responsible AI principles [16], and emerging governance structures (e.g., many participants mentioned that with the rise of LLMs, review boards and working groups are slowly established towards handling the risks that LLMs pose) instead illustrate benevolence and integrity. Marketing campaigns and other release of communication pieces served to communicate these cues to certain trustors (particularly customers of the developer organizations). On the other hand, we did not find any trustworthiness cues on the side of the organizations that adopt LLMs, neither towards their employees nor towards their end-users, nor any benevolence cues inter-organizational (e.g., around job displacement) and towards indirect stakeholders (e.g., artists whose content could be used as training data [9]).

On the trustor's side, how these cues are perceived remains to be better understood. For instance, many of our participants, in their capacity as trustors, discussed the absence of benevolence from certain organizations, and instead emphasized the profit motive of these organizations. Additionally, some participants acknowledged benevolence but recognized the practical impossibility of exercising benevolence and developing meaningful governance processes in the race toward LLM development and adoption. For instance, while data stewards exist, whether they can comprehensively fulfill their duty and whether rigorous evaluations of their work are conducted and communicated (e.g., percentage of copyrighted data still included in training datasets) were different questions.

*3.2.2   Not All **Individual** Trustees Display All ABI Properties.* Many trustors discussed trust towards employees of the same or different organizations, referring to their expertise for building LLMs (ability), their integrity in building the LLMs, and general benevolence. Ability was referred to via the publication of academic papers in prestigious venues, corporate recognition, the belief that certain teams had the knowledge and could teach others about trustworthy LLMs, and the responsible AI-related communication efforts and actions organized by a few individuals. In terms of benevolence, participants discussed their own motivation and the dedication of their colleagues towards social good and developing ethical systems. Discussions about integrity only revolved around the end-users of LLMs. For instance, organizations employing these users or indirect stakeholders discussed users' integrity in terms of not querying LLMs

with prompts to copy the styles of living artists. We did not identify any other integrity consideration for any other type of participant, nor questions of benevolence and ability from the consumers towards providers or indirect stakeholders.

3.2.3   ABI Properties of **LLMs** *Are Conflated With the Properties of Other Trustees.* Research has long investigated the factors that make AI trustworthy and that communicate its trustworthiness [33]. In our interviews, the same factors and cues surfaced overall, particularly in terms of ability. *P40 "Everyone's gonna want something different. We're gonna get to a place where it is gonna come to trust. I trust that this is going to protect my information. I trust that this is going to be a good response. I trust this could be reliable. And if I can't trust it, I'm not going to use it, whether I'm the one that's supporting it, bringing it into the company or an end user. If I can't trust this response, I'm gonna call the help desk every single time."* We found that the benevolence of the LLM-powered system has been conflated with the benevolence of the producer organizations (particularly the deployer one) towards the consumer (similarly to [25, 33]), but also the benevolence of the consumer organization towards its individual users. In terms of integrity, LLM integrity revolves around "the degree to which the operational or decision process of the model is appropriate to achieve the users' goal" [33]. The primary discussions we identified around this topic were about receiving indications about the documents used for the outputs of retrieval-augmented generation use-cases. This was especially relevant to the consumer organizations, but we did not find such forms of AI integrity (nor others such as explainability) mentioned by any other participants.

3.2.4   Discussion. Our results showed a plurality of factors that impact trust relations. These factors align with the ability, benevolence, integrity framework proposed by Mayer et al. [35] and applied to AI from a communication perspective by Liao and Sundar [33]. Naturally, the exact factors of trustworthiness and the cues that signaled these factors were however different across different trusted entities, and consequently different from those prior studies have investigated. This confirms the importance of disentangling the LLM supply chain, the entities that shape it, and the exact trust relations that traverse it, to more comprehensively identify trust factors. While these factors had not been explicitly discussed in the past in the context of AI, a few studies [17, 25] have hinted at the impact of the organizational context on the appreciation of an AI system. Here, the "organizational context" is a trust relation where the trustee might not be the LLM but instead an employee or organization, and this relation impacts trust towards the LLM.

