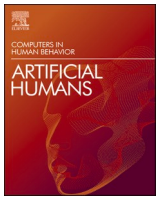
Computers in Human Behavior: Artificial Humans 2 (2024) 100043

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



Trust in artificial intelligence: Literature review and main path analysis☆

Bruno Miranda Henrique [a,*], Eugene Santos Jr. [a,**]

[a] Thayer School of Engineering, Dartmouth College, Hanover - NH, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Trust
Trust calibration
Literature review
Main path analysis

ABSTRACT

Artificial Intelligence (AI) is present in various modern systems, but it is still subject to acceptance in many fields. Medical diagnosis, autonomous driving cars, recommender systems and robotics are examples of areas in which some humans distrust AI technology, which ultimately leads to low acceptance rates. Conversely, those same applications can have humans who over rely on AI, acting as recommended by the systems with no criticism regarding the risks of a wrong decision. Therefore, there is an optimal balance with respect to trust in AI, achieved by calibration of expectations and capabilities. In this context, the literature about factors influencing trust in AI and its calibration is scattered among research fields, with no objective summaries of the overall evolution of the theme. In order to close this gap, this paper contributes a literature review of the most influential papers on the subject of trust in AI, selected by quantitative methods. It also proposes a Main Path Analysis of the literature, highlighting how the theme has evolved over the years. As results, researchers will find an overview on trust in AI based on the most important papers objectively selected and also tendencies and opportunities for future research.

1. Introduction

Artificial Intelligence (AI) can be regarded as a common part of modern society. This technology can be found in many applications, from movies and products recommender systems to self-driving cars (Albrecht & Stone, 2018). Also, it has many forms and flavors, like automation machines (Parasuraman & Riley, 1997), chatbots (Araujo, 2018), decision-support aids (Buehler & Weisswange, 2020) or even being part of teams (Wilder et al., 2020). In these cases, AI members are referred to as autonomous agents, in the sense that they might have their own beliefs and goals (Castelfranchi, 1998; Castelfranchi & Falcone, 1998) and, therefore, act on their own. However, in any case, AI agents are subject to human relationships, acceptance, ignorance and trust (Castelfranchi & Falcone, 2000). In fact, trust in AI can be determinant of its acceptance (Kelly et al., 2023; Sun & Botev, 2021). Within this context, this paper assesses how the literature has considered trust in the relationship between humans and AI.

Trust itself is not easily defined. Deutsch (1960) relates trust to personal and situational characteristics in a relationship. By his turn, Mayer et al. (1995) define trust in terms of acceptance to be vulnerable to the actions of an agent. As another example, Pinyol and Sabater-Mir (2011) argue that trust is the process of deciding to interact with an agent. Regardless of definition, because trust involves a relationship, it can be studied in the context of autonomous agents interactions. Thus, trust in AI is a valid research area and has been approached by different methods. Within this theme, initial researchers focused on trust towards automation (Lee & Moray, 1992; Muir, 1994) and implications of reliance in acceptance. However, as machines evolved and gained autonomous characteristics, trust is now studied towards AI, like in Araujo et al. (2020) and Zhang et al. (2023).

With many possible approaches to assess trust in AI, authors seek references in many scientific fields, such as Psychology, Management and Computational Sciences. With that, the literature on trust in AI encompasses various research areas, making it a challenge to compile the most influential papers about it. Therefore, some authors provide important surveys, but most analyze only special parts of the literature. For example, Lee and See (2004) and Hoff and Bashir (2014) provide a literature review on trust regarding automation; Pinyol and Sabater-Mir (2011) compile trust models regarding reputation; Adadi and Berrada (2018) conduct an extensive review on methods and models of eXplainable Artificial Intelligence (xAI), a topic closely related to trust; and Rapp et al. (2021) specialize on trust and other relationships

☆ This document was a collaborative effort.

* Corresponding author.
** Corresponding author.
E-mail addresses: bruno.miranda.henrique.th@dartmouth.edu (B.M. Henrique), eugene.santos.jr@dartmouth.edu (E. Santos).

https://doi.org/10.1016/j.chbah.2024.100043
Received 4 July 2023; Received in revised form 5 January 2024; Accepted 10 January 2024
Available online 17 January 2024
2949-8821/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

regarding chatbots. The most general and comprehensive literature review we are aware of was provided by Glikson and Woolley (2020), but they only go as far as 1999 in the past and restrict their search to empirical papers. Therefore, we propose an objective review on the most important works systematically compiled from all the literature on trust in AI, with no restrictions. By systematically reviewing the literature, we answer the research question of highlighting the most influential works on the theme, main approaches to it and also the research opportunities on trust in AI.

In order to keep this literature review as broad as possible, but still specialized on trust in AI, we do not impose any restrictions regarding research fields or year of publication in our search. Also, to avoid any biases and keep our objectivity regarding the most influential works in a concise way, we only rely on quantitative methods for literature selection. With those guidelines, in order to achieve a systematic literature review, we follow the methods used by Henrique et al. (2019). Therefore, from all compiled documents, we select for the review: the most cited papers; the papers with the greatest bibliographic coupling; and the papers with most occurring co-citation relationships. Also, we propose a main path of the literature on trust in AI, that is, a view of how the research on the theme evolved through time (Henrique et al., 2018). After analyzing that evolution, we provide comments on the most recent papers in the theme and also on gaps and research opportunities.

In summary, this literature review contributes to the research on trust in AI by providing a systematic and objective compilation of the most influential papers on the theme, conducting a Main Path Analysis (MPA) of the literature, addressing common approaches and identifying research gaps. With that, we intend this paper to be a guide through the minimum literature necessary for a new researcher to grasp what has been done in the theme. However, seasoned researchers will also benefit from the identified common approaches and potential opportunities for further research. In order to structure our contribution, the next section provides a brief theoretical background on AI and trust; Section 3 describes the methods used in the bibliographic search; Section 4 presents a bibliometric analysis of search results; the review of selected literature is conducted on Section 5; Section 6 presents discussions, approaches and research opportunities; and Section 7 concludes the paper.

2. Brief theoretical background

AI had many seminal concepts before becoming a huge research field, such as Alan Turing's ideas of intelligent machines (Turing, 1950, pp. 433–460), the artificial neuron model of McCulloch and Pitts (1943) and the Hebbian theory of learning (Hebb, 1949), but it was the Dartmouth College Summer workshop of 1956 that marked the official use of the term *artificial intelligence*. Since then, AI became a research field on its own, separated from Control Theory, Operations Research and Decision Theory. However, AI draws concepts and ideas from those fields and a number of others, such as Biology, Psychology, Communications, Game Theory, Mathematics, Logic and Probability Theory (Buchanan, 2005).

Although clearly attached to Computer Science, AI has practical applications in several high stakes domains, like medicine (Longoni et al., 2019), aviation (Guevarra et al., 2023), self-driving cars (Liu et al., 2020) and manufacturing (Li et al., 2017), just to cite a few. In such domains, it is imperative that AI systems are not only accurate and responsive, but also trustworthy. Hence, without a minimum level of human trust in AI, systems applying this technology would not be used at all, preventing humans to rely on AI in areas in which it has superior performance (Kelly et al., 2023; Sun & Botev, 2021). On the other hand, over reliance on faulty or less accurate AI can lead to disastrous outcomes (Dzindolet et al., 2003).

Closely related to AI, automation refers to partial or full replacement of human action by machines or systems (Parasuraman et al., 2000). It can occur in various levels and domains, such as smart homes (Brush et al., 2011), industrial communications (Wollschlaeger et al., 2017) and manufacturing (Lu et al., 2020). Although early automation was based on static rules and actions, modern intelligent automation is enabled by underlying AI systems (Glikson & Woolley, 2020). For instance, AI has been used to solve problems in home automation (Evans, 1991), agriculture (Jha et al., 2019) and healthcare (Davenport & Kalakota, 2019). Therefore, as done by Glikson and Woolley (2020), our review on trust in AI is broad enough to include previous works considering human trust toward automation perceived as intelligent systems.

