Rethinking Trust in AI Assistants for Software Development: A Critical Review

Sebastian Baltes, Timo Speith, Brenda Chiteri, Seyedmoein Mohsenimofidi, Shalini Chakraborty, Daniel Buschek
University of Bayreuth, Germany
{sebastian.baltes, timo.speith, brenda.chiteri, s.mohsenimofidi, s.chakraborty, daniel.buschek}@uni-bayreuth.de

***Abstract*—Trust is a fundamental concept in human decision-making and collaboration that has long been studied in philosophy and psychology. However, software engineering (SE) articles often use the term 'trust' informally—providing an explicit definition or embedding results in established trust models is rare. In SE research on AI assistants, this practice culminates in equating trust with the likelihood of accepting generated content, which does not capture the full complexity of the trust concept. Without a common definition, true secondary research on trust is impossible. The objectives of our research were: (1) to present the psychological and philosophical foundations of human trust, (2) to systematically study how trust is conceptualized in SE and the related disciplines human-computer interaction and information systems, and (3) to discuss limitations of equating trust with content acceptance, outlining how SE research can adopt existing trust models to overcome the widespread informal use of the term 'trust'. We conducted a literature review across disciplines and a critical review of recent SE articles focusing on conceptualizations of trust. We found that trust is rarely defined or conceptualized in SE articles. Related disciplines commonly embed their methodology and results in established trust models, clearly distinguishing, for example, between initial trust and trust formation and discussing whether and when trust can be applied to AI assistants. Our study reveals a significant maturity gap of trust research in SE compared to related disciplines. We provide concrete recommendations on how SE researchers can adopt established trust models and instruments to study trust in AI assistants beyond the acceptance of generated software artifacts.**

***Index Terms*—Software Engineering, AI Assistants, Large Language Models, Human-Computer Interaction, Information Systems, Trust, Literature Review, Critical Review**

I. Introduction

Since the launch of ChatGPT in November 2022 and the subsequent hype around generative artificial intelligence (GenAI), AI assistants that help humans generate images, text, or source code have become an integral part of the daily work of many knowledge workers around the world [94]. GenAI tools including ChatGPT and GitHub Copilot support software engineers in tasks such as understanding existing code, fixing bugs, or implementing new features [72], [135]. GitHub claims that, as of February 2023, Copilot is "*behind an average of 46 percent of developers' code across all programming languages.*"[1] The IDE integration of GitHub Copilot and similar AI development assistants allows tool vendors to collect fine-grained usage data to understand which suggestions developers accept, which they reject, and which they modify before committing a change [163].

[1]GitHub Blog: GitHub Copilot now has a better AI model

Previous work coauthored by employees of companies with commercial interests in increasing tool adoption, including Microsoft, Google, and Amazon, investigated what they refer to as **trust** in AI assistants [24], [31], [64], [101], [117], [149]. In some of those articles, the concept of *trust* is operationalized as tool adoption in general or, more specifically, as a high **acceptance rate** for generated content. In short, they equate trust with a high probability of developers accepting generated software artifacts. Compared to existing conceptualizations of trust in foundational disciplines such as psychology and philosophy, but also more applied disciplines such as human-computer interaction (HCI) and information systems (IS), we consider this view to be **too simplistic**. It **disregards decades of research on trust**.

Although embedding software engineering (SE) research in existing models and theories of trust is—as our literature review will show—not widespread, our community does have a history of discussing **trust in source code**. For example, in his Turing Award lecture in 1984, Ken Thompson addressed the morality of trust, stating: "*You can't trust code that you did not totally create yourself... No amount of source-level verification or scrutiny will protect you from using untrusted code*" [142]. Given the widespread adoption of AI-based code generation and the persistent issue of hallucinations [14], it becomes imperative to adopt a more **holistic perspective on trust**, one that extends beyond the notion of accepting AI-generated content. Combining Ken Thompson's statement with the increasing amount of generated code, this leads to a critical question: *Can developers genuinely trust generated software artifacts or is there a conceptual difference between trust and mere acceptance of generated content?*

Our analysis of conceptualizations of trust in other disciplines reveals that viewing trust as mere acceptance of generated content is **problematic** for two key reasons. First, it lacks a solid theoretical foundation because it is unclear whether what is being measured is actually aligned with common conceptualizations of trust. People's behavior, and hence their likelihood of accepting generated content, can be influenced not only by the perceived trustworthiness of the AI they are using, but also by other factors. Factors that are not necessarily related to trust [78], [92] include people's specific situation (e.g., time pressure), the type of AI system they are interacting with, the specific task they are performing and the potential consequences of decisions. Second, viewing trust as mere acceptance of generated content is potentially harmful. Even if trust could be measured through acceptance, this would not imply that the measured trust is appropriate trust. Given the potential maintainability challenges posed by generated code [85], the security risks associated with misplaced trust in AI-generated code [108], and the potential negative impact on the developers' learning [126], [156], research needs to distinguish **appropriate** from **inappropriate trust**.

We advocate for a broader interdisciplinary approach to understanding trust in SE. Note that, in this paper, we focus on **human trust**, not trust in a more technical sense (e.g., chains of trust in a security context). We address the following research questions:

RQ1 How is trust conceptualized in foundational disciplines and applied disciplines related to SE?
RQ2 How is trust conceptualized in SE?
RQ3 How can we adopt existing trust models for SE research to study trust in AI assistants?

After introducing the foundations of human trust based on research in psychology and philosophy (Section II), we present the results of a literature review on conceptualizations of trust (Section III-B) in HCI and IS (**RQ1**) as well as SE (**RQ2**). We complement that literature review with a critical review of trust conceptualizations (Section III-C) in recent SE articles (**RQ2**) and then synthesize the results of those two literature reviews into a discussion on how to rethink trust in AI assistants for software development (**RQ3**), providing actionable recommendations (Sections IV and VI).

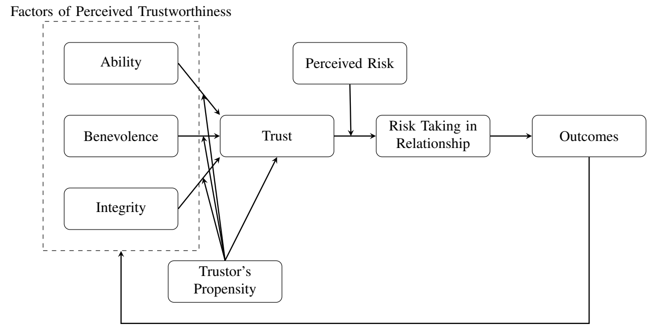

Fig. 1. Trust model by Mayer *et al.* (own illustration based on [92]).

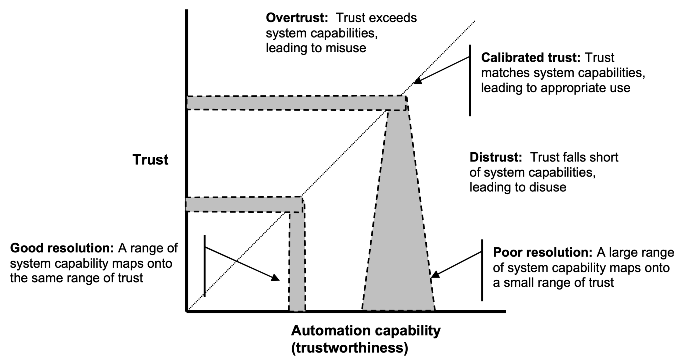

Fig. 2. Trust model by Lee and See (see Figure 2 in [78]).

II. Background: Foundations of Trust

Trust and the related concept of *trustworthiness* have been discussed and conceptualized long before disciplines such as SE, HCI, or IS recognized these concepts in their scientific discourses. Therefore, before we dive into our cross-disciplinary literature review of trust, we outline how philosophy and psychology, that is, disciplines which have been studying the foundations of human behavior for a long time, conceptualize *human trust*.

First, it is important to distinguish *trust* from *trustworthiness*. Both philosophy and psychology agree that *trust* is an attitude that one entity (the trustor) has towards another entity (the trustee) that it hopes will be *trustworthy*. Hence, leaving *perceived trustworthiness* aside, *trustworthiness* is considered a property of the trustee and not an attitude [78], [98], [99]. Furthermore, both disciplines agree that *trust* is a contextual matter: a trustor may trust a trustee in one situation or for one specific task while, at the same time, not trusting them in another situation or for another task. Typically, the trustor is human, while it is debatable whether the trustee needs to be human as well (see Section II-B for details).

A. Psychology

In psychology, two models have been very influential in the discourse on trust. Those models have also been adopted in other fields. In 1995, Mayer *et al.*, surveying the literature on trust, created the *ability, benevolence, and integrity model of trust* [92] (see Figure 1). They defined trust as *"the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party"* [92, p. 712].

In this model, *ability* refers to the trustee's competence in a given domain, *benevolence* captures the trustee's (assumed) goodwill and positive intentions towards the trustor, and *integrity* refers to the trustee's adherence to ethical principles the trustor agrees with. The model suggests that trust is high when all three components of trustworthiness are perceived as high by the trustor, whereas deficiencies in any factor can weaken trust. How much (changes in) the factors affect trust depends on the trustor's *propensity to trust*. This propensity also influences the trustor's *initial level of trust*. Trust can manifest itself in specific situations, depending on the level of trust and the *risk* perceived by a trustor when they trust the trustee to do a certain task. Depending on the outcome of this risk-taking, the perception of the three factors can change.

Lee and See take this model, expand it, and apply it to automation contexts [78]. They introduce the concepts of *calibration*, *resolution*, and *specificity* (see Figure 2). *Calibration* concerns the crucial question of whether the trust a trustor has corresponds to the real capabilities, that is, the actual trustworthiness of a system. Ill-calibrated trust may lead to disuse (in the case of *dis-/undertrust*) or misuse (in the case of *overtrust*) of a system. *Resolution* is about how changes in system capabilities change levels of trust. For example, with low resolution, high changes in capability would lead to small changes in trust. Finally, *specificity* reflects how granular the trust is. Specificity emphasizes the context-dependence of trust by stating how much trust depends on the capabilities of a particular subsystem or a particular situation. According to Lee and See, good calibration, high resolution, and high specificity of trust can mitigate misuse and disuse of automation systems and enhance human-automation partnerships. Although they also introduce a model on the interaction of context, agent characteristics, and cognitive properties with the appropriateness of trust, their work is mostly cited for making the above distinctions.