3.3   Trustor-Related Factors: Trust, Appropriate Trust & Reliance also Depend on the Trustor and Context

3.3.1   Trust Varies Across Trustors. We notice that the extent of trust across trustors towards the same trustee varies. We found that both contextual and personal factors mediate trust relations. In terms of context, not only the stakes of the trust activity impact the relation (e.g., whether adopting the LLM might create financial risks or reputational damage to the consumer organization), but also the expectations the trustor has towards the trustee. For instance, we find that different consumer organizations have different expectations towards the ability of the LLM provided by the deployer organization depending on the environment in which individual users will use this system. Particularly, when individual users are employees of the consumer organization, we primarily find expectations in terms of accuracy, while in cases where the users are external to the organization, we also find expectations concerning the offensiveness of the outputs of the AI system.

In terms of personal factors, the natural propensity for trust in organizations and in technology, the ways individuals perceive AI systems, their advantages, but also their potential concerns, e.g., about dataset biases, also play a role. Besides, the knowledge and awareness of the participants about LLMs, how they function, and what harm they can cause also impacted the types of trustworthiness cues they were looking for and their potential distrust towards LLMs.

P34 "Continuing to educate everybody on what's the best use of AI and what is trustworthy AI, it will allow each team to make better calls of what to use."

3.3.2   The Supply Chain is Prone to (Un)Calibrated Trust. Reflecting on all the factors we surfaced, we found various avenues that lead to potentially non-calibrated trust. The ways trust relations affect each other is one of the key risk areas. On the side of the trustor, for instance, we found that both the temporality of the supply chain, the organizational versus interpersonal trust, and the impact of different trust activities can sometimes lead to blindly trusting an LLM despite its characteristics not meeting trustworthiness expectations. For instance, consumer organizations explained trusting deployer organizations for developing trustworthy traditional machine learning systems, and for this reason, they would also trust the deployers for developing trustworthy LLMs. *P48 "That's one of the things that establishes [developer-org] as an authority on it. We don't have this relationship with other vendors, so how do they expect adoption. So I think just engaging us in things like this really helps establish you as an authority. Maybe it's a bad thing, and it's kind of that segregation of duties: we've built this trust, and therefore we're going to trust you."* Within chained trust relations, we also identified potentially wrong beliefs (e.g., about other's trust relations), that led to inappropriate trust.

On the side of the trustee, we found that the trustee often needs the trust and reliance of the trustor for various reasons. Upstream, this is for job retention and climbing the corporate ladder, prompting the development of LLMs and their use and hence economic growth, and downstream for getting access to the technical artifacts. In this context, the trustee might have the incentive to develop unwarranted trust. This incentive sometimes plays a role in the beliefs of the trustor in the integrity of the trustee. It can also lead the trustee to display untruthful or misguiding trustworthiness cues [8, 12, 14]. *P18 "There's some sort of inherent trust that it is someone else's problem down the line, and that's why I go back to: the person writing the requirement needs to make sure that they're considering these things."* Yet, the need for trust also becomes, in certain cases, a motivation for trustworthiness. Organizations recognizing that trustworthiness could become their competitive advantage compared to simply increasing the accuracy of an LLM might put more effort into developing such trustworthy LLMs. *P15 "If you throw solutions which lose customers, we lose everything. So we are very mindful. We want to be responsible. We want trust. It's not just customers, even internal people."*

Note that we also found situations where one entity relies on another one without trusting it. For instance, in the case of the user organization, they might want to buy the services of an LLM to remain competitive (vis-a-vis competitors who would also do so) while knowing that this LLM is not yet trustworthy.

3.3.3   Discussion. These results show that trust is not only impacted by the characteristics of the trustee but also by those of the trustor and the context in which it is. While the characteristics of the trustor that impacts their trust relationship are overall not different from prior work (e.g., except their general propensity to trust organizations), the LLM supply chain brings new contextual factors of importance. To the best of our knowledge, no prior work has discussed these contextual factors. Finally, we found that these contextual factors and personal factors can lead to both calibrated and non-calibrated trust, and trust is not always present when reliance is. To the best of our knowledge, prior work [12] had not identified the intentional (or not) ways to deceive trustors. Furthermore, while prior works have made clear the distinction between trust as an attitude and the behaviors that can result [31], accounting for the complexity of LLM supply chains has allowed us to identify and discuss organizational strings and powers that can impact trust and related behaviors.