3. Methods of the bibliometric analysis

The methods used in this literature review are largely based on techniques presented in Henrique et al. (2018, 2019). In order to compile a list of articles to review in a structured and quantitative way, Henrique et al. (2019) list papers with the greatest bibliographic coupling, most co-citation occurrences and papers that are part of the MPA as described in Henrique et al. (2018). They also list the most cited papers in their field, as well as the most recent ones. Following those methods as guidelines, we start by gathering all the publications listed in the Scopus database, searching for *artificial intelligence trust* in Abstracts, Key Words and Titles. However, examining the initial results, some heuristics became clear: this limited search excluded documents related to trust and *machine learning* algorithms, which are practical building blocks of AI. Also, some authors don't refer to single humans or artificial agents in their research, but to the broader term *team*. Finally, as it will be discussed in the paragraphs dedicated to the reviews, some papers refer to trust as a construct to be *calibrated* between humans and AI.

Based on the initial heuristics of a narrow search, we changed our search terms in order to keep our results as broad as possible. Thus, we conducted our final search on the Scopus database with the keywords *artificial intelligence*, *machine*,[1] *trust*, *calibration*, *team* and possible combinations between them. We also included papers listed by Google Scholar, using the same keywords. The resulting list of papers can be evaluated as a valid representation of the literature by using Lotka's Law (Saam & Reiter, 1999), according to which the frequency of publications by authors follows a special distribution, in any scientific field. Therefore, we compare the publication frequency of our compiled list with the theoretical distribution expected by Lotka's Law as a method to validate the sample as a significant portion of the entire literature.

Aside from reviewing the most cited and the most recent publications, we also include in this literature review the publications with the greatest bibliographic coupling in the database. According to Kessler (1963), when papers share the same references, they are said to be coupled bibliographically. In this way, papers based on a common literature have a bigger chance to represent a given research field. Also, bibliographic coupling helps with the representativeness of those papers that do not have enough publication time to have high amounts of citations, but are based on frequently cited literature. However, bibliographic coupling results in static citation networks, giving no insights on how the research field evolves over time (Small, 1973). To that end, co-citation networks are commonly used. According to Small (1973), papers that are frequently cited together are in a co-citation relation and the resulting network should show dynamic relationships in the literature. Therefore, we also list the papers with the most co-citation occurrences.

Finally, to gain insights on how the research field of trust in AI has evolved, we propose an MPA of the literature. According to the groundbreaking work of Hummon and Doreian (1989), the most important research papers of a given literature database can be highlighted in a timeline path. That represents how the flow of knowledge of

[1] The term *learning*, as in Machine Learning, was excluded in order to eliminate results from the research field of Education. However, the terms *machine* and *trust* resulted in the retrieval of the desired documents about Machine Learning and trust.

a given field evolved through time (Liu et al., 2013). With that, main paths are powerful tools for new researchers to gain understanding of their research fields as well as for experienced researchers to account for previous methods and conclusions. According to the tutorial presented by Henrique et al. (2018), research papers can be organized in a network, with connections made by their references. In that way, multiple paths can be calculated between a starting node, called source, and an ending node, called sink. Then, the most important path regarding citations can be found by the approach proposed by Batagelj (2003). We refer the reader to Henrique et al. (2018, 2019) for more details on how to build the main path of the literature.

4. Results of the bibliometric analysis

A search on the Scopus database using the terms defined in Section 3, on 1/18/2023, returned 293 papers. We searched in Abstracts, Key Words and Titles. However, no further restrictions were applied on the original search results, i. e. we did not filter results by publication year, knowledge areas or any other constraints. The same search on Google Scholar returned 466 documents, also searching in Abstracts, Key Words and Titles. However, after filtering duplicates of papers already retrieved by Scopus and papers for which we could not find complete information (such as references), we ended up with 177 documents retrieved exclusively from Google Scholar. Therefore, summing up the searches from both databases, our whole compilation had 470 papers related to trust in AI. Those papers have some information summarized on Table 1. As shown on the table, the years of publications go back to 1960 and the documents are scattered among 272 periodicals (journals, reports and conferences). Also, the distribution of publications per year is shown in Fig. 1, denoting that before 2017, less than 10 papers were published each year. However, Fig. 1 indicates a clear growing trend in the number of publications, illustrating the increasing importance of this research field.[2]

As stated in Section 3, we resort to Lotka's Law to validate our compiled documents database as a significant representation of the literature on trust in AI. The law theorizes that the frequency of n documents by author published each year should be proportional to $1/n^d$, where $d \approx 2$ (Saam & Reiter, 1999). This theoretical distribution is shown in the continuous line in Fig. 2 and the observed distribution is marked as circles, approximated by a coefficient of 2.33 with $R^2 \approx 0.81$ and a proportionality constant of approximately 0.23. Also, the Kolmogorov-Smirnoff test for difference between both distributions, theoretical and observed, had a p-value of 0.63, failing to reject the hypothesis that they are the same distribution. In summary, there is evidence that the compiled database approximates a meaningful portion of the whole literature on trust in AI.

Table 2 shows the venues with the highest publications count on trust in AI. The top of that table is occupied by a series of books that contains important sub-series in Computer Sciences, such as Lecture Notes in Artificial Intelligence, Transactions on Computational Collective Intelligence and Transactions on Computational Science. Because it is a collection of journals and conferences, it is no surprise it has most of the published works in the database of this review. However, it is interesting to note the single venue with most publications, which is

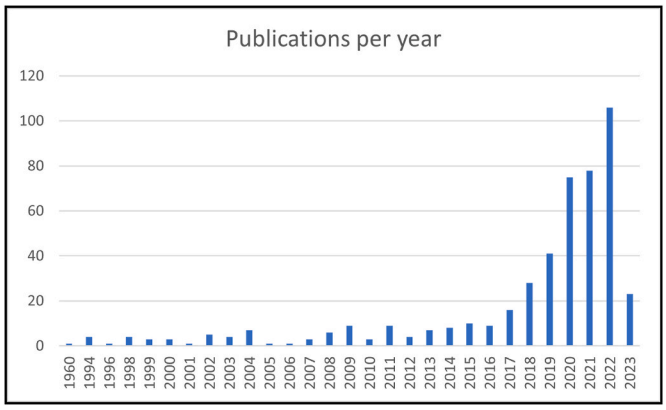


Fig. 1. Frequency of publication of the articles compiled in this literature review. The search on Scopus and Google Scholar databases was conducted on 1/18/2023, which explains the lower number of documents from 2023.

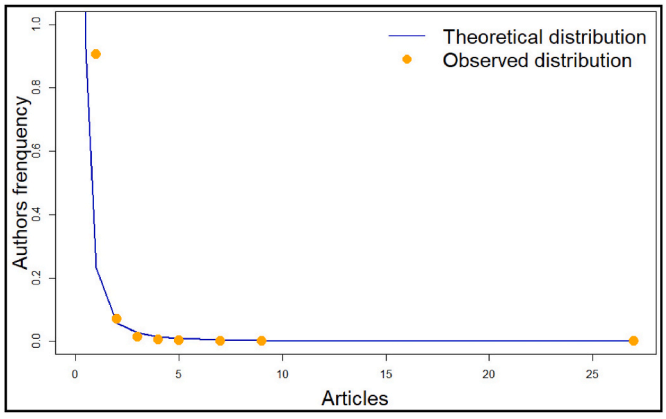


Fig. 2. Frequency of publication, by author, in the database compiled. The continuous line represents Lotka's theoretical distribution. The circles mark the distribution observed in the database of the present work.

Table 1
Description of the database of articles used in the bibliometrics.