B. Philosophy

McLeod analyzes philosophical debates on trust and distill three requirements for trust [98]. In summary, trust requires that one can:

1) be *vulnerable* to others, in particular to betrayal,
2) rely on others to be *competent* to do what one wishes to trust them to do, and
3) rely on them to be *willing* to do it.

For most philosophers, trust is a kind of *reliance*, although it involves "*some extra factor*" [98]. This extra factor is often determined by how philosophers interpret the three requirements listed above.

For example, regarding the factor of *vulnerability to betrayal*, trust is often seen as pre-supposing *anti-monitoring* (see, e.g., [10], [71]). The idea here is that if a person monitors an entity, this person cannot be betrayed by that entity. At most, the person could be disappointed if the entity does not perform as intended. Here, there is an interesting difference to Mayer *et al.*'s theory of trust, which explicitly decouples monitoring from trust. However, their theory is flexible enough to model monitoring situations. For example, an employee may devalue the *benevolence* of (and thus the trust in) their employer due to constant monitoring.

The factors of *competence* and *willingness* (i.e., motivation) have received the most attention in the debate, with many competing theories that exist [98]. These theories range from the trustee acting out of self-interest [51] or goodwill [10], [66] to theories of normative expectation [43]. In summary, while there is a minimum consensus of the three requirements mentioned above, philosophers differ so much in the exact formulations of these requirements that some even assume that there are different types of trust. This would mean that trust is not one form of reliance but different ones [30], [121], [131].

Philosophers place more emphasis on the *preconditions of trust* or the characteristics of its *formation* than psychologists. Therefore, in addition to the concept of *calibrated trust* (which they call *well-grounded* trust), they also introduce the concepts of *justified* and *plausible* trust [98]. *Justification* is about the indicators that have been consulted for trust formation, that is, whether it is based on good evidence. If a person happens to trust the system in the right way, but does so by chance without any indications, then philosophers would not speak of justified trust. Finally, *plausibility* is about whether the preconditions for trust are satisfied. For example, if two parties are antagonistic, there can be no trust between them.

Justified trust and *well-grounded* trust are independent of each other. A person might be justified in their trust, for instance, by receiving reports on the trustee's performance. However, these reports might be forged, significantly overestimating actual performance, so that the ensuing trust would be ungrounded. Similarly, the above example of the trustor trusting the trustee to the right level by chance is a case of well-grounded but unjustified trust.

C. Philosophy and Psychology on Trust and AI

Philosophical and psychological discussions about trust continue to this day, often in specific areas of application. AI is one of them, especially in recent times. There are two discussions that are worth highlighting here. In particular, with regard to the *plausibility of trust*, there are discussions as to whether it makes sense to apply the term trust to AI at all. From a philosophical point of view, AI has no motivation with which it could betray us.

This discussion even made it into the media when the *Guidelines for Trustworthy AI* of a high-level AI expert group set up by the European Commission were published. One of the members of this group, the German philosopher Thomas Metzinger, gave an interview shortly before the guidelines were published, in which he complained about the strong involvement of companies in the design of the guidelines. He claimed that these would use the principle of *trustworthy AI* for "*ethics washing,*" because in his opinion, the term trustworthy AI is conceptual nonsense.[2] While some authors agree with Metzinger in that the concept of trust cannot be applied to AI systems, for example, since they do not possess emotive states or cannot be held responsible [37], [116], others defend the possibility to trust AI, for instance, by showing how philosophical theories of trust can be adapted to AI [30], [60], [104], [118], [146] or by differentiating trust in AI from trust in humans [160] (see [38] for an overview of the debate).

Irrespective of the discussion as to whether the concept of trust can be meaningfully applied to AI, the *Guidelines for Trustworthy AI* had strong effects on the subsequent operationalization of trust in AI. By specifying requirements that are essential for trustworthiness, they provide fertile ground for psychological and philosophical research to operationalize these requirements more precisely (see, e.g., [82], [91]).

The other noteworthy discussion revolves around the relationship between explainable AI (XAI) and trust in AI. Increasing trust is one of—if not the–primary goal of XAI [68], [77]. Kästner *et al.* refer to this as the *explainability-trust hypothesis* [68]. However, studies show that providing explanations does not consistently increase trust, and in many cases, even decreases it [68], [77]. Here, trust calibration plays an important role, as explanations that reveal system errors should lead to lower trust. Unfortunately, psychological biases influence this process, as people sometimes trust a system simply because of the illusion of an explanation [40]. On the same topic, the discussion about the plausibility of trust on the philosophical side is worth noting. For example, Ferrario and Loi argue, drawing on the concept of anti-monitoring, that explainability cannot contribute to trust at all, as it constitutes a form of monitoring [41].

[2] EU guidelines: Ethics washing made in Europe (Tagesspiegel)

III. Literature Review

After outlining the foundations of trust, we now turn to the methodology and results of our review of trust conceptualizations across the disciplines IS, HCI, and SE.

A. Methodology

The goal of our research was to understand how trust is conceptualized in SE (**RQ2**) compared to other disciplines (**RQ1**) and, based on these conceptualizations, to explore how existing trust models can be adapted to study AI assistants in software development (**RQ3**). To achieve this, we conducted a literature review across disciplines.

1) Literature review across disciplines: We decided to use *Google Scholar* for the literature review, because portals such as the *ACM Digital Library* and the *IEEE Xplore Digital Library*, which are commonly used for literature reviews in SE, do not cover common publication venues in IS. We used the Python library *scholarly*[3] to automate the retrieval of articles using a common search query across disciplines:

```
DISCIPLINE (trust OR trustworthiness
 OR trustworthy)
```

The values for DISCIPLINE were 'software engineering', 'human-computer interaction', and '(management OR information systems)'. We selected HCI and IS because they are closely related to SE, but we expected them to have a more mature view on trust and its conceptualizations, in particular because both HCI and IS have a strong connection to psychology. For the IS search queries, we added 'management' because information systems and management have overlapping publication venues. In fact, the most reputable IS venue is called *Management Information Systems Quarterly* (see the rankings linked from the inclusion criteria listed below). We validated this assumption with an exploratory literature review of trust-related articles in those disciplines before we moved to the systematic approach described in this section. Since our goal was to apply existing trust models to AI assistants, we further retrieved articles for the same queries, adding:

```
(ai OR genai OR artificial intelligence)
```

For each discipline, we created two lists based on the queries mentioned above (e.g., HCI and HCI-AI, see Table II) and retrieved 50 articles from each list, selecting those with the highest search rank as of March 2025, resulting in a total of 300 articles to review. We then identified and marked duplicates across the lists. We assigned one author to each discipline who then reviewed the articles according to the inclusion and exclusion criteria we defined. We organized weekly meetings to discuss unclear cases and applied the following criteria for the literature review across disciplines, which we conducted to address **RQ1** and **RQ2**:

- **Include** papers published in the target discipline (use common rankings such as *ICORE*[4] for SE or the *ABDC JQL*[5] for IS; do not filter by rating, only exclude venues that are not listed in discipline-specific rankings).
- **Include** papers published in the main or research tracks of conferences and regular journal papers. Include literature reviews and special issues on human trust (analyzed separately, see Section III-B4).
- **Exclude** other editorials, short papers (<5 pages), vision papers, textbooks, and dissertations.
- **Exclude** papers that do not discuss human trust but trust in a technical sense (e.g., in a security context).

For SE, we included 13 unique primary research papers (one paper was present in both lists), seven literature reviews (of which two were also part of the IS list; only three were SE-specific), and one special issue that overlaps with IS. For HCI, we included 22 unique primary research papers (five papers were present in both lists) and five literature reviews (two present in both lists; two overlapping with IS). For IS, we included 24 unique primary research papers (no overlap), four literature reviews (two also present in the SE lists, three without a strong IS focus), and three special issues on trust (one overlapping with SE). Among the excluded SE papers were two HCI papers, which we added to the list of HCI papers (hence we have 24 HCI research papers in total). Similarly, the HCI list contained one IS paper, which we added to the list of IS papers (hence we have 25 IS research papers in total). Across the SE, HCI, and IS lists, we identified 15 articles that were published in philosophy venues (including ethics and fairness) or psychology venues. We read those papers and either integrated into the background section (see Section II) or the discussion section (see Section IV).

Table II lists all the included papers. Row PH+PS lists all philosophy and psychology papers that we integrated as mentioned above. For the other papers, we downloaded the corresponding PDF files, searched for all occurrences of the substring "trust," and extracted all paragraphs that conceptualized trust either directly or indirectly. We also kept track of which trust models the papers referred to (if any) and whether they customized (adapted or extended) existing trust models. We report the results of this analysis in Section III-B.

Many of the excluded SE papers were either not full papers or published in non-SE venues. This, in conjunction with the fact that only few SE papers contained explicit conceptualizations of trust, contributed to the decision to follow up with a critical review of recently published articles on trust in leading SE venues (see below).

2) Critical review of recent SE literature: A critical review focuses on evaluating key issues and analyzing a sample of studies with shared characteristics [12]. The issue we intended to evaluate is the lack of trust conceptualization in SE research (related to **RQ2**). To this end, we used *dblp SPARQL*[6] to retrieve all articles that met the inclusion/exclusion criteria:

- **Include** papers in leading SE conferences (ICSE, ASE, FSE) and journals (TSE, TOSEM) published recently, i.e., just before or after the GenAI hype (2022-2024).
- **Include** papers published in the main or research tracks of conferences and regular full journal papers.

[3]https://github.com/scholarly-python-package/scholarly
[4]https://portal.core.edu.au/conf-ranks/
[5]https://abdc.edu.au/abdc-journal-quality-list/
[6]https://sparql.dblp.org/

TABLE I
COMMON TRUST MODELS ACROSS DISCIPLINES.

Trust Model	Summary	HCI	HCI-AI	IS	IS-AI
Mayer et al. (1995) [92]	Multi-disciplinary model of organizational trust.	[28], [33], [50], [106], [111], [114], [134]	[15], [50], [111], [134]	[81], [133], [139], [161]	[7], [36], [123], [141]
Lee&See (2004) [78]	Trust in automation systems and technology.	[28], [106], [111], [134], [151]	[15], [111], [134], [137], [159]	[133]	[7], [141]
McKnight et al. (2011) [95]	Model of trust in technology grounded in system-level traits	[50], [134]	[50], [134]	None	[141]
Gulati et al. (2019) [50]	Empirically validated model for measuring intelligent systems trust.	[111], [134]	[111], [134]	None	None
McAllister (1995) [93]	Dual-factor model of organizational interpersonal trust.	[106], [114]	[15]	None	None
Rousseau et al. (1998) [115]	Integration of organizational trust across disciplines.	[28], [114]	None	[32]	None
Madsen&Gregor (2000) [88]	Human-computer trust model, focus on affective & cognitive trust.	[76], [106]	[15]	None	None
McKnight et al. (2002) [97]	Model of initial trust in e-commerce settings.	None	None	[32], [55], [80], [81], [140]	[147]
Gefen et al. (2003) [45]	Model of trust formation in e-commerce settings.	None	None	[55], [133], [138]	[145]

- **Exclude** editorials, special issues, short papers, vision papers, papers from co-located events, etc.