4   IMPLICATIONS & FUTURE RESEARCH

Our preliminary results bear implications for practitioners, researchers, and policy-makers, and raise diverse research questions for each of these stakeholders.

How Does Trust Fit in LLM Governance? Our findings showed the complexity of trust relations and hinted at the challenges toward establishing calibrated trust across the supply chain. Trust enables collaborations and collaborations are necessary for the development and adoption of LLMs. Furthermore, calibrated trust should enable appropriate collaborations, ideally leading to compliant and ethical LLM supply chains.

This suggests the importance of bringing research on trust and AI collaborations to policy-makers and LLM governance researchers. Ideally, they should account for the variety of trustors and trustees across the supply chain, and for the challenges, needs, and driving forces each of them has. How can we account for trust relations while ensuring compliance of LLMs to their upcoming regulations? If it is possible to achieve a certain level of trust and appropriate trust within and across organizations, what is the most appropriate model of LLM supply chain governance? How do current notions of accountability [42] and distributed responsibility [40] fair with the complex AI supply chain we identified, and particularly with the various trust relations required in this supply chain? Furthermore, future policies would need to acknowledge the various factors that might impact trust, and especially calibrated or unwarranted trust across the different entities in the LLM supply chain. Hence, how can we ensure that each trustee develops their ability, benevolence, and integrity appropriately, and that the trustworthiness cues they put forward are truthful? How do such trustworthiness cues fit with the potential trade secrets of each organization in the supply chain? From the trustor's side, do we need to foster more reflexivity among each entity of the supply chain, before trusting them to build compliant systems and interact with these systems appropriately? Or should we instead establish stricter requirements and assessment processes? Can internal assessments of LLMs be truthful if they rely on trust? Prior research on trust and governance outside AI contexts could be a good start to answer these questions [19].

What Kind of Trust Is Desirable and Achievable? Our findings also suggest taking a step back and asking whether any type of trust is achievable and desirable anywhere in the supply chain, and what type of trust do we actually want to foster in each trust loci. On the one hand, subjectivity surfaced across the supply chain. We found that trustors have different expectations concerning the ability, integrity, and benevolence of an LLM and its supply chain. We also found that they expect different values to be accounted for in LLMs. Questions of subjective expectations for the outputs of LLMs and whether they can feasibly be met have already been raised [26]. Research has further shown the impact of trustors sharing the trustee's values on trust [37]. Bringing these prior results to the LLM supply chain could hint at the potential impossibility of building trust across the entire supply chain. This suggests that we should be cautious in defining the type of trust and trustworthiness cues needed across the supply chain. Instead of relying on each trustor's individual expectations, we might for instance need to foster discussions between supply chain entities to agree on a common set of expectations for the LLM supply chain, and to subsequently base trustworthiness cues on such commonly-accepted expectations. On the other hand, because the supply chain is so complex, it revealed impossible for each trustor to develop a well-informed trust in each relevant trustee: their visibility on upstream and downstream parts of the supply chain is limited [55]. As a result, we might want to distinguish different trust relations and different desirable processes to foster trust in these relations. For instance, we could investigate how to develop informed, reflexive, trust among entities that are close in the supply chain, while accepting "blind" trust among entities that are more distant as this is also necessary to the development of the LLM supply chain.