Characteristic	Value
Number of articles.	470
Periodicals.	272
Period of the publications.	1960–2023
Average number of citations per article.	43.36
Authors.	1425
Authors with single-author articles.	72
Single-author articles.	102
Authors per article.	3.49

Table 2
The 10 journals/conferences with the highest number of articles in the database searched. *Including Subseries Lecture Notes In Artificial Intelligence and Computational Collective Intelligence.

Periodical	Number of Articles
Lecture Notes In Computer Science[a]	46
Computers In Human Behavior	15
Advances In Intelligent Systems And Computing	10
IJCAI International Joint Conference On Artificial Intelligence	9
AAAI Association for the Advancement of Artificial Intelligence Spring Symposium	8
International Journal of Human-Computer Interaction	8
Human Factors	6
Journal Of Medical Internet Research	6
International Joint Conference On Autonomous Agents And Multiagent Systems	6
Frontiers in Psychology	5

[a] Including Subseries Lecture Notes In Artificial Intelligence and Computational Collective Intelligence.

[2] The search on the databases was conducted in January 2023, which explains the low number of documents from that year.

Computers in Human Behavior. This journal specializes in a psychological perspective of computers usage,[3] indicating that research regarding trust in AI finds a common ground of reference in Psychology points of view. Also, as shown in Table 2 for the compiled database of this work, Computers in Human Behavior has even more publications on trust in AI than IJCAI and AAAI, highly reputable conferences on Artificial Intelligence.

5. Review of the selected literature

In this section, we analyze the documents for our structured literature review on trust in AI. We strictly followed the quantitative methods described in Section 3 to avoid any qualitative bias when selecting the papers. Therefore, we start by briefly commenting on the most cited papers in Section 5.1, followed by the papers with greatest bibliographic coupling, in Section 5.2, and co-citation relationships, in Section 5.3. Then, we present our proposal for a main path of the literature in Section 5.4, that is, our vision of how the knowledge of trust in AI had flowed through time. Finally, we end this section with the most recent papers, in Section 5.5, which represents the current state of research in the field.

5.1. Most-cited articles

Before diving into the most cited papers in this literature review, it is worth clarifying our citations count. We restrict the citations within the papers in our compiled database. This means that we only consider citations between papers inside our complete set of 470 papers. The reason we do not use Scopus and Google Scholar citation counts is that those numbers include citations from any source, regardless of research field. Therefore, the Scopus and Google Scholar numbers potentially include citations from research fields unrelated to the specific theme of trust in AI, which could insert bias into our papers selection.

As seen in Table 3, the most cited paper identified in our collection is Lee and See (2004). It is an early approach to model the trust relationship between humans and machines, a theoretical approach to trust dynamics between humans and automation. Not explicitly working with a broader concept of AI, the authors base their model on previous literature of Organizations, Psychology and Cognitive Processes, extending concepts to automation. In that context, Lee and See (2004) provide some guidance on how to achieve adequate levels of trust, which resonates with the idea of trust calibration at a very high level.

The most cited paper regarding the definition of trust is Mayer et al. (1995). Although they do not describe specific relationships between humans and machines, their definition is general enough to be applied when referencing trust in AI. Therefore, many authors (Chong et al., 2022; Kim & Song, 2021; Schmidt et al., 2020) accept Mayer et al. (1995)'s definition of trust. They define trust in terms of the acceptance of some vulnerability regarding the expected action of another agent. This idea of trust is directly extended by papers specifically assessing trust in AI. Although some authors provide their own variations (de Visser et al., 2016; Lee & See, 2004), the general idea of Mayer et al. (1995) holds true in most of them.

Hoff and Bashir (2014) research human trust in automation. In order to make a solid approach, they conducted a survey considering only previous works that reported experiments with humans and some measurement of trust in automation. Within that research, Hoff and Bashir (2014) do not consider AI directly, but the resulting factors that influence trust can be related to artificial agents as well. For example, they consider the effects of past experiences of a hypothetical human with automation, which can be related to how trust evolves with the relation of the human with an autonomous agent. Finally, Hoff and Bashir (2014) suggest that automation should include anthropomorphic characteristics and that improvements of performance and reliability can lead to trust increase.

Table 3
The 10 articles most cited in the compiled database. The number of citations refers to the citations in the compiled database according to the search conducted in this survey.

References	Title	Journal	Citations
Lee and See (2004)	Trust In Automation: Designing For Appropriate Reliance	Human Factors	42
Mayer et al. (1995)	An Integrative Model Of Organizational Trust	Academy Of Management Review	24
Hoff and Bashir (2014)	Trust In Automation: Integrating Empirical Evidence On Factors That Influence Trust	Human Factors	10
Adadi and Berrada (2018)	Peeking Inside The Black-Box: A Survey On Explainable Artificial Intelligence (XAI) (2018)	IEEE Access	9
Nass and Moon (2000)	Machines and Mindlessness: Social Responses to Computers	Journal of Social Issues	8
Araujo (2018)	Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions	Computers in Human Behavior	7
Glikson and Woolley (2020)	Human Trust in Artificial Intelligence: Review of Empirical Research	Academy of Management	7
Lee and Moray (1992)	Trust, control strategies and allocation of function in human-machine systems	Ergonomics	7
Pak et al. (2012)	Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults	Ergonomics	7
Castelfranchi (1998)	Modeling social action for AI agents	Artificial Intelligence	5

In general, authors analyze various constructs influencing the relationships between humans and machines in cooperation contexts, with more or less levels of interaction between agents. Among those constructs, transparency is often mentioned. In this regard, xAI is a prolific research field, with its own body of knowledge. However, the field still lacks a firm definition of its subject, according to Adadi and Berrada (2018). Then, to propose definitions for main concepts and essential understandings related to xAI, Adadi and Berrada (2018) survey 381 papers, establishing key terms and proposing a classification for the literature on the theme.

Nass and Moon (2000) is also a majorly cited paper regarding how humans interact with machines. They report experiments with human subjects and computers in situations of cooperation, similar to other experiment setups of human-machine teaming. In those situations, according to Nass and Moon (2000), people tend to apply social rules towards the machine. As examples, people might be polite to computers; or they might attribute a gender for a machine. A particularly interesting finding in the experiments reported in Nass and Moon (2000) was that when the computer was framed as being from the same ethnicity of human subjects, it was perceived as more trustworthy. This technique of exploring human cues when designing artificial agents was applied by Araujo (2018). In that paper, also present in Table 3, the author experiments with human perceptions about a chatbot with human-like characteristics. The results indicate positive effects in building relationships with humans when the chatbot is designed to incorporate some anthropomorphism.

As mentioned in Section 1, Glikson and Woolley (2020) provide a broad literature review on trust in AI, organizing the research areas in

[3] https://www.sciencedirect.com/journal/computers-in-human-behavior.

robotics, virtual agents and embedded agents. Within each area, they define dimensions that can affect cognitive trust (based on rationality and knowledge about the artificial agent) and emotional trust (related to anthropomorphism, mood and immediacy behaviors). The dimensions can be read as characteristics pertaining to each AI form or implementation and also environment and task domains. They are listed as tangibility, transparency, reliability, immediacy behaviors and the task itself. Glikson and Woolley (2020) discuss that, when developing an AI agent, not only its level of intelligence is important, but also how it is represented should be taken into consideration.

The earliest paper in Table 3 is Lee and Moray (1992). Similarly to Hoff and Bashir (2014), Lee and Moray (1992) report empirical experiments involving humans and their interaction with automation tools. The authors examine how trust can influence human decisions when they are supervising an industry controller device, in a simulated environment. In that context, trust was measured by questionnaires. Their results suggest that performance and failures from the automation influence trust in a dynamic manner. Another important contribution of Lee and Moray (1992) is the proposition of a mathematical model of trust. They propose a moving average vector of performance and faults, each with their respective coefficients. Therefore, trust is modeled as a dynamic time series model, fitted to data collected in a specific domain.