For the 2670 papers that met our inclusion criteria, we retrieved all abstracts from the ACM and IEEE digital libraries and searched for the substring "trust" in the title and abstract. Of the 2670 papers, 58 mentioned trust. We downloaded the corresponding PDF files and manually searched for and extracted implicit and explicit conceptualizations of trust. We report the results of this analysis in Section III-C.

3) Data availability: We provide all retrieval and analysis scripts and all metadata we retrieved from *Google Scholar* and *dblp* as part of our supplementary material [13]. Moreover, we share the annotated lists of papers, including exclusion reasons, and all extracted excerpts. All included research papers are referenced in Table II.

B. Trust Conceptualizations across Disciplines

In this section, we report the results of our literature review across disciplines. Note that with this review, our goal was not to achieve a complete coverage of all articles that describe conceptualizations of trust in the target disciplines. Our goal was to identify whether articles in the selected disciplines commonly refer to established trust models, with the goal of comparing SE articles to related disciplines. Although we acknowledge this as a limitation of our research, the difference between SE articles and HCI/IS articles was so clear that our work nevertheless provides an important contribution to the field of SE. Table I lists the trust models that at least two articles referred to. The models proposed by Mayer *et al.* [92] and Lee and See [78] are used across HCI and IS, while McKnight *et al.* [97] and Gefen *et al.* [45] are IS-specific. The other models listed in the table are primarily referenced in HCI, but less frequently than the Mayer *et al.* [92] and Lee and See [78] models.

*1) **Information Systems and Management (IS)**:* In IS research, trust has been studied since the 1990s [132]. The articles we included in our review of IS literature were published in the time span from 2000 to 2024. Of the 24 research papers we analyzed, only four did not embed their studies in existing trust models. It was common for IS papers to customize existing trust models (13 out of 25 articles).

TABLE II
INCLUDED ARTICLES OF OUR LITERATURE REVIEW; CR: CRITICAL REVIEW, ADDED: FROM SEARCH RESULTS OF OTHER DISCIPLINE.

Area	Research Articles	Literature Reviews	Special Issues	Added
HCI	[1], [27], [28], [33], [50], [59], [65], [76], [106], [111], [114], [128], [134]	[8], [49], [63], [102]	None	[151]
HCI-AI	[6], [15], [28], [44], [50], [58], [59], [111], [128]–[130], [134], [137], [155], [159]	[8], [28], [99]	[67]	[127]
IS	[5], [32], [48], [55], [80], [81], [90], [107], [133], [138]–[140], [161]	None	[17], [18]	None
IS-AI	[7], [35], [36], [52], [62], [74], [119], [123], [141], [143], [145]	[47], [69], [86], [157]	[87]	[147]
SE	[16], [20], [21], [23], [29], [53], [61], [64], [70], [83], [144], [150]	[26], [57]	None	None
SE-AI	[3], [64]	[69], [79], [84], [86], [162]	[87]	None
SE-CR	[4], [22], [75], [105], [153]	None	None	None
PH+PS	[30], [38], [54], [60], [82], [91], [104], [109], [110], [116], [124], [149], [152], [156]	None	None	None

Several articles discussed that "*trust is a dynamic concept that develops over time*" [81]. The factors and processes that lead to *trust formation* change gradually [92], and can include aspects such as people's perceptions about a company or brand [133]. This means that trust in a company can transfer to trust in a particular tool, a concept called *trust* *transference* [136], [148]. Saffarizadeh *et al.* applied this concept to AI agents and noted that "*trust in an AI creator can be transferred to its AI agent when there is a perceived meaningful association between them*" [119].

IS articles clearly distinguish *initial trust*, which is important to "*create a trustworthy first impression of a new system and encourage adoption*" [81], from the before-mentioned process of trust formation and later stages such as *knowledge-based trust*, "*where the individual knows the other party well enough to predict the party's behavior in a situation*" [141]. From a tool perspective, the later phases of trust are sometimes referred to as "*postadoption*", where aspects such as *predictability* and *helpfulness* of a tool [113] play an important role [140]. McKnight *et al.*'s model is frequently referred to as a common IS model for initial trust, whereas Gefen *et al.*'s model focuses on trust formation. In McKnight *et al.*'s model, based on Mayer *et al.*, trust in vendors of e-commerce websites is a multidimensional construct centered around *trusting beliefs*, that is, "*perceptions of the competence, benevolence, and integrity of the vendor*" and *trusting intentions*, that is, the "*willingness to depend*" or, in other words, the "*decision to make oneself vulnerable to the vendor*" [97]. Gefen *et al.* models consumers' online purchase intentions as a combination of perceived usefulness, ease-of-use, and trust in the vendor [45]. They show that online trust is built through "*(1) a belief that the vendor has nothing to gain by cheating, (2) a belief that there are safety mechanisms built into the web site, and (3) by having a typical interface, (4) one that is, moreover, easy to use.*"

An important observation by Mayer *et al.* is repeated in IS articles: humans differ in their general tendency to be willing to trust others [81], which is sometimes referred to as their *propensity* [32], [92] or *disposition* to trust [80]. These individual differences are a potential confounding factor in studies with developers investigating trust in AI assistants.

Two other central constructs are *trusting intention*, which is "*the trustor's willingness to depend on the trustee*" [81] and a person's *trusting beliefs* [141], that is "*favorable object-specific beliefs*" [140] that capture "*the trustor's perception of whether the trustee has the desired attributes to be trusted*" [81]. IS literature also discusses how the trusting beliefs proposed by Mayer *et al.* (benevolence, ability, and integrity) can foster technology acceptance [19]. Troshani *et al.* further categorize those trusting beliefs into *cognitive trust* and *affective trust* [145] (see also Komiak and Benbasat). Cognitive trust captures "*a customer's confidence or willingness to rely on a service provider's competence and reliability*" while *affective trust* is the "*confidence that one places in another party on the basis of feelings generated by the level of care and concern as demonstrated by the other party.*" In summary, competence (i.e., ability in Mayer *et al.*'s model) and predictability influence cognitive trust while benevolence and integrity influence affective trust [145].

The IS literature discusses the validity of applying models of human trust to technology [32], [81], [119], which is also discussed in philosophy as the *plausibility of trust* (see Section II-C). An aspect to consider is that humans have long been known to anthropomorphize technology [103]. To study trust relationships between humans and human-like tools, researchers have adapted the concept of *interpersonal trust* to technology, for example, to study trust in automation or autonomous systems (see the description of Lee and See's trust model in Section III-B2). However, researchers have also argued that such models do not necessarily apply to GenAI tools due to their nondeterminism [7].

Articles further discuss that increasing automation and complexity leads to systems being perceived as opaque and sophisticated, hence trust becomes more and more important [78]. This situation leads to uncertainty and a *perceived risk* of using an automation tool, which could hinder adoption [25], [140]. However, Chopra and Wallace also write that "*trust can only arise when there exists a state of dependence between the trustor and trustee, and when acting on this dependence entails risk, i.e., the trustor possesses uncertainty about the outcomes and vulnerability to a potential loss if the outcomes are undesirable*"[32]. This is certainly true for software engineers who use GenAI tools to develop software. Again, the role of risk and vulnerability appears also in the psychology literature, in particular Mayer *et al.*, as well as in the philosophical debate, which we have summarized in Section II-B.

Two opposing situations in the context of AI for decision-making are *mistakenly denied trust* and *unfounded trust* [123]. Schmidt *et al.* summarize the situation as follows: "*If humans increasingly leverage AI to inform, derive, and justify decisions, it also becomes important to quantify when, how, why, and under which conditions they tend to overly trust or mistrust those systems*" [123]. This view corresponds to the concepts of *undertrust* and *overtrust* as described by Lee and See [78].

Regarding *trust quantification*, Schmidt *et al.* note that "*trust is measured differently in different research fields*" [123]. However, they also mention that a common approach used across disciplines is to ask study subjects about their *trusting beliefs* or *trusting intentions* towards a certain entity, for which there are standardized survey instruments (see, e.g. [46]). Validated measures exist also to assess subjects' *disposition to trust* [96]. Of course, such instruments cannot capture the actual critical situations in which a study subject made itself "*vulnerable to the actions of another party*" [92]. Therefore, answering survey questions can never fully capture trust. Schmidt *et al.* mention *behavioral trust* as a stronger indicator, referring to situations such as parents actually leaving their children with a certain babysitter as an indicator of trusting the babysitter [123]. However, they also mention that human behavior and decision-making are complex and affected by factors such as monetary constraints or lack of alternatives. Therefore, researchers must be careful when using behavior as a proxy for trust, considering that there might be multiple drivers for the observed behavior.

For decision-making, an exemplary operationalization of trust based on observed behavior is calculating the probability that users follow a model's predictions based on usage data [112], [122], which is closely related to our motivating example of software developers accepting code suggestions. Schmidt *et al.* argue that this operationalization might work well as a behavioral proxy of trust when comparing different treatments (e.g., different ways of presenting AI suggestions) or comparing different groups based on the same treatment [123]. However, factors such as participant knowledge and confidence, and (perceived) task difficulty are known to affect advice acceptance [158]. Moreover, over time, the usage of tools moderates the bias of users towards self-reliance in decision-making and could further influence the observed behavior [39].

*2) **Human-Computer Interaction (HCI)**:* Trust has been studied in HCI since the late 1980s [100]. Our review of HCI literature includes articles published between 1996 and 2024. Among the 24 unique HCI research papers we included, eight articles embedded their trust definitions in Mayer *et al.*'s model. Three articles explicitly cite the definition, one paraphrases it, and four reference components such as *ability*, *integrity*, or *benevolence* without adopting the complete model. Furthermore, nine papers refer to Lee and See's trust model, and another five papers incorporate elements from both models, showing an overlap in conceptual foundations. Among the twelve papers building upon existing trust models, six extended the models either by developing new models or by adopting the definitions and further exploring specific aspects of trust. The remaining six papers refer to the models without modification. In contrast, eleven papers do not explicitly refer to existing trust models.