Studying New Trust Relationships in the LLM Supply Chain. Finally, the novel character of our findings confirms the importance of adopting empirical, qualitative methodologies, grounded in the realities of the study participants to further investigate trust. With current assumptions in existing, quantitative, studies about trust in AI (trustors, trustees, activities, and impacting factors) [36], we would not have been able to identify these new complexities. Our study is however only the start of the exploration of trust within the LLM supply chain. We will need to investigate in more detail each of the trust relations and their impacting factors. For such research, we envision several practical and conceptual challenges. In terms of recruitment of participants, we do realize the difficulty of entering in contact with diverse stakeholders within organizations, and of discussing questions of trust that might be confidential. This might become especially challenging for quantitative studies that would require more participants. Yet, starting by conducting quantitative studies among populations of LLM users (as done currently) and AI developers before broadening to other entities, and accounting for the new factors, would already provide impactful insights. We also envision that the type of organization in which the participants work could impact their trust, and it will be important to understand the relevant dimensions of organizations (e.g., AI maturity [43]). Finally, researchers will need to account for the diverse designations that are used across organizations to refer to potentially similar or different roles vis-a-vis the LLM supply chain, and will also need to pay attention to terminological confusions in the way they address these different roles (varying professional cultures and jargons across organizational teams [22]).

Another challenge will revolve around adapting the ways to measure trust, and the ways to present the impacting factors. To set up the studies about the trust of the user of an AI system, the reliance behavior on individual outputs of the AI system [52] is often used as a proxy. However, the new trust relations we identified do not necessarily involve reliance choices over these different outputs, but instead might revolve around the decision to adopt or to provide access to a technical artifact, or to delegate such decision to another, potentially more knowledgeable, individual in the same organization. Therefore, whether and how these behaviors could be used as new proxy measures remains to be investigated. Additionally, careful thought should also be put into developing ways to convey potential relevant factors to the trustees in the studies. Particularly, this will require understanding how to present the values of the organizations within the supply chain (e.g., presenting them vaguely, in an explicit list, in a ranked list, with or without information about the purpose of the LLM), as our study showed that the participants not only weigh various advantages and disadvantages of LLMs differently, but also prioritize values differently (similarly to [24]), and that these prioritizations might depend on the purpose of the LLM and the stakes they have in it.

REFERENCES

[1] BSA (The Software Alliance). [n. d.]. *AI Developers and Deployers: An Important Distinction.* https://www.bsa.org/files/policy-filings/03162023aidevdep.pdf

[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] 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.

[4] 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.

[5] 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.

[6] Agathe Balayn and Seda Gürses. 2021. Beyond Debiasing: Regulating AI and its inequalities. *EDRi Report. https://edri. org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online. pdf* (2021).

[7] 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.

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

[9] Washington Post Editorial Board. 2023. Opinion | AI could threaten creators — but only if humans let it. https://www.washingtonpost.com/opinions/2023/11/24/ai-llm-intellectual-property-crisis/

[10] 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.*

[11] 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/

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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

[19] Ranjay Gulati and Jack A Nickerson. 2008. Interorganizational trust, governance choice, and exchange performance. *Organization science* 19, 5 (2008), 688–708.

[20] 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.

[21] 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.

[22] Pippa Hall. 2005. Interprofessional teamwork: Professional cultures as barriers. *Journal of Interprofessional care* 19, sup1 (2005), 188–196.

[23] 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).*

[24] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How different groups prioritize ethical values for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 310–323.

[25] 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.

[26] 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).

[27] 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.

[28] 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.

[29] 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.

[30] 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).

[31] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[32] 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.

[33] 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.

[34] Tambiama Madiega. 2021. Artificial intelligence act. *European Parliament: European Parliamentary Research Service* (2021).

[35] 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.

[36] 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).

[37] 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.

[38] 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.

[39] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6. 126–135.

[40] Will Orr and Jenny L Davis. 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society* 23, 5 (2020), 719–735.

[41] 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).

[42] 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.

[43] 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.

[44] 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.

[45] Paul Smart, Brian Pickering, Michael Boniface, and Wendy Hall. 2021. Risk Models of National Identity Systems: A Conceptual Model of Trust and Trustworthiness. (2021).

[46] 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).

[47] Hwee Hoon Tan and Augustine KH Lim. 2009. Trust in coworkers and trust in organizations. *the Journal of Psychology* 143, 1 (2009), 45–66.

[48] 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.

[49] Gaurav Tejpal, RK Garg, and Anish Sachdeva. 2013. Trust among supply chain partners: a review. *Measuring Business Excellence* 17, 1 (2013), 51–71.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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/

[55] 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.

[56] 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).