Pak et al. (2012) report an empirical study to measure if anthropomorphic characteristics introduced in a decision aid system would increase trust. Specifically, anthropomorphic characteristics are introduced in a system supporting decisions regarding health diagnostics, with the effects on trust measured with respect to anthropomorphic cues. They find that even without changing the reliability of the system, just by introducing a picture of a person on the display, trust perception is altered, especially for young subjects.

As observed from this Section, the majority of the analyzed papers in this review treat trust in the context of human-machine collaboration. Some of the papers analyze the setup of human-machine (AI) teams (Schelble, Lopez, et al., 2022; Zhang et al., 2023), while others work with AI as aids in decision-making scenarios (Gomez et al., 2023; Pak et al., 2012; Rieger et al., 2022). However, Castelfranchi (1998), one of the most cited authors regarding autonomous agents, establishes AI entities as collaborators, essentially different from mere tools of automation. This is a major shift on how AI is framed, especially in modern applications such as robotics, multi-agent systems and human-machine teaming.

5.2. Articles with the greatest bibliographic coupling

Table 4 shows the literature with the greatest bibliographic coupling in the surveyed documents of this literature review. It starts with Castelfranchi and Falcone (2010), a book with the focus on human interaction. In that context, Castelfranchi and Falcone (2010) build their trust model dynamics borrowing concepts from Psychology and Sociology. This is a common approach of human-computer trust modeling, that is, building from previously accepted constructs applicable to human relationships and then extending to relations involving artificial autonomous agents (Lee & See, 2004; Muir, 1994). That is also the approach of de Visser et al. (2016), who research human factors influencing trust. However, differently from previous empirical papers on trust, de Visser et al. (2016) report results regarding trust resilience. In that aspect, the authors experiment with how much trust relationships between humans and cognitive agents can resist in deteriorated conditions. These trust deterioration scenarios are simulated by artificial reliability variations. In those conditions, human subjects show positive effects on trust resilience when the agent displays some anthropomorphic cues.

As done by de Visser et al. (2016), Oksanen et al. (2020) also experiment with trust between humans and artificial agents. However, their focus is on first impressions and differences about trust when the agent is perceived as a robot or as an unspecific artificial intelligence entity. In their experiments, Oksanen et al. (2020) used a trust game

Table 4
The 10 articles with the greatest bibliographic coupling among all the articles searched.

Reference	Title	Journal
Castelfranchi and Falcone (2010)	Trust theory: A socio-cognitive and computational model	(Book) John Wiley & Sons
de Visser et al. (2016)	Almost human: Anthropomorphism increases trust resilience in cognitive agents	Journal of Experimental Psychology: Applied
de Visser et al. (2019)	Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams	International Journal of Social Robotics
Oksanen et al. (2020)	Trust Toward Robots and Artificial Intelligence: An Experimental Approach to Human–Technology Interactions Online	Frontiers in Psychology
Saβmannshausen et al. (2021)	Trust in artificial intelligence within production management – an exploration of antecedents	Ergonomics
Kim and Song (2021)	How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair	Telematics and Informatics
Sun and Botev (2021)	Intelligent autonomous agents and trust in virtual reality	Computers in Human Behavior Reports
Dazeley et al. (2021)	Levels of explainable artificial intelligence for human-aligned conversational explanations	Artificial Intelligence
Zerilli et al. (2022)	How transparency modulates trust in artificial intelligence	Patterns
Lukyanenko et al. (2022)	Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities	Electronic Markets

including two artificial agents, one using a human name, which is an anthropomorphic cue, and other with an artificial nickname. However, not all human participants were informed which opponents were robots or any other form of artificial intelligence, that is, trust was measured regarding their own perceptions. According to Oksanen et al. (2020), the agent with an artificial nickname was more trusted than one with a human name. A possible explanation, proposed by the authors, is that the agent with a human name could be perceived as an actual human participant, implying some negativism towards them. This argument seems contrary to other studies that indicate improvements on trust in the presence of anthropomorphism (de Visser et al., 2016; Kim & Song, 2021; Pak et al., 2012; Waytz et al., 2014). Aside from that result, Oksanen et al. (2020) also bring evidence about the relevance of human personality on trust relationships with AI, which motivates further investigation on methods for dynamic trust calibration.

Advancing the theory on human-machine teaming and trust calibration, de Visser et al. (2019) propose a model concerning social capabilities of robots in the context of collaboration with humans. They seek understanding from human teams literature, especially regarding patterns of healthy and unhealthy relationships. As a result, their built model is centered on relationship equity, which is influenced by perceptions of behavior of team members and regulations with respect to them. However, the model for trust calibration in de Visser et al. (2019) is not implementation-ready, serving only as a framework for discussions and further insights for future research. Similarly, Saβmannshausen et al. (2021) also propose a trust model, but their focus is on the antecedent variables of trust in AI in a production management environment. They interviewed experts in a qualitative research framework, testing hypothesis involving variables related to trust, such as AI ability, comprehensibility, predictability and errors. Unlike other studies, Saβmannshausen et al. (2021) highlight the influence of digital affinity as an antecedent of trust. In fact, few studies explore digital affinity or technology literacy as factors influencing trust in AI, indicating an opportunity for further investigations. Another example is

found in Oksanen et al. (2020).

Sun and Botev (2021) bring yet another recent model proposal for trust, especially regarding acceptance of agents in virtual reality environments. Following other papers, they overview concepts scattered in the literature and put them together in a theoretical high-level framework of trust. This framework contains factors such as capability, reliability, usability, divided into dimensions of analysis. Nevertheless, the reader might be cautious regarding the proposed division, since it is just an abstraction for understanding trust as proposed by the authors, not meant for an implementation framework. Therefore, the factors described by Sun and Botev (2021) as capable of influencing trust should not be seen as an exhaustive list. Another framework built from a survey on the literature is given by Lukyanenko et al. (2022). The authors seek a deeper understanding of trust in general, then extend the constructs to trust in AI. Alongside identified research opportunities, Lukyanenko et al. (2022) bring a unique view about trust. According to their paper, trust can be seen as a mechanism used by humans to reduce complexity in relationships, abstracting the inner details about systems with which they interact. In that view, trust enables engagement between humans and between humans and AI.

Advancing the research on trust resilience, following similar ideas from de Visser et al. (2016), Kim and Song (2021) conduct experiments regarding trust restoration in decision-making support agents. Specifically, they experiment with human subjects and AI advisors in games of investment. In that scenario, the advisor would commit mistakes and try to regain trust by apologizing with human-like behaviors, acknowledging its limitations, or by framing itself as an artificial agent and referring to external factors for blame. The results showed bigger improvements in trust repair when the advisor apologized using human-like behavior, assuming the blame. However, results like these should be handled with care, especially regarding possible generalizations. As soundly stressed by Zerilli et al. (2022), most empirical research on trust in AI is conducted in domain-specific applications and with only one type of AI agent, which makes results valid only for those given special conditions. Although they are hardly generalizable, few authors bring up that warning.

Yet another review paper listed on Table 4 is Dazeley et al. (2021). It is a recent take on the literature of xAI, but with an interesting approach of assessing explanations in a five-level conversational framework that can lead to trust in AI. In this context, Dazeley et al. (2021)'s main proposal is that the agent must assess the context of the explainee in order to calibrate the adequate level of explanations it provides. Their model is based on a conversational approach, meaning that as the agent provides explanations, the explainee asks for more information or clarifications interactively. Within this interaction, the agent perceives the explainee's understanding and calibrates the level of explanations accordingly. In summary, Dazeley et al. (2021) don't show practical solutions or implementations in order to calibrate or achieve trust, but they propose a comprehensive framework to advance xAI into considering the dynamics of user interaction. Under this prism, xAI techniques implemented following Dazeley et al. (2021)'s framework can contribute to dynamic assessment of trust between humans and machines.