Several HCI articles conceptualize trust as a *context-dependent phenomenon* shaped by system behavior, user interaction, and situational factors. This perspective is especially prevalent in domains such as AI agents, automated decision support systems, and human-robot interaction. For example, Sousa *et al.* [134] propose the *human-centered trustworthy framework*, which adapts Gulati *et al.*'s human–computer trust model [50] and integrates trustworthiness components from Mayer *et al.* [92], along with constructs from McKnight *et al.* [95] and Lee and See [78]. In this framework, trust is approached as a multilevel concept that includes *user predispositions*, *interface features* such as transparency and control, and *institutional factors* such as data governance and accountability. This layered view underscores the relational and evolving nature of trust in human–AI interaction.

Pinto *et al.* [111] developed and validated a human–robot interaction trust scale by adapting Gulati *et al.*'s model [50]. Trust is conceptualized following Mayer *et al.* [92]'s definition as the *willingness to be vulnerable to another party*, and is further informed by Lee and See's view of *trust development over time* [78]. Their adapted model incorporates dimensions such as *reciprocity*, *competence*, *benevolence*, *predictability*, *honesty*, *trust predisposition*, and *willingness to be vulnerable*, tailored to human–robot interaction.

Nothdurft *et al.* [106] conceptualize trust as a multidimensional and dynamic construct that evolves through interaction. They draw on Mayer *et al.* [92] to describe trustworthiness in terms of *ability*, *integrity*, and *benevolence*, and incorporate Lee and See's view of trust in automation [78] as a *calibrated attitude* shaped by system performance and user experience. Their conceptualization also reflects Madsen and Gregor's adaptation [88] of McAllister's definition [93], which emphasizes user confidence and willingness to act on the recommendations of intelligent systems. Building on this foundation, Nothdurft *et al.* [106] propose an adaptive explanation architecture that adjusts system transparency and feedback to support appropriate levels of trust in automated systems.

Riegelsberger *et al.* [114] conceptualize trust as a *psychological state* characterized by a *positive expectation* that one's *vulnerability* will not be exploited. Like Nothdurft *et al.* [106], they build on Mayer *et al.* [92] and McAllister [93], but extend the discussion by emphasizing the *social grounding* of trust. Their framework is embedded in a trustor–trustee model adapted from the *trust game* [9], where trust is shaped by *contextual cues*, *reputation*, and *internalized norms*. They argue that trust in mediated interactions is not only influenced by *individual traits* but also by *perceived trustworthiness* of systems shaped through interface design and social context.

In contrast, several papers reference foundational trust models without adapting or extending them. Banovic *et al.* [15] cite Mayer *et al.* [92], McAllister [93] and Madsen and Gregor [88] to examine how untrustworthy AI systems can still elicit user trust, but do not propose a new framework or modify the referenced models. Similarly, Yokoi *et al.* [159] apply Lee and See [78] definition to investigate trust in AI-based medical decision-making, without extending it conceptually. Suen and Hung [137] adopt Lee and See's trust model to frame user trust in AI-based video interviews. Although the model is applied to a specific domain, its theoretical structure remains unchanged. Kraus *et al.* [76] primarily build on Madsen and Gregor [88] trust model, while only briefly mentioning Mayer *et al.* and Lee and See. These models are not embedded in their conceptual framework and rather serve as peripheral citations. Similarly, Cai *et al.* [28] adopt Mayer *et al.*'s and Lee and See's definitions of trust to conceptualize user trust in human–computer interaction, but use the models primarily to validate the Chinese version of the human–computer trust scale in this new cultural context.

*3) **Software Engineering (SE)**:* Trust is a critical element in SE that influences the adoption of software tools, processes, and collaborative practices. However, our review of 13 articles published between 2002 and 2024 reveals that SE research typically does not engage with established foundational trust models. In fact, only one paper [61] cited Mayer *et al.*'s trust definition and based their research on a related trust model. Furthermore, one paper [150] implicitly referred to established trust models by mentioning *risk*, that is, the aspect of *vulnerability* (see Section II-A). The other papers either did not explicitly define trust or used custom definitions, diverging from disciplines such as IS and HCI, where established trust models are more commonly applied.

A key example of SE's customized approach to trust is the PICSE framework proposed by Johnson *et al.* [64]. It describes a structured approach to understanding trust in software tools, identifying five key factors: *personal*, *interaction*, *control*, *system*, and *expectation*. The authors refer to Fogg and Tseng's credibility framework (HCI) [42] and Rousseau *et al.*'s trust literature review (IS) [115]. However, these trust models did not contribute to the PICSE framework—they are only discussed retrospectively.

Jalali *et al.* adapted Schultz's situational trust model [125], which extends Mayer *et al.*'s model, to analyze trust dynamics in global software engineering [61]. In line with HCI and IS research, their model underlines the dynamic nature of trust, distinguishing *initial trust building*, where trust forms based on past interactions, from *trust evolution*, where trust fluctuates during the project based on expectations and behaviors.

Wang and Redmiles applied evolutionary game theory to examine trust and cooperation in distributed teams. Their study does not explicitly model trust development or draw on established trust theories [150]. Similarly, Calinescu *et al.* propose dynamic assurance cases for engineering trustworthy self-adaptive software, emphasizing system reliability and correctness rather than interpersonal trust [29]. Akbar *et al.* take a different approach, presenting a taxonomy of decision-making challenges in trustworthy AI [3]. Although these studies address trust-related concerns, they do so in domain-specific ways, emphasizing technical, ethical, or structural factors rather than explicitly modeling trust development based on established theories.

Bratthall and Jørgensen address trust in single data source software engineering case studies [23]. Bertram *et al.* investigate *trust perceptions* in code review through an eye-tracking study, demonstrating how the perceived origin of code (human versus machine-generated) influences review behaviors and trust judgments [20]. However, both papers do not explicitly reference or build upon established trust models.

*4) **Literature Reviews and Special Issues**:* Among the included papers were eleven literature reviews. In addition, we found three editorials for special issues on trust. Describing these articles in detail is beyond the scope of this paper. However, we want to briefly summarize them because they reflect the scientific discourse on trust.

We found five literature reviews of trust that were **not limited to one of our target disciplines** SE, HCI, or IS. Kaur *et al.*'s review centers around requirements for trustworthy AI [69]. They refer to trust definitions across disciplines, including psychology, sociology, and economy, mentioning *integrity* and *reliability* as an "*agreement*" across these disciplines. They define trustworthy AI as a "*framework to ensure that a system is worthy of being trusted based on the evidence concerning its stated requirements*" [69]. Yang and Wibowo examined 142 studies published between 2015 and 2022, identifying components, influencing factors, and outcomes of users' trust in AI [157]. Li *et al.*'s review outlines aspects of AI trustworthiness such as *robustness*, *explainability*, *transparency*, and *fairness*, organized along the lifecycle of AI systems [79]. However, they do not discuss established trust definitions or models. Gulati *et al.* reviewed 47 studies on trust in technology published in IS and HCI venues and found 17 different theories that were integrated from disciplines such as psychology and economics [49]. They conclude that "*the intricacies of how trust is formed and maintained in online environments still necessitates further investigation*" and that the development of standardized and empirically validated instruments for trust measurement is crucial. Finally, Lockey *et al.*'s review focuses on *antecedents of trust* in AI [86]. They identify five AI-specific trust challenges: *transparency and explainability*, *accuracy and reliability*, *automation versus augmentation*, *anthropomorphism and embodiment*, and *mass data extraction* [86]. An interesting observation regarding AI assistants for software development is that "*over-anthropomorphism may lead to overestimation of the AI's capabilities*" [86], increasing risk [34], decreasing trust [11], and potentially leading to manipulation [120]. This is certainly an underexplored perspective on AI assistants for software development, especially when focusing on human factors of software security.

Three literature reviews focused on trust research in **HCI**. Jeon's review discussed "*the effects of emotions on trust in the context of technology use*" [63]. Their review indicates that positive emotions lead to higher trust. Bach *et al.* reviewed 23 empirical studies that focus on user trust in AI-enabled systems [8]. They describe trust definitions, factors that influence user trust, and measurement methods, reporting that seven studies explicitly defined trust, six of which referring to Mayer *et al.* [92] or Lee and See [78]. Eight studies conceptualized trust, while nine studies neither defined nor conceptualized trust. They found surveys, interviews, and focus groups to be the most commonly used methods to assess user trust. Bach *et al.* conclude that it is important to select a definition of user trust that aligns with the specific study context. Mehrotra *et al.* reviewed 65 articles to assess the various definitions of appropriate trust in human-AI interaction, providing an overview of concepts and definitions and identifying similarities and differences between them [99]. They conclude that four common methods for building appropriate trust are improving *system transparency*, considering user *cognition and perception*, using *guidelines or models* to achieve calibrated trust in AI, and considering the entire *continuum of trust*, including *over-, under-, mis-, and dis-trust*.

One literature review had a clear **IS** focus. Glikson and Woolley synthesized two decades of empirical research on the determinants of human trust in AI across disciplines [47]. The authors motivate how trust in AI differs from other technologies and identify an AI's capabilities as an important antecedent of trust development. In addition, they propose a framework that identifies factors that influence users' *cognitive* and *emotional* trust. Regarding the development of *cognitive trust*, they highlight the importance of an AI's *tangibility*, *transparency*, *reliability*, and *immediacy*. For *emotional trust*, on the other hand, and AI's *anthropomorphism* is a major factor. Glikson and Woolley also critically reflect on past studies by pointing to the diversity of trust measure used and the lack of longer-term studies in "*higher stakes environments*" [47].

For IS, we further identified three special issues on trust. One of them was published in 2010 in their flagship journal *MIS Quarterly*, focusing on "*novel perspectives*" on trust in IS. The fact that this special issue was published 15 years ago is another indication of the maturity of trust research in that discipline. The other IS-specific special issues focused on trust in online environments [18] and trust in AI [87].

For **SE**, we included two literature reviews. Liu *et al.* conducted a tertiary study reviewing 141 secondary studies on trustworthy AI and software, describing trustworthiness as a "*highly abstract concept comprising related quality at-*

tributes" [84]. The review focuses on these quality attributes and compares them between trustworthy AI and other trustworthy software. The conceptualization of trust and trustworthiness is not in focus. Finally, Hou and Jansen reviewed 112 articles on trust in software ecosystems, identifying different definitions of concepts such as *software trust*, *system trust*, or *software service trust*. However, their focus was on compiling definitions of specific trust forms, not unifying or aligning them [57].

C. Critical Review of Trust Conceptualizations in Software Engineering Research

Of the 58 articles that we analyzed as part of our critical review, only one article referred to an established trust model, and only three articles provided trust definitions. This underlines the problematic situation suggested by the literature review presented in Section III-B3. There is a significant gap in the maturity of trust research in SE compared to IS and HCI. Trust as a concept is often discussed in a context-specific manner without embedding concepts and definitions in established trust theories or models. This lack of embedding in established trust theories from other disciplines hinders interdisciplinary comparisons and research towards a deeper understanding of what constitutes trust in SE-specific settings.

Noller *et al.* are the only authors who explicitly reference an established trust model, incorporating Lee and See's model of human-automation trust [105]. Alami *et al.* define trust as "*the unyielding belief that the person is truthful and reliable*" and in the context of their research as "*the willingness of the community to rely on the contributor, a prerequisite for considering her code change*" [4]. This conceptualization frames trust as a social and reputation-based construct, essential in collaborative environments like open source software projects, where trust is built through ongoing contributions and credibility. Although this framing is reasonable, an additional embedding in existing trust models would allow researchers to work out the particularities of trust in open-source projects in comparison to trust in other settings.

Winter *et al.* [153] implicitly define trust by referencing dispositional trust as proposed by Hoff and Bashir [56], which describes an individual's inherent tendency to trust automation, independent of context or system performance. This positions trust as a psychological trait that varies between individuals. However, it is important to note that trust was only a peripheral topic, not a central focus of their study. Kou *et al.* discuss trust in the context of large language models (LLMs) for code generation, suggesting that users assess trustworthiness based on the model's explainability and how well its predictions align with their expectations [75]. However, they neither explicitly define trust nor consider the broader explainability discussion (see Section II). Instead, they cite previous work by Boggust *et al.*, who explored how users determine whether to trust a machine learning model based on its ability to convey behavior in a way that aligns with their expectations [22]. Neither study provides a formal definition of trust; rather, both frame it in terms of user perceptions and interpretability. Given the amount of related work on trustworthiness and explainability of AI-based systems in philosophy (see Section II-C) and IS (see Section III-B1), it is surprising that this article is again not broadly embedded in established trust conceptualizations.

TABLE III
SUMMARY OF THE TENETS OF TRUST AS DISCUSSED IN PHILOSOPHY AND PSYCHOLOGY COMPARED TO SOFTWARE ENGINEERING, WHERE TRUST HAS BEEN OPERATIONALIZED AS ARTIFACT ACCEPTANCE.

Tenets of trust in philosophy (PH) and psychology (PS)	**Operationalizing trust as artifact acceptance...**
Trust is an attitude, not a behavior (PH, PS).	...treats trust as a behavior.
Trust is context-dependent (PH, PS).	...just focuses on one context.
Trust should be well-calibrated or well-grounded (PH, PS).	...does not measure calibration or grounding.
Trust should have a high resolution (PS).	...does not measure resolution.
Trust should have a high specificity (PS).	...could at most measure low specificity.
Trust should be justified (PH).	... ignores justification.
Trust should be plausible (PH).	... ignores plausibility.

IV. DISCUSSION: RETHINKING TRUST IN AI ASSISTANTS

To answer **RQ3**, we want to return to the title of our paper, that is, **rethinking trust in AI assistants for software development**. Our research was motivated by the fact that several SE research articles suggest approaches to increase 'trust' with the aim of increasing tool adoption (see Section I). Although previous research in other fields such as IS suggests that trust is indeed related to tool adoption and usage (see Section III-B1), the corresponding IS research articles thoroughly define and conceptualize trust and related concepts such as *trust transference*, *trusting beliefs*, or a user's *disposition to trust*. These conceptualizations originate in mature trust models from psychology, in particular the one proposed by Mayer *et al.* in 1995 [92]. However, as Thiebes *et al.* note, "*trust in general is a complex phenomenon that has sparked many scholarly debates in recent decades*", therefore "*it is not surprising that the conceptualization of trust in AI and what makes AI trustworthy—as of today—remains inconclusive and highly discussed in research and practice*" [141]. It is unfortunate that the SE research community is not a major contributor to that interdisciplinary discourse because, as our review of SE literature has shown, SE research (and practice, see [110]) on trust mainly uses ad hoc trust definitions without embedding studies in established trust models.

The detachment of SE research from established trust models is problematic, as evidenced by the equation of trust with the acceptance of generated content, which is shortsighted for several reasons. Drawing on the foundations of trust from psychology and philosophy (see Section II), Table III shows how such an operationalization disregards the tenets of trust discussed in these disciplines. First, both disciplines see trust as an *attitude* and not a *behavior*. Therefore, measuring acceptance disregards factors that influence a person's behavior, such as their situation (e.g., time pressure), the type of AI system they are interacting with, their specific task, the potential consequences of decisions, and many more. Still, even if acceptance measured trust, it would not measure *calibrated* or *warranted trust*. Hence, the measured trust could be completely detached from the actual capabilities of the AI. However, given the fact that GenAI-based software development assistants will continue to hallucinate [14], the potentially resulting maintainability challenges [85], security risks [108], and negative impact on learning [126], [156], it would not only be morally but also pragmatically better to find out whether developers' trust in such tools is actually calibrated or warranted. After all, maintainability issues or security vulnerabilities introduced by relying on generated code could cause major financial and reputational damage to individuals and companies.

It is still useful to record statistics such as acceptance rates and make them widely available to researchers. Currently, AI tool vendors basically have a monopoly on fine-grained usage data in real-world situations. Studying such data in detail is crucial, as acceptance of AI suggestions can be a factor by which people form their trust; that is, in Mayer *et al.*'s model, acceptance rates can contribute to the *perceived ability* of the system. For example, Wang *et al.* developed and tested design concepts, including a dashboard showing a user's personal GitHub Copilot acceptance rate, finding that it helped developers align their expectations with an AI assistant's ability [149].

In general, the trust dimensions proposed by Mayer *et al.* [92], that is, *ability*, *benevolence*, and *integrity* can be easily transferred to AI assistants for software development. As mentioned before, *ability* requires AI assistants to generate correct and reliable code that is appropriate for a given context. *Benevolence* reflects interaction patterns that AI vendors implement to ensure that AI assistants are perceived as supportive, and *integrity* models the data curation and guardrails that AI vendors and regulatory bodies implement to avoid biases or problematic code that, for example, contains security vulnerabilities. Concerning the last two points, there is now more and more research that argues that, for a system to be adequately trusted, it is not just the system that needs to function correctly. Rather, for a system to be perceived as benevolent and integer, there must be regulation, auditing processes, and established standards that govern the entire engineering process [89], [109], [152]. Winfield and Jirotka, for instance, argue that ethical governance, consisting of ethical codes of conduct, ethics training, responsible innovation, and transparency, is essential to build trust in AI systems [152]. Similarly, Manzini *et al.* argue that user trust requires evidence on the functionality of an AI assistant at the level of its design, the practices of the developing organization, and oversight by external bodies [89].

On the other hand, it is interesting to see how IS and HCI primarily focus on the psychological side of trust. We posit that discussion in SE can also benefit from incorporating philosophical insights on trust. For example, the discussion as to whether trust in AI is even conceptually possible is relevant in the sense that it critically questions whether engineering for trust is not just a marketing phenomenon or, in the worst case, even ethicswashing (see Section II-C). A similar phenomenon can be observed in the debates on XAI, where the expectation is that XAI invariably leads to greater trust, something that could not be proven in studies [68]. In this sense, the concept of *justification* is also important for SE, as it can help ensure that trust in AI assistants for software development is formed appropriately, rather than, for example, being the result of misleading marketing rhetoric.

It is surprising that the trust conceptualizations in SE are still so underdeveloped, as tools from SE can be extremely important for creating trustworthy systems. For example, Ahuja *et al.* argue that SE processes and tools, in particular a detailed requirements analysis, monitoring instruments, and automated testing, are essential to create systems that can really be trusted [2].

We close our discussion by pointing to potential trust instruments and measurements that could be adapted in SE research. As already mentioned, operationalizing trust is difficult; one needs to be specific about the actual concepts being captured (e.g., *trusting beliefs* or the general *propensity to trust*). SE researchers can use validated instruments from other disciplines (see Section III-B1), but should keep in mind that focusing on individual aspects, such as *trusting beliefs* or *behavioral metrics*, is usually not sufficient.

V. Threats to Validity

As with any empirical study, our methodology is subject to certain limitations that may affect the completeness, accuracy, and generalizability of our findings. In the following, we identify and discuss potential threats to the validity of our methodology, including biases in paper selection, interpretation of trust conceptualizations, and the applicability of findings across disciplines. We classify the validity threats in three categories as proposed by Wohlin *et al.* [154].

Internal Validity concerns whether our study design and execution introduced biases that may affect the correctness of our findings. A potential threat to internal validity is that we may have missed indirect or implicit definitions of trust as well as references to trust models during our literature review. Additionally, by focusing on the highest-ranked articles from our queries, we may have unintentionally excluded relevant papers ranked lower. We acknowledge the selective inclusion based on an opaque ranking algorithm provided by Google Scholar as a limitation. However, Google Scholar is used by academics across disciplines and allowed us to use the same sampling approach for all disciplines. Future work should extend our literature review by using discipline-specific retrieval approaches. In addition, we complemented the review of SE literature by conducting a critical review of recently published SE papers, which confirmed the results of our review based on Google Scholar search results.

External Validity concerns the generalizability of our findings. Our literature review focused on SE, HCI, and IS; our description of the foundations of trust is rooted in philosophy and psychology. We did not cover trust conceptualizations in other disciplines, such as sociology. However, given that our literature review primarily revealed trust models rooted in psychology, these foundations seem to be the most relevant ones for the considered disciplines. As we only considered English-language publications, our findings may not reflect cultural variations in trust.

Construct Validity concerns whether we accurately captured and analyzed the concept of trust as intended. Given the complexity of trust and its multidimensional nature, one potential threat is our own limited understanding of the concept. To mitigate this, we recruited coauthors from philosophy and HCI and collected feedback from another colleague with a background in psychology and IS.

VI. Conclusion

Our review of SE, IS, and HCI literature has shown that SE articles, contrary to IS and HCI, often use the term trust informally and that providing an explicit definition of trust or embedding results in established trust models is rare. However, without a common definition, true secondary research on trust is impossible. Moreover, as discussed above, too simplistic operationalizations of trust are problematic as well. Therefore, to conclude our paper, we provide actionable recommendations for SE researchers planning to study trust in AI assistants. Our recommendations are based on a thorough description of the foundations of trust (Section II) and a literature review across disciplines (Section III-B).