5.3. Articles with the greatest co-citation relationships

Table 5 lists the articles with the greatest co-citation among the compiled database of this literature review. The first reference, Parasuraman and Riley (1997), does not address specifically the trust aspects of teaming humans and autonomous agents together, but the paper establishes solid definitions of misuse and abuse of automation by humans. They discuss automation as tasks previously done by humans, who assume roles of operators. According to the authors, the interaction between humans and automation is influenced by factors such as trust. In that context, they relate misuse with over-trust on the automation and abuse is related to the usage of automation regardless of performance and possible consequences. Finally, Parasuraman and Riley (1997) provide discussions on the rational use of automation, which can be loosely related to trust calibration.

Table 5
The 10 articles with the greatest co-citation relationship among those searched.

Reference	Title	Journal
Parasuraman and Riley (1997)	Humans and Automation: Use, Misuse, Disuse, Abuse	Human Factors: The Journal of the Human Factors and Ergonomics Society
Dzindolet et al. (2003)	The role of trust in automation reliance	International Journal of Human-Computer Studies
Mcknight et al. (2011)	Trust in a specific technology	ACM Transactions on Management Information Systems
Hancock et al. (2011)	A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction	Human Factors
Waytz et al. (2014)	The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle	Journal of Experimental Social Psychology
Dietvorst et al. (2015)	Algorithm aversion: People erroneously avoid algorithms after seeing them err	Journal of Experimental Psychology: General
Salem et al. (2015)	Would You Trust a (Faulty) Robot?	Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction
Hengstler et al. (2016)	Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices	Technological Forecasting and Social Change
Longoni et al. (2019)	Resistance to Medical Artificial Intelligence	Journal of Consumer Research
Jobin et al. (2019)	The global landscape of AI ethics guidelines	Nature Machine Intelligence

Following Parasuraman and Riley (1997) in the necessity of calibrating the correct levels of reliance in automation, Dzindolet et al. (2003) conduct experiments involving image detection by humans with the aid of an autonomous system to examine the dynamics of trust in the occurrence of errors. They measured trust by using questionnaires. In that regard, trust is shown to decrease with errors. However, providing explanations for those errors might improve trust relationship with automation. Then, examining results, the authors recommend users' training on a given automation system with respect to its capabilities and limitations, which can be interpreted as initial calibration of trust in the expected results of the system.

It should be noted that Parasuraman and Riley (1997) and Dzindolet et al. (2003) do not consider trust in AI directly, but trust towards automation agents or technological tools for automation. Since automation can be a practical application of AI, as described in Section 2, the results on those papers are relevant for this literature review. The same applies to Hancock et al. (2011), who searched the literature for the factors influencing interactions between humans and robots. They applied quantitative analysis to show that trust is but one of the important elements influencing interactions between humans and robots, but there is a growing concern about it. They also point out that trust is related to reliability, type and behavior of the robots.

Mcknight et al. (2011) also don't present results specific to trust in AI, but their research outputs are relevant to the field. They conduct surveys to examine how properties of a specific technology influence trust building, regardless of human factors related to it. They surveyed students interacting with technology, in that case, a spreadsheet software, and concluded that trust can be divided into initial and knowledge-based trust. Although their questionnaires are run only on a very specific population interacting with one piece of technology and without any risk associated, results seem to agree that trust is dynamic and changes with respect to how humans experience the technology.

Adding to the research on trust in AI involving anthropomorphism,

Waytz et al. (2014) report increased levels of trust when autonomous cars exhibit human-like characteristics. Although their experiments are very context-specific, one can relate the results to other types of autonomous systems. In their paper, Waytz et al. (2014) study humans driving cars, autonomous or not, and their self-reported trust on the vehicle. However, some subjects get to experience cars exhibiting some anthropomorphism, such as human voice, name and gender. In that empirical setup, people reported higher levels of trust when they perceived the car with human-like capabilities. With those results, Waytz et al. (2014) suggest a link between trust and the human perception of machines having minds of their own.

The phenomenon of algorithm aversion is explored in Dietvorst et al. (2015). The authors are motivated by the lack of studies seeking to understand why some people have aversion to algorithms, even in situations in which they perform better than humans. The context of their experiments was the acceptance or not of forecasts given by a human medical provider, an artificial tool or a human provider assisted by an artificial agent. Interestingly, Dietvorst et al. (2015) show that people have low confidence in algorithms, even knowing when they outperform humans. The levels of confidence may be even lower when humans observe errors made by the artificial support system. By their turn, Salem et al. (2015) also investigate errors made by artificial agents. Specifically, they report empirical results about how trust in robots is affected in the presence of errors. As expected, higher rates of reliability and trustworthiness were given when the robot operated correctly.

Hengstler et al. (2016) draw insights into trust in AI by examining case studies in the fields of transportation and medical technologies. The paper distinguishes itself from the others in Table 5 because it shifts the focus of analysis to the firm promoting or developing a given AI technology. With that, trust in AI is analyzed with respect to trust in the technology itself and in the firm. Hengstler et al. (2016) argue that AI technologies are initially characterized by high levels of perceived risk and that they suffer initial resistance. Then, firms developing the technology have an important role in fostering acceptance, especially regarding their communications. Still according to Hengstler et al. (2016), some factors that promote trust are operational security and data security. Resistance to AI is also the object of Longoni et al. (2019), specifically in the medical area. Similarly to Dietvorst et al. (2015),

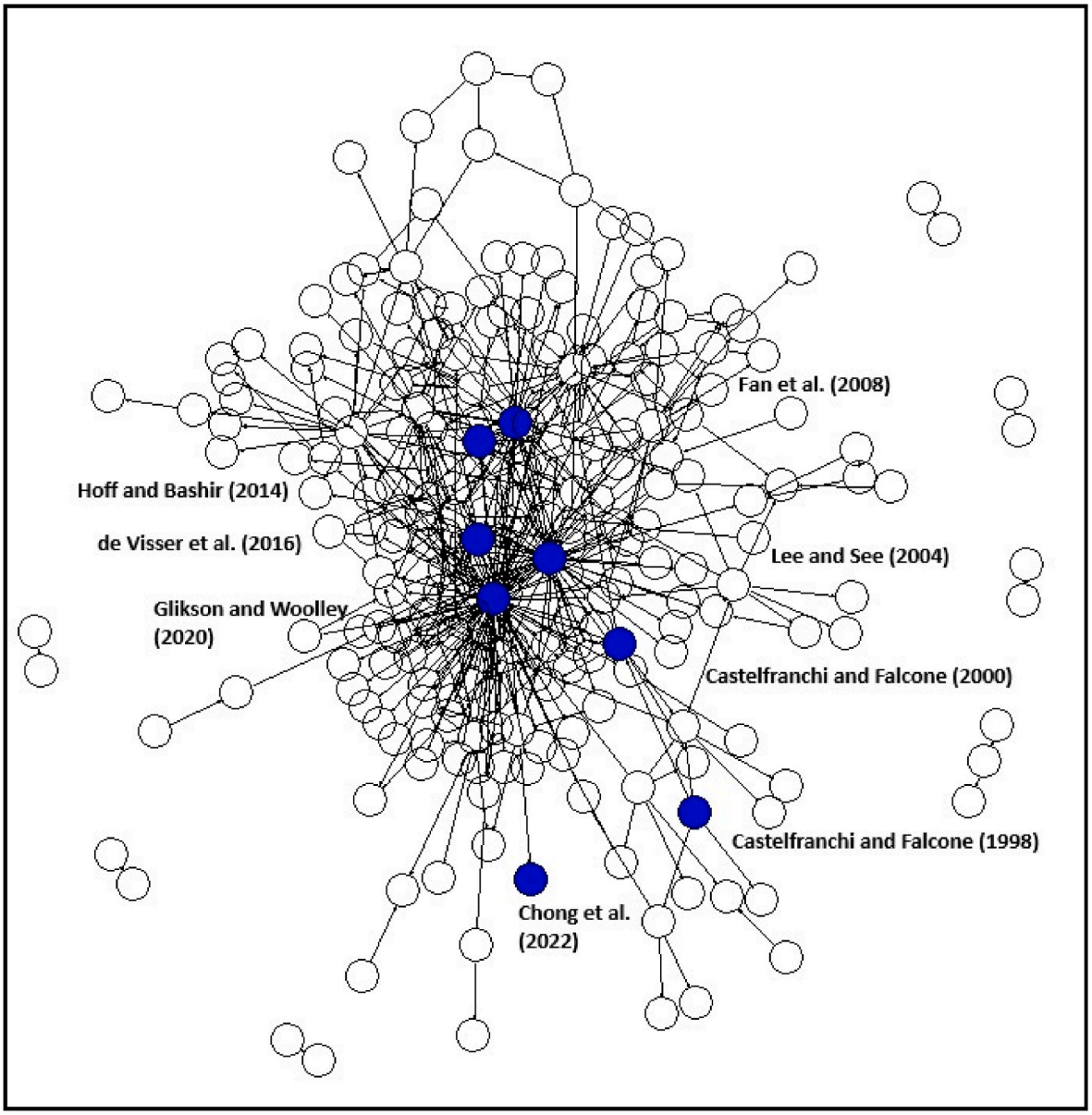


Fig. 3. Citation network of all the documents compiled in this review. Each node represents a paper and the edges are citations between them. The highlighted nodes are the Main Path of the Literature, according to the methods described in Henrique et al. (2018).

Longoni et al. (2019) applied questionnaires to measure human preferences between human medical providers and artificial agents. Again, results indicate preferences towards human providers, even when the alternative is a more accurate AI. A possible explanation presented by Longoni et al. (2019) is that people might consider that machines neglect the perceived uniqueness of them (people tend to see themselves as unique with respect to their characteristics, circumstances and symptoms. On the other hand, machines are objective, not taking into consideration that vision of uniqueness).

The last paper in Table 5 is Jobin et al. (2019). It discusses ethical AI and the existing proposals of guidelines towards that. The authors examine 84 documents from various international bodies and committees in order to analyze points of agreement between them and where they diverge. The analyzed documents constitute what they call soft law, since they only have persuasive character, serving as guides. One of the main findings of Jobin et al. (2019) is that no ethical principle is common to all documents but there is some convergence around transparency, justice and fairness, non-maleficence, responsibility and privacy. Other recurrent principles observed were accountability, beneficence, freedom and autonomy, sustainability, dignity, solidarity and trust.

5.4. Main path

Fig. 3 illustrates the complete citation network built with the papers in the compiled database. Each node represents a paper and the edges between nodes are their references to each other. In that figure, the proposed main path of the literature is highlighted and the papers that are part of it are shown in Table 6. The path was uncovered using the papers database compiled in this work and the methods described in Section 3. According to previous discussion, those papers should give insights on the flow of scientific knowledge regarding trust in AI. However, it should be noted that the path proposed is highly dependent on the papers compilation and on definitions of source and sink nodes, as discussed in Henrique et al. (2018). Therefore, multiple paths are possible, and the following paragraphs discuss only the most important one, selected using the MPA described by Batagelj (2003).

Table 6
Articles that are part of the main path of the literature searched.

Reference	Title	Journal
Castelfranchi and Falcone (1998)	Towards a theory of delegation for agent-based systems	Robotics and Autonomous Systems
Castelfranchi and Falcone (2000)	Trust and control: A dialectic link	Applied Artificial Intelligence
Lee and See (2004)	Trust in Automation: Designing for Appropriate Reliance	Human Factors: The Journal of the Human Factors and Ergonomics Society
Fan et al. (2008)	The influence of agent reliability on trust in human-agent collaboration	Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction
Hoff and Bashir (2014)	Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust	Human Factors: The Journal of the Human Factors and Ergonomics Society
de Visser et al. (2016)	Almost human: Anthropomorphism increases trust resilience in cognitive agents.	Journal of Experimental Psychology: Applied
Glikson and Woolley (2020)	Human Trust in Artificial Intelligence: Review of Empirical Research	Academy of Management
Chong et al. (2022)	Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice	Computers in Human Behavior

The proposed main path of the literature starts with Castelfranchi and Falcone (1998). The authors argue that the notions of task and delegation are not clear and previous literature has some disagreements. However, according to the authors, previous works converge on using tasks on behalf of something or someone when relating to agents. In order to propose a theory of task delegation and adoption by agents, the authors propose ontologies for the concepts of agent, plan, task, delegation and adoption. The main point of Castelfranchi and Falcone (1998) with respect to the literature of trust in AI is that the agent, especially the artificial intelligence agent, is treated as a collaborator, potentially with goals of its own; it is an autonomous agent that has a task to be adopted, with potential conflicts, and delegated by another agent by contracts in a collaborative environment. Then, shortly after publishing that seminal paper, the same authors discuss control as a contributing factor in building trust in autonomous agents (Castelfranchi & Falcone, 2000). According to the authors, previous works considered control as a restriction to trust. However, Castelfranchi and Falcone (2000) argue that control and trust are not mutually exclusive and can coexist in the relationship with the autonomous agent.

Following Table 6, Lee and See (2004) innovate surveying the literature on trust in human relationships and extending the constructs to the case of trust in automation. As commented in 5.1, Lee and See (2004) propose a theoretical model of the dynamics of trust in the context of automation. By their turn, Fan et al. (2008) conduct experiments to show how the reliability of autonomous agents can influence human trust. Fan et al. (2008) definitively depart from simple automation agents and frame the experiments with teams in which the artificial agents are members aiding in decision-making. In that context, the authors introduce systematic errors on agents' decisions, showing that this can impact human members' trust in the artificial member of the team. This paper is also important for explicitly acknowledging the importance of calibrating the expected performance of the autonomous agent and the impacts of those expectations on trust.

Similarly to Lee and See (2004), Hoff and Bashir (2014) also search previous literature for concepts that can influence trust in automation. As described in Section 5.1, Hoff and Bashir (2014) suggest the inclusion of anthropomorphism in autonomous agents' behaviors in order to achieve higher levels of trust. This tendency of discussing how human-like behaviors can improve trust is observed in various works, including de Visser et al. (2016), the next paper in the main path of the literature. Their work on the influence of anthropomorphic cues on trust resilience was briefly reviewed in Section 5.2. Following Table 6, the literature review of Glikson and Woolley (2020) summarizes the previously researched factors that influence trust in AI. No definitive model of trust or trust calibration is provided by Glikson and Woolley (2020), but their comprehensive survey provides a guide for future research and implementations.

Finally, the main path of the researched literature converges on the work of Chong et al. (2022). This paper addresses not only confidence of humans in AI, but also in themselves. The authors conducted experiments with chess players about their acceptance or denial of suggestions of moves given by an AI system. Specifically, they investigate how AI accuracy and feedback about moves affect confidence levels; and decision patterns of players that made correct decisions as to accept or ignore the suggestions made by the AI. In that context, we argue that one of the main merits of Chong et al. (2022) is the tailoring of a mathematical equation relating confidence, experience and acceptance or denial of suggestions. Although not generalizable, their equation mapped observed behaviors on the empirical experiments with chess players and led to quantitative results. As an example of result, poor performance of AI showed to be detrimental to confidence, in both the AI and in the humans themselves. However, perhaps the most striking contribution of Chong et al. (2022) is the indication that human confidence in themselves might be more important than the confidence in the AI

system when deciding to accept or ignore a recommendation from the AI.