1) Before designing concrete studies on trust an AI assistants, carefully read the literature on **trust beyond SE**. We hope that our article serves as a useful starting point for this.
2) When reporting the study design and results in papers, embed design and results in **established trust models** (see Table I); customize them only if necessary.
3) Use **established terminology** when discussing trust, for example, distinguishing *overtrust*/*undertrust* from *calibrated trust* or distinguishing *initial trust* from *trust formation* (see Section II and Section III-B1).
4) Do not equate artifact acceptance with trust. When designing empirical studies to assess developers' trust in AI assistants, clearly outline which **aspects of trust** are being measured using which instruments (e.g., *cognitive trust* versus *affective trust*). Use **validated instruments** whenever possible (see trust operationalizations discussed in Section III-B1) and openly discuss their limitations.
5) Consider participants' *trust disposition* and *trusting beliefs* as **confounding factors**; consider the *transferable* nature of trust. For example, trust in a particular AI vendor can propagate to a specific AI assistant that this vendor offers.
6) Clearly describe the **study context** and reflect on its generalizability. Given how central *risk* and being *vulnerable* are for trusting behavior, a study in a low risk scenario might not generalize to a high risk situation, for example, in the context of safety-critical software.

We hope that our paper contributes to more mature and interdisciplinarily embedded SE trust research in the future.

References

[1] E. Adar, D. S. Tan, and J. Teevan, "Benevolent deception in human computer interaction," in *ACM CHI 2013*, 2013.

[2] M. K. Ahuja, M. Belaid, P. Bernabé, *et al.*, "Opening the software engineering toolbox for the assessment of trustworthy AI," in *NeHuAI 2020*, ser. CEUR Workshop Proceedings, vol. 2659, 2020.

[3] M. A. Akbar, A. A. Khan, S. Mahmood, S. Rafi, and S. Demi, "Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges," *Softw. Pract. Exp.*, vol. 54, no. 9, 2024.

[4] A. Alami, R. Pardo, M. L. Cohn, and A. Wasowski, "Pull request governance in open source communities," *IEEE Trans. Software Eng.*, vol. 48, no. 12, 2022.

[5] D. K. Allen, D. Colligan, A. Finnie, and T. Kern, "Trust, power and interorganizational information systems: The case of the electronic trading community translease," *Inf. Syst. J.*, vol. 10, no. 1, 2000.

[6] S. Amershi, D. S. Weld, M. Vorvoreanu, *et al.*, "Guidelines for human-ai interaction," in *ACM CHI 2019*, S. A. Brewster, G. Fitzpatrick, A. L. Cox, and V. Kostakos, Eds., 2019, p. 3.

[7] O. Asan, A. E. Bayrak, and A. Choudhury, "Artificial intelligence and human trust in healthcare: Focus on clinicians," *J. Med. Internet. Res.*, vol. 22, no. 6, 2020.

[8] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. C. Sousa, "A systematic literature review of user trust in ai-enabled systems: An HCI perspective," *Int. J. Hum. Comput. Interact.*, vol. 40, no. 5, 2024.

[9] M. Bacharach and D. Gambetta, "Trust in signs.," in (Russell Sage foundation series on trust, vol. 2.), Russell Sage foundation series on trust, vol. 2. New York, NY, US: Russell Sage Foundation, 2001.

[10] A. Baier, "Trust and antitrust," *Ethics*, vol. 96, no. 2, 1986.

[11] A. L. Baker, E. Phillips, D. Ullman, and J. R. Keebler, "Toward an understanding of trust repair in human-robot interaction: Current research and future directions," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 4, 2018.

[12] S. Baltes and P. Ralph, "Sampling in software engineering research: A critical review and guidelines," *Empir. Softw. Eng.*, vol. 27, no. 4, 2022.

[13] S. Baltes, T. Speith, B. Chiteri, S. Mohsenimofidi, S. Chakraborty, and D. Buschek, *Rethinking trust in ai assistants for software development: A critical review (supplementary material)*, Zenodo, Apr. 2025. DOI: 10.5281/zenodo.15031354. [Online]. Available: https://doi.org/10.5281/zenodo.15031354.

[14] S. Banerjee, A. Agarwal, and S. Singla, *Llms will always hallucinate, and we need to live with this*, 2024. arXiv: 2409.05746.

[15] N. Banovic, Z. Yang, A. Ramesh, and A. Liu, "Being trustworthy is not enough: How untrustworthy artificial intelligence (AI) can deceive the end-users and gain their trust," *Proc. ACM Hum. Comput. Interact.*, vol. 7, no. CSCW1, 2023.

[16] S. Becker, W. Hasselbring, A. Paul, *et al.*, "Trustworthy software systems: A discussion of basic concepts and terminology," *ACM SIGSOFT SEN*, vol. 31, no. 6, 2006.

[17] I. Benbasat, D. Gefen, and P. A. Pavlou, "Introduction to the special issue on novel perspectives on trust in information systems," *MIS Quarterly*, vol. 34, no. 2, 2010.

[18] I. Benbasat, D. Gefen, and P. A. Pavlou, "Special issue: Trust in online environments," *J. Manag. Inf. Syst.*, vol. 24, no. 4, 2008.

[19] I. Benbasat and W. Wang, "Trust in and adoption of online recommendation agents," *J. Assoc. Inf. Syst.*, vol. 6, no. 3, 2005.

[20] I. Bertram, J. Hong, Y. Huang, W. Weimer, and Z. Sharafi, "Trustworthiness perceptions in code review: An eye-tracking study," in *ACM/IEEE ESEM 2020*, 2020.

[21] V. D. Bianco, L. Lavazza, S. Morasca, and D. Taibi, "A survey on open source software trustworthiness," *IEEE Softw.*, vol. 28, no. 5, 2011.

[22] A. W. Boggust, B. Hoover, A. Satyanarayan, and H. Strobelt, "Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior," in *ACM CHI 2022*, 2022.

[23] L. Bratthall and M. Jørgensen, "Can you trust a single data source exploratory software engineering case study?" *Empir. Softw. Eng.*, vol. 7, no. 1, 2002.

[24] A. Brown, S. D'Angelo, A. Murillo, C. Jaspan, and C. Green, "Identifying the factors that influence trust in ai code completion," in *AIware 2024*, 2024.

[25] H. G. Brown, M. S. Poole, and T. L. Rodgers, "Interpersonal traits, complementarity, and trust in virtual collaboration," *J. Manag. Inf. Syst.*, vol. 20, no. 4, 2004.

[26] B. Buhnova, D. Halasz, D. Iqbal, and H. Bangui, "Survey on trust in software engineering for autonomous dynamic ecosystems," in *ACM SAC 2023*, 2023.

[27] J. Burgoon, J. Bonito, B. Bengtsson, C. Cederberg, M. Lundeberg, and L. Allspach, "Interactivity in human–computer interaction: A study of credibility, understanding, and influence," *Computers in Human Behavior*, vol. 16, no. 6, 2000.

[28] J. Cai, R. Wu, H. Hua, G. Sun, Y. Zhang, and X. F. and, "Psychometric properties of the chinese version of human-computer trust (hct) scale in the chinese sample of adults," *Int. J. Hum.-Comput. Int.*, vol. 41, no. 1, 2025.

[29] R. Calinescu, D. Weyns, S. Gerasimou, M. U. Iftikhar, I. Habli, and T. Kelly, "Engineering trustworthy self-adaptive software with dynamic assurance cases," *IEEE Trans. Software Eng.*, vol. 44, no. 11, 2018.

[30] M. Chen, "Trust and trust-engineering in artificial intelligence research: Theory and praxis," *Philosophy & Technology*, vol. 34, no. 4, 2021.

[31] R. Cheng, R. Wang, T. Zimmermann, and D. Ford, ""it would work for me too": How online communities shape software developers' trust in ai-powered code generation tools," *ACM Trans. Interact. Intell. Syst.*, vol. 14, no. 2, 2024.

[32] K. Chopra and W. A. Wallace, "Trust in electronic environments," in *HICSS 2003*, 2003.

[33] C. L. Corritore, B. Kracher, and S. Wiedenbeck, "On-line trust: Concepts, evolving themes, a model," *Int. J. Hum. Comput. Stud.*, vol. 58, no. 6, 2003.

[34] K. E. Culley and P. Madhavan, "A note of caution regarding anthropomorphism in HCI agents," *Comput. Hum. Behav.*, vol. 29, no. 3, 2013.

[35] S. D. Das, P. K. Bala, and A. N. Mishra, "Towards defining a trustworthy artificial intelligence system development maturity model," *J. Comput. Inf. Syst.*, vol. 64, no. 6, 2024.

[36] A. R. Dennis, A. Lakhiwal, and A. Sachdeva, "Ai agents as team members: Effects on satisfaction, conflict, trustworthiness, and willingness to work with," *J. Manag. Inf. Syst.*, vol. 40, no. 2, 2023.

[37] J. Dorsch and O. Deroy, "Quasi-metacognitive machines: Why we don't need morally trustworthy ai and communicating reliability is enough," *Philosophy & Technology*, vol. 37, no. 2, 2024.

[38] J. M. Durán and G. Pozzi, "Trust and trustworthiness in ai," *Philosophy & Technology*, vol. 38, no. 1, 2025.

[39] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Hum. Comput. Stud.*, vol. 58, no. 6, 2003.

[40] M. Eiband, D. Buschek, A. Kremer, and H. Hussmann, "The impact of placebic explanations on trust in intelligent systems," in *Extended Abstracts of ACM CHI*, 2019.

[41] A. Ferrario and M. Loi, "How explainability contributes to trust in ai," in *ACM FAccT 2022*, 2022.

[42] B. J. Fogg and H. Tseng, "The elements of computer credibility," in *CHI 1999*, 1999.

[43] K. Frost-Arnold, "Imposters, tricksters, and trustworthiness as an epistemic virtue," *Hypatia*, vol. 29, no. 4, 2014.

[44] Ö. Ö. Garibay, B. Winslow, S. Andolina, *et al.*, "Six human-centered ai grand challenges," *Int. J. Hum. Comput. Interact.*, vol. 39, no. 3, 2023.

[45] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and TAM in online shopping: An integrated model," *MIS Quarterly*, vol. 27, no. 1, 2003.

[46] D. Gefen and D. Straub, "Managing user trust in b2c e-services," *e-Service*, vol. 2, no. 2, 2003.

[47] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Acad. Manag. Ann.*, vol. 14, no. 2, 2020.

[48] M. Grimsley and A. Meehan, "E-government information systems: Evaluation-led design for public value and client trust," *Eur. J. Inf. Syst.*, vol. 16, no. 2, 2007.

[49] S. Gulati, J. McDonagh, S. Sousa, and D. Lamas, "Trust models and theories in human–computer interaction: A systematic literature review," *Computers in Human Behavior Reports*, vol. 16, p. 100 495, 2024.

[50] S. Gulati, S. C. Sousa, and D. Lamas, "Design, development and evaluation of a human-computer trust scale," *Behav. Inf. Technol.*, vol. 38, no. 10, 2019.

[51] R. Hardin, "Trust and trustworthiness," *Russell Sage Foundation*, 2002.

[52] A. Hasija and T. L. Esper, "In artificial intelligence (ai) we trust: A qualitative investigation of ai technology acceptance," *J. Bus. Logist.*, vol. 43, no. 3, 2022.

[53] A. E. Hassan, D. Lin, G. K. Rajbahadur, *et al.*, "Rethinking software engineering in the era of foundation models: A curated catalogue of challenges in the development of trustworthy fmware," in *ACM FSE 2024 Companion Proceedings*, 2024.

[54] H. He, J. Gray, A. Cangelosi, Q. Meng, T. M. McGinnity, and J. Mehnen, "The challenges and opportunities of human-centered AI for trustworthy robots and autonomous systems," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 4, 2022.

[55] H. Hoehle, S. Huff, and S. Goode, "The role of continuous trust in information systems continuance," *J. Comput. Inf. Syst.*, vol. 52, no. 4, 2012.

[56] K. A. Hoff and M. N. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Hum. Factors*, vol. 57, no. 3, 2015.

[57] F. Hou and S. Jansen, "A systematic literature review on trust in the software ecosystem," *Empir. Softw. Eng.*, vol. 28, no. 1, p. 8, 2023.

[58] W. Huo, Z. Zhang, J. Qu, *et al.*, "Speciesism and preference of human-artificial intelligence interaction: A study on medical artificial intelligence," *Int. J. Hum. Comput. Interact.*, vol. 40, no. 11, 2024.

[59] W. Huo, G. Zheng, J. Yan, L. Sun, and L. Han, "Interacting with medical artificial intelligence: Integrating self-responsibility attribution, human-computer trust, and personality," *Comput. Hum. Behav.*, vol. 132, 2022.

[60] A. Jacovi, A. Marasovic, T. Miller, and Y. Goldberg, "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI," in *ACM FAccT 2021*, 2021.

[61] S. Jalali, Ç. Gencel, and D. Smite, "Trust dynamics in global software engineering," in *ACM/IEEE ESEM 2010)*, 2010.

[62] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski, "Data governance: Organizing data for trustworthy artificial intelligence," *Gov. Inf. Q.*, vol. 37, no. 3, 2020.

[63] M. Jeon, "The effects of emotions on trust in human-computer interaction: A survey and prospect," *Int. J. Hum. Comput. Interact.*, vol. 40, no. 22, 2024.

[64] B. Johnson, C. Bird, D. Ford, N. Forsgren, and T. Zimmermann, "Make your tools sparkle with trust: The PICSE framework for trust in software tools," in *SEIP@ICSE 2023*, 2023.

[65] A. N. Joinson, U. Reips, T. Buchanan, and C. B. P. Schofield, "Privacy, trust, and self-disclosure online," *Hum. Comput. Interact.*, vol. 25, no. 1, 2010.

[66] K. Jones, "Second-hand moral knowledge," *The Journal of Philosophy*, vol. 96, no. 2, 1999.

[67] L. P. R. Jr., G. Bansal, N. P. Melville, and T. Stafford, "Introduction to the special issue on AI fairness, trust, and ethics," *AIS Trans. Hum. Comput. Interact.*, vol. 12, no. 4, 2020.

[68] L. Kästner, M. Langer, V. Lazar, A. Schomäcker, T. Speith, and S. Sterz, "On the relation of trust and explainability: Why to engineer for trustworthiness," in *IEEE REW 2021*, 2021.

[69] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: A review," *ACM Comput. Surv.*, vol. 55, no. 2, 2023.

[70] I. Keivanloo and J. Rilling, "Software trustworthiness 2.0 - A semantic web enabled global source code analysis approach," *J. Syst. Softw.*, vol. 89, 2014.

[71] A. Keren, "Trust and belief: A preemptive reasons account," *Synthese*, vol. 191, no. 12, 2014.

[72] R. Khojah, M. Mohamad, P. Leitner, and F. G. de Oliveira Neto, "Beyond code generation: An observational study of chatgpt usage in software engineering practice," *Proc. ACM Softw. Eng.*, vol. 1, no. FSE, 2024.

[73] S. Y. X. Komiak and I. Benbasat, "The effects of personalization and familiarity on trust and adoption of recommendation agents," *MIS Quarterly*, vol. 30, no. 4, 2006.

[74] P. Korzyński, S. C. e Silva, A. M. Górska, and G. Mazurek, "Trust in ai and top management support in generative-ai adoption," *J. Comput. Inf. Syst.*, 2024.

[75] B. Kou, S. Chen, Z. Wang, L. Ma, and T. Zhang, "Do large language models pay similar attention like human programmers when generating code?" *Proc. ACM Softw. Eng.*, vol. 1, no. FSE, 2024.

[76] M. Kraus, N. Wagner, and W. Minker, "Effects of proactive dialogue strategies on human-computer trust," in *ACM UMAP 2020*, 2020.

[77] M. Langer, D. Oster, T. Speith, *et al.*, "What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research," *Artificial Intelligence*, vol. 296, 2021.

[78] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, 2004.

[79] B. Li, P. Qi, B. Liu, *et al.*, "Trustworthy AI: from principles to practices," *ACM Comput. Surv.*, vol. 55, no. 9, 2023.

[80] X. Li, T. J. Hess, and J. S. Valacich, "Using attitude and social influence to develop an extended trust model for information systems," *Data Base*, vol. 37, no. 2-3, 2006.

[81] X. Li, T. J. Hess, and J. S. Valacich, "Why do we trust new technology? A study of initial trust formation with organizational information systems," *J. Strateg. Inf. Syst.*, vol. 17, no. 1, 2008.

[82] Y. Li, B. Wu, Y. Huang, and S. Luan, "Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-ai trust," *Front. Psychol.*, vol. 15, 2024.

[83] N. Limam and R. Boutaba, "Assessing software service quality and trustworthiness at selection time," *IEEE Trans. Software Eng.*, vol. 36, no. 4, 2010.

[84] B. Liu, G. Li, H. Zhang, Y. Jin, Z. Wang, and D. Shao, "The gap between trustworthy AI research and trustworthy software research: A tertiary study," *ACM Comput. Surv.*, vol. 57, no. 3, 2025.

[85] Y. Liu, T. Le-Cong, R. Widyasari, *et al.*, "Refining chatgpt-generated code: Characterizing and mitigating code quality issues," *ACM Trans. Softw. Eng. Methodol.*, vol. 33, no. 5, 2024.

[86] S. Lockey, N. M. Gillespie, D. Holm, and I. A. Someh, "A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions," in *HICSS 2021*, 2021.

[87] R. Lukyanenko, W. Maass, and V. C. Storey, "Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities," *Electron. Mark.*, vol. 32, no. 4, 2022.

[88] M. Madsen and S. Gregor, "Measuring human-computer trust," in *11th Australasian Conference on Information Systems*, vol. 53, 2000.

[89] A. Manzini, G. Keeling, N. Marchal, K. R. McKee, V. Rieser, and I. Gabriel, "Should users trust advanced ai assistants? justified trust as a function of competence and alignment," in *ACM FAccT 2024*, 2024.

[90] J. Mao, J. Lee, and C. Deng, "Vendors' perspectives on trust and control in offshore information systems outsourcing," *Inf. Manag.*, vol. 45, no. 7, 2008.

[91] J. Mattioli, H. Sohier, A. Delaborde, *et al.*, "An overview of key trustworthiness attributes and kpis for trusted ml-based systems engineering," *AI and Ethics*, vol. 4, no. 1, 2024.

[92] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Acad. Manag. Rev.*, vol. 20, no. 3, 1995.

[93] D. J. McAllister, "Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations," *Academy of management journal*, vol. 38, no. 1, pp. 24–59, 1995.

[94] D. McGeorge, *The ChatGPT Revolution: How to Simplify Your Work and Life Admin with AI*. Wiley, 2023.

[95] D. H. McKnight, M. Carter, J. B. Thatcher, and P. F. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Trans. Manag. Inf. Syst.*, vol. 2, no. 2, 12:1–12:25, 2011. DOI: 10.1145/1985347.1985353. [Online]. Available: https://doi.org/10.1145/1985347.1985353.

[96] D. H. McKnight, V. Choudhury, and C. J. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Inf. Syst. Res.*, vol. 13, no. 3, 2002.

[97] D. H. McKnight, V. Choudhury, and C. J. Kacmar, "The impact of initial consumer trust on intentions to transact with a web site: A trust building model," *J. Strateg. Inf. Syst.*, vol. 11, no. 3-4, 2002.

[98] C. McLeod, "Trust," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Fall 2023, Metaphysics Research Lab, 2023.

[99] S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, and M. L. Tielman, "A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges," *ACM J. Responsib. Comput.*, vol. 1, no. 4, Nov. 2024.

[100] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man Mach. Stud.*, vol. 27, no. 5-6, 1987.

[101] A. Murillo, A. Elizondo, S. D'Angelo, *et al.*, "Understanding and designing for trust in ai-powered developer tooling," *IEEE Softw.*, vol. 41, no. 6, 2024.

[102] B. A. Myers, J. D. Hollan, and I. F. Cruz, "Strategic directions in human-computer interaction," *ACM Comput. Surv.*, vol. 28, no. 4, 1996.

[103] C. Nass, Y. Moon, B. J. Fogg, B. Reeves, and D. C. Dryer, "Can computer personalities be human personalities?" *Int. J. Hum. Comput. Stud.*, vol. 43, no. 2, 1995.

[104] P. J. Nickel, "Trust in medical artificial intelligence: A discretionary account," *Ethics Inf. Technol.*, vol. 24, no. 1, p. 7, 2022.

[105] Y. Noller, R. Shariffdeen, X. Gao, and A. Roychoudhury, "Trust enhancement issues in program repair," in *ICSE 2022*, 2022.