5.5. Most recent articles

The most recent articles of the literature database compiled for this review, as described in Section 3, are listed in Table 7. These papers represent the most recent production regarding trust in AI. Therefore, reviewing them after studying the proposed main path can elicit what are the research trends and possible gaps in the field. In summary, the main path reviewed in Section 5.4 starts by discussing automation and autonomous agents, surveying possible factors influencing how humans build trust in those agents. Then, it becomes clear that anthropomorphism may be used to increase trust or even to repair it. Also, as seen in the next paragraphs, recent empirical research puts great focus on human-machine teaming.

Candrian and Scherer (2022) analyze risks and premiums involved when tasks are delegated to machines. It is interesting to point out that the task delegation framework was theoretically discussed by Castelfranchi and Falcone (1998), as commented in Section 5.4. However, Candrian and Scherer (2022) conduct experiments to evaluate if people feel more comfortable in delegating tasks to AI or to human agents. Their results indicate that people prefer to delegate tasks to AI when those tasks are perceived as better approached algorithmically. This indicates that the trust necessary to delegate tasks to artificial agents is domain-dependent. Therefore, according to Candrian and Scherer (2022), in order to delegate tasks, people expect a payoff, which can be higher or lower depending on the domain. This implies that delegation to AI in a domain where humans are believed to achieve better results requires higher expected payoff.

Table 7
The 10 most recent articles in the compiled database.

Reference	Title	Journal
Naiseh et al. (2023)	How the different explanation classes impact trust calibration: The case of clinical decision support systems	International Journal of Human-Computer Studies
Leichtmann et al. (2023)	Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task	Computers in Human Behavior
Bernardo and Seva (2023)	Affective Design Analysis of Explainable Artificial Intelligence (XAI): A User-Centric Perspective	Informatics
Gomez et al. (2023)	Mitigating Knowledge Imbalance In Ai-Advised Decision-Making Through Collaborative User Involvement	International Journal Of Human Computer Studies
Kelly et al. (2023)	What Factors Contribute To The Acceptance Of Artificial Intelligence? A Systematic Review	Telematics And Informatics
Zhang et al. (2023)	Trust In An Ai Versus A Human Teammate: The Effects Of Teammate Identity And Performance On Human-Ai Cooperation	Computers In Human Behavior
Rieger et al. (2022)	Challenging Presumed Technological Superiority When Working With (Artificial) Colleagues	Scientific Reports
Schelble, Lopez, et al. (2022)	Let's Think Together! Assessing Shared Mental Models, Performance, And Trust In HumanAgent Teams	Proceedings Of The ACM On Human-Computer Interaction
Candrian and Scherer (2022)	Rise Of The Machines: Delegating Decisions To Autonomous Ai	Computers In Human Behavior
Karran et al. (2022)	Designing For Confidence: The Impact Of Visualizing Artificial Intelligence Decisions	Frontiers In Neuroscience

The intuitive idea that AI might have better performance than humans in specific domains is known as MABA-HABA (machines are better at – humans are better at) (Glikson & Woolley, 2020). As described before, this idea seems to hold in the practical experiments of Candrian and Scherer (2022). However, Rieger et al. (2022) challenge that idea. They experiment with human subjects making decisions with the aid of AI partners in the domains of X-Ray, loan assignments and visual detection. Their results show that humans tend to prefer other humans as partners, with respect to trust and responsibility in the presence of errors. However, when asked to flip positions, that is, be the subject of an assessment, humans would rather be evaluated by AI agents. This result could be related to the perception that AI is less biased than a human evaluator.

Another recent paper that reports results of experiments on teaming humans with AI and its effects on trust is Zhang et al. (2023). Their paper is focused on how the agent is framed, either as human or as artificial agent. Following previous authors, Zhang et al. (2023) build a scenario in which humans must issue a decision with the aid of a recommendation given by an artificial teammate. However, some participants were deceived: they were led to believe they were working with another human as teammate; but, in reality, they had an artificial agent as partner. As result, Zhang et al. (2023) show that humans had higher levels of trust when recommendations were issued by a teammate truthfully framed as AI, when compared with an AI teammate deceitfully identified as human. Also, regarding the practice of deceive, trust levels were significantly lowered. In summary, when deceived, humans accepted less recommendations from the AI.

As it can be concluded by previous comments, human-machine teaming has become the common ground for researching trust in artificial intelligence. Therefore, factors that influence trust are proposed and tested with some reference to the team members, such as accuracy, reliability, apologies and explanations. In a recent approach, Schelble, Flathmann, et al. (2022) vary team composition in order to assess trust, performance and shared mental models as perceived by human teammates. In their experiments, teams of 3 agents each would complete tasks in a simulation game. However, the composition of teams would be all-humans, human-human-AI or human-AI-AI. Results were as mixed as the teams compositions, but it is an interesting approach to study how team composition can affect human perception. For instance, according to Schelble, Flathmann, et al. (2022), in teams with 2 humans and 1 AI, humans perceived higher team cognition between themselves and not so much with the AI. This can be a very domain-specific result, but it deserves further research.

As seen from the previous reviewed papers, most experiments involve decision-making by a human accepting or rejecting a recommendation issued by an autonomous agent. This is a one-way approach, in which information flows only from the AI to the human in the form of a recommendation or opinion. Innovating in their approach, Gomez et al. (2023) consider information flowing also from the human to the AI. Therefore, in situations that the autonomous agent is less accurate than expected, it can have input from the human agent, improving the overall performance of the team. Using this framework in their experiments, Gomez et al. (2023) report that users have higher trust levels when involved in the process of recommendation generation by the AI. The authors argue this can be seen as a form of trust calibration.

Improvement on trust in artificial intelligence by xAI is still an active research theme. For instance, the recent approach of Karran et al. (2022) uses visualization schemes to show areas of pictures strongly considered by a classifier. With that approach for explanation on the AI output, users' confidence can be improved by reducing epistemic uncertainty. By their turn, Leichtmann et al. (2023) use xAI to improve user trust in an app. They conduct experiments in a mushroom picking virtual environment, with the aid of a visual AI system capable of providing explanations about its output. The results indicate that the users who received explanations had trust levels better calibrated.

The effects of xAI approaches on trust calibration were investigated

in the empirical research of Naiseh et al. (2023). Their experiments involved health specialists used to clinical decision support systems and different classes of xAI. They found that example-based and counterfactual explanations were the classes perceived as understandable, contributing to trust calibration. However, results obtained by Naiseh et al. (2023) indicate that the presence of explanations can also lead to over-reliance. Still on the context of xAI, Bernardo and Seva (2023) study emotions that can influence trust calibration with xAI, such as excitement, surprise or amazement. Interestingly, Bernardo and Seva (2023) conclude that xAI contributes similarly to other cues in the interaction of humans and AI for trust calibration. Also, according to Bernardo and Seva (2023), the design of the explanation system can change its effectiveness.

Finally, the search for factors that influence AI trust and acceptance is still going, as shown by the recent review conducted by Kelly et al. (2023). One of the main contributions of their review is a model for AI technology acceptance, with predictors such as perceived usefulness, performance, cultural factors and trust.

6. Discussion

In the previous sections, we provided a structured review of the literature on trust in AI. To that end, we employed objective and quantitative methods to come up with the most important works on the theme, without restrictions on year of publication or knowledge area. Then, we divided the papers into segments of the most cited, greatest bibliographic coupling, highest co-citation and the most recent papers. Also, we proposed a main path of how the literature has evolved through time. In summary, this section presents our main findings after analyzing the literature of the previous sections, as well as identified research opportunities in the theme.

Our first remark drawn from the previous analysis is that trust in AI does not have a common, unique definition. For example, Muir (1994) describes trust as an expectation from one agent towards another agent; Castelfranchi and Falcone (2000) relate trust to a belief that an agent will act on behalf of another, through delegation and goal adoption. However, the most accepted trust definition comes from Mayer et al. (1995), who build on the perspective of an agent, the trustor, accepting to be vulnerable to the actions of a trustee. Note that this definition implies that the trustor has something to lose depending on the behavior of the trustee, but still, it assumes the risk. With this respect, other authors have proposed variations of this definition (Israelsen & Ahmed, 2019; Lee & See, 2004), but they do not collide with Mayer et al. (1995)'s.