[106] F. Nothdurft, G. Bertrand, H. Lang, and W. Minker, "Adaptive explanation architecture for maintaining human-computer trust," in *IEEE COMPSAC 2012*, X. Bai, F. Belli, E. Bertino, *et al.*, Eds., 2012.

[107] J.-G. Park and J. Lee, "Knowledge sharing in information systems development projects: Explicating the role of dependence and trust," *Int. J. Proj. Manag.*, vol. 32, no. 1, 2014.

[108] N. Perry, M. Srivastava, D. Kumar, and D. Boneh, "Do users write more insecure code with AI assistants?" In *ACM SIGSAC CCS 2023*, 2023.

[109] F. T. Piller, M. Srour, and T. J. Marion, "Generative ai, innovation, and trust," *J. Appl. Behav. Sci.*, vol. 60, no. 4, 2024.

[110] S. Pink, E. Quilty, J. Grundy, and R. Hoda, "Trust, artificial intelligence and software practitioners: An interdisciplinary agenda," *AI & SOCIETY*, 2024.

[111] A. Pinto, S. Sousa, A. Simões, and J. Santos, "A trust scale for human-robot interaction: Translation, adaptation, and validation of a human computer trust scale," *Hum. Behav. Emerg. Tech.*, vol. 2022, no. 1, 2022.

[112] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. M. Wallach, "Manipulating and measuring model interpretability," in *ACM CHI 2021*, 2021.

[113] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships.," *Journal of Personality and Social Psychology*, vol. 49, no. 1, 1985.

[114] J. Riegelsberger, M. A. Sasse, and J. D. McCarthy, "The mechanics of trust: A framework for research and design," *Int. J. Hum. Comput. Stud.*, vol. 62, no. 3, 2005.

[115] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-

discipline view of trust," *Academy of management review*, vol. 23, no. 3, pp. 393–404, 1998.

[116] M. Ryan, "In ai we trust: Ethics, artificial intelligence, and reliability," *Sci. Eng. Ethics*, vol. 26, no. 5, 2020.

[117] S. Sabouri, P. Eibl, X. Zhou, *et al.*, "Trust dynamics in ai-assisted development: Definitions, factors, and implications," in *ICSE 2025*, 2025.

[118] A. Safdari, "Toward an empathy-based trust in human-otheroid relations," *AI & SOCIETY*, 2024.

[119] K. Saffarizadeh, M. Keil, and L. Maruping, "Relationship between trust in the ai creator and trust in ai systems: The crucial role of ai alignment and steerability," *J. Manag. Inf. Syst.*, vol. 41, no. 3, 2024.

[120] A. Salles, K. Evers, and M. F. and, "Anthropomorphism in ai," *AJOB Neuroscience*, vol. 11, no. 2, 2020.

[121] N. Scheman, "Trust and trustworthiness," in *The Routledge Handbook of Trust and Philosophy*, Routledge, 2020.

[122] P. Schmidt and F. Bießmann, *Quantifying interpretability and trust in machine learning systems*, 2019. arXiv: 1901.08558.

[123] P. Schmidt, F. Bießmann, and T. Teubner, "Transparency and trust in artificial intelligence systems," *J. Decis. Syst.*, vol. 29, no. 4, 2020.

[124] J. R. Schoenherr, R. Abbas, K. Michael, P. Rivas, and T. D. Anderson, "Designing ai using a human-centered approach: Explainability and accuracy toward trustworthiness," *IEEE Trans. Technol. Soc.*, vol. 4, no. 1, 2023.

[125] C. D. Schultz, "A trust framework model for situational contexts," in *ACM PST 2006*, 2006.

[126] E. Shein, "The impact of AI on computer science education," *Commun. ACM*, vol. 67, no. 9, 2024.

[127] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, 2020.

[128] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *Int. J. Hum. Comput. Interact.*, vol. 36, no. 6, 2020.

[129] B. Shneiderman, "Human-centered artificial intelligence: Three fresh ideas," *AIS Transactions on Human-Computer Interaction*, vol. 12, no. 3, pp. 109–124, 2020.

[130] A. Silva, M. Schrum, E. Hedlund-Botti, N. Gopalan, and M. C. Gombolay, "Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction," *Int. J. Hum. Comput. Interact.*, vol. 39, no. 7, 2023.

[131] T. W. Simpson, "What is trust?" *Pacific Philosophical Quarterly*, vol. 93, no. 4, 2012.

[132] M. Söllner, "What we really know about antecedents of trust: A critical review of the empirical information systems literature on trust," *Psychology of Trust: New Research*, 2013.

[133] M. Söllner, A. Hoffmann, and J. M. Leimeister, "Why different trust relationships matter for information systems users," *Eur. J. Inf. Syst.*, vol. 25, no. 3, 2016.

[134] S. C. Sousa, D. Lamas, J. Cravino, and P. Martins, "Human-centered trustworthy framework: A human-computer interaction perspective," *IEEE Computer*, vol. 57, no. 3, 2024.

[135] G. Sridhara, R. H. G., and S. Mazumdar, *Chatgpt: A study on its utility for ubiquitous software engineering tasks*, 2023. arXiv: 2305.16837.

[136] K. J. Stewart, "How hypertext links influence consumer perceptions to build and degrade trust online," *J. Manag. Inf. Syst.*, vol. 23, no. 1, 2006.

[137] H. Suen and K. Hung, "Building trust in automatic video interviews using various AI interfaces: Tangibility, immediacy, and transparency," *Comput. Hum. Behav.*, vol. 143, 2023.

[138] C. Tam, A. Loureiro, and T. Oliveira, "The individual performance outcome behind e-commerce," *Internet Research*, vol. 30, no. 2, 2020.

[139] C. Tansley and S. Newell, "Project social capital, leadership and trust," *Journal of Managerial Psychology*, vol. 22, no. 4, 2007.

[140] J. B. Thatcher, D. H. McKnight, E. W. Baker, R. E. Arsal, and N. H. Roberts, "The role of trust in postadoption IT exploration: An empirical examination of knowledge management systems," *IEEE Trans. Engineering Management*, vol. 58, no. 1, 2011.

[141] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electron. Mark.*, vol. 31, no. 2, 2021.

[142] K. Thompson, "Reflections on trusting trust," *Commun. ACM*, vol. 27, no. 8, 1984.

[143] C. Tomitza, M. Schaschek, L. Straub, and A. Winkelmann, "What is the minimum to trust ai? - A requirement analysis for (generative) ai-based texts," in *Wirtschaftsinformatik (WI))*, 2023, p. 35.

[144] E. H. Trainer, B. Al-Ani, and D. F. Redmiles, "Impact of collaborative traces on trustworthiness," in *CHASE 2011*, 2011, pp. 40–47.

[145] I. Troshani, S. R. Hill, C. Sherman, and D. Arthur, "Do we trust in ai? role of anthropomorphism and intelligence," *J. Comput. Inf. Syst.*, vol. 61, no. 5, 2021.

[146] J. Viehoff, "Making trust safe for ai? non-agential trust as a conceptual engineering problem," *Philosophy & Technology*, vol. 36, no. 4, 2023.

[147] M. Virvou, G. A. Tsihrintzis, and E. Tsichrintzi, "VIRTSI: A novel trust dynamics model enhancing artificial intelligence collaboration with human users - insights from a chatgpt evaluation study," *Inf. Sci.*, vol. 675, p. 120 759, 2024. DOI: 10.1016/J.INS.2024.120759. [Online]. Available: https://doi.org/10.1016/j.ins.2024.120759.

[148] N. Wang, X. Shen, and Y. Sun, "Transition of electronic word-of-mouth services from web to mobile context: A trust transfer perspective," *Decis. Support Syst.*, vol. 54, no. 3, 2013.

[149] R. Wang, R. Cheng, D. Ford, and T. Zimmermann, "Investigating and designing for trust in ai-powered code generation tools," in *ACM FAccT 2024*, 2024.

[150] Y. Wang and D. F. Redmiles, “Cheap talk, cooperation, and trust in global software engineering - an evolutionary game theory model with empirical support,” *Empir. Softw. Eng.*, vol. 21, no. 6, 2016.

[151] D. G. Widder, L. Dabbish, J. D. Herbsleb, A. Holloway, and S. Davidoff, “Trust in collaborative automation in high stakes software engineering work: A case study at NASA,” in *ACM CHI 2021*, 2021.

[152] A. F. Winfield and M. Jirotka, “Ethical governance is essential to building trust in robotics and artificial intelligence systems,” *Philos. Trans. R. Soc. A*, vol. 376, no. 2133, 2018.

[153] E. Winter, D. Bowes, S. Counsell, *et al.*, “How do developers really feel about bug fixing? directions for automatic program repair,” *IEEE Trans. Software Eng.*, vol. 49, no. 4, 2023.

[154] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Springer, 2012.

[155] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, “Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI,” *Int. J. Hum. Comput. Interact.*, vol. 39, no. 3, 2023.

[156] L. Yan, S. Greiff, Z. Teuber, and D. Gašević, “Promises and challenges of generative artificial intelligence for human learning,” *Nature Human Behaviour*, vol. 8, no. 10, 2024.

[157] R. Yang and S. Wibowo, “User trust in artificial intelligence: A comprehensive conceptual framework,” *Electron. Mark.*, vol. 32, no. 4, 2022.

[158] I. Yaniv, “Receiving other people’s advice: Influence and benefit,” *Organizational Behavior and Human Decision Processes*, vol. 93, no. 1, 2004.

[159] R. Yokoi, Y. Eguchi, T. Fujita, and K. Nakayachi, “Artificial intelligence is trusted less than a doctor in medical treatment decisions: Influence of perceived care and value similarity,” *Int. J. Hum. Comput. Interact.*, vol. 37, no. 10, 2021.

[160] G. Zanotti, M. Petrolo, D. Chiffi, and V. Schiaffonati, “Keep trusting! a plea for the notion of trustworthy ai,” *AI & SOCIETY*, 2023.

[161] M. Zhang, X. Zhao, and M. Lyles, “Effects of absorptive capacity, trust and information systems on product innovation,” *Int. J. Oper. Prod. Manag.*, vol. 38, no. 2, 2018.

[162] P. Zhang, S. Ding, and Q. Zhao, “Exploiting blockchain to make AI trustworthy: A software development lifecycle view,” *ACM Comput. Surv.*, vol. 56, no. 7, 2024.

[163] A. Ziegler, E. Kalliamvakou, X. A. Li, *et al.*, “Measuring github copilot’s impact on productivity,” *Commun. ACM*, vol. 67, no. 3, 2024.