Another interesting remark is that not all papers analyzed in this literature review restrict themselves to a very rigid definition of AI, nor they seek one. It is not a matter of common agreement between authors, as commented by Kelly et al. (2023). In this context, some authors analyze trust in AI with respect to general automation (Hoff & Bashir, 2014; Lee & See, 2004; Muir, 1994), while others consider automation provided by an artificially intelligent agent (de Visser et al., 2016; Dzindolet et al., 2003; Fan et al., 2008) or a cognitive agent (Israelsen & Ahmed, 2019; Rapp et al., 2021; Schelble, Flathmann, et al., 2022). However, the works converge on human trust in an agent to perform some given task on his/her behalf or that is part of a team with a common goal. Note that by agent we mean an artificial machine, regardless of its possession of real intelligence or not. This broader view of trust in AI bypasses the discussions of what constitutes intelligence on artificial tools, focusing on the human aspect of trust towards what they consider to be AI.

As discussed, agreement on what intelligence is, in the context of artificial agents, is not required when researching trust in AI. From the analyzed literature, artificial agents are broadly defined as automated entities capable of choices. For instance, they can be recommender systems (Chong et al., 2022), self-driving car agents (Hengstler et al., 2016; Waytz et al., 2014), or machine learning classifiers (Karran et al., 2022; Schmidt et al., 2020). Although AI should not be confused with simple automation (Glikson & Woolley, 2020), it can be used to provide task automation (de Visser et al., 2016; Rieger et al., 2022). Furthermore, AI has been investigated with respect to trust when mixed among humans in collaborative efforts (Gomez et al., 2023; Schelble, Flathmann, et al., 2022; Zhang et al., 2023). Thus, the specific form of AI technology and its definition can be very flexible in the investigation of trust. More recent approaches include AI as Bayesian networks agents emulating theories of minds (Westby & Riedl, 2023) and as Large Language Models (LLMs) generating responses to users' prompts (Ouyang et al., 2022).

Still on the forms of AI investigated in the context of trust, earlier papers examined them as simple tools of automation or aid (Lee & Moray, 1992; Lee & See, 2004; Muir, 1994). By that point of view, humans act as supervisors to machines, which are seen as tools to achieve goals. Opposite to that, most recent papers tend to frame autonomous agents as part of teams, alongside humans (Cohen et al., 2021; Fan et al., 2008; Schelble, Lopez, et al., 2022; Zhang et al., 2023). That perspective implies that artificial agents can also have goals of their own, can choose to adopt common team goals or even work independently. In this regard, autonomous agents are referred to as cognitive agents, as in de Visser et al. (2014), de Visser et al. (2016), Glikson and Woolley (2020) and Zhang et al. (2023). In fact, human-machine teaming is a fast-growing research field in AI, especially considering cognitive agents, as also concluded in the review work of Pinyol and Sabater-Mir (2011).

Similar to trust definition, literature has different models of trust in AI. They are diverse and built by different approaches, but often they are reasoned by surveying the literature related to human relations, searching for factors that can influence trust. For example, as seen in Section 5.1, Lee and See (2004) conduct an extensive survey over diverse research areas to enumerate constructs, such as reputation, risk and confidence, arranging them together in a theoretical model of trust that can be extended to automation. That research pattern can also be observed in Muir (1994), Castelfranchi and Falcone (2000), Hoff and Bashir (2014), Israelsen and Ahmed (2019), Sun and Botev (2021) and Kelly et al. (2023). In this context, few authors attempt quantitative models, such as the linear model based on fault and performance proposed by Lee and Moray (1992). We argue that there is an opportunity for future research to explore quantitative models of trust in AI.

A pattern observed when authors search for factors of trust is the recurrent reference to anthropomorphism. Nass and Moon (2000), for example, show that people tend to apply social rules to computers and perceive them as having personalities. With that, it seems reasonable to research human behaviors that contribute to increasing trust and try to model them in AI agents. In this regard, Hoff and Bashir (2014) recommend increasing anthropomorphic characteristics in order for an autonomous agent to achieve higher levels of trust. Also, the empirical results from Waytz et al. (2014), de Visser et al. (2016), Araujo (2018), Kim and Song (2021) agree that trust is increased when autonomous agents display human-like behaviors or characteristics.

Regarding empirical papers, most researchers of trust in machines and AI resort to decision-making experiments. In those scenarios, a human agent has to make a decision after receiving a recommendation from an autonomous machine, usually an artificial intelligent agent such as a classifier. Within this type of experiment, trust is measured by how many recommendations are accepted by the human subjects. Examples of decision-making experiments with the aid of AI can be found in Dzindolet et al. (2003), Fan et al. (2008) and de Visser et al. (2016).

As seen from Adadi and Berrada (2018), Dazeley et al. (2021) and Karran et al. (2022), research on trust in AI can be approached by the broad field of xAI. This is related to improving transparency of the outcomes and inner workings of intelligent agents, as well as having explanations for their behavior and decisions in a humanly understandable way. In this regard, Hoff and Bashir (2014) conclude that transparency is a major factor influencing trust. More recently, Paleja

et al. (2021) show results of xAI contributing to increase human-machine teams situational awareness, which relates to higher levels of trust. However, caution is advised when trying to generalize results of xAI research. For instance, experiments conducted by Schmidt et al. (2020) concluded that transparency does not increase trust in some cases. Those conclusions are aligned with the empirical research of Alufaisan et al. (2021), who present mixed results regarding advantages of employing xAI.

Finally, some papers analyzed in this review study the importance of calibration regarding expected behavior of AI. Specifically, trust calibration refers to adequate levels of trust in an autonomous agent (Hoff & Bashir, 2014; Parasuraman & Riley, 1997). For example, failure to calibrate trust in a recommender system may imply errors of interpretation, misuse of information and poor decisions. In order to achieve adequate trust calibration, some authors provide general recommendations. For instance, Dzindolet et al. (2003) suggest initial training of users on the system's capabilities; de Visser et al. (2014) recommend displaying measures of risk and uncertainty of the autonomous agent's recommendation; and Karran et al. (2022) approach calibration with visual explanations. In this respect, Karran et al. (2022) show that xAI can be used to seek trust calibration, but Schmidt et al. (2020) argue that explanations can, in fact, harm user's decision-making capacity. Also, the work of Naiseh et al. (2021) is a recent approach on how xAI can be used to improve trust on AI, with discussions of potential problems.

Nevertheless, to the best of our knowledge, few authors proposed dynamic ways to calibrate trust, like Gomez et al. (2023) who incorporate user's inputs in situations when the AI is known to be less accurate. Dynamic calibration refers to achieving the optimal level of trust on a given system, relying on it when it is most efficient but having control over it when its performance might be poor. It has the potential to improve user experience, avoiding over reliance on the AI system and mistrust in domains in which the AI has better performance than a human agent. Also, trust calibrated dynamically can lower user training costs and foster adoption and acceptance. Furthermore, we did not identify attempts to mathematically calibrate trust in AI. This means important research opportunities, not only for innovations in ways AI adapt to human behavior but also for proposing objective measures related to trust in form of equations that can be implemented algorithmically.

Summarizing our findings on this literature review, it is clear that rigid definitions of intelligence in artificial systems or agents are not a pre-requisite to investigate trust in AI. It seems to be enough that humans perceive the artificial system as intelligent and somewhat autonomous. In fact, some researchers use no AI technology at all, investigating interactions between humans and what they are led to believe to be AI (Cohen et al., 2021; Schelble, Lopez, et al., 2022); Schelble, Lopez, et al., 2022 method called Wizard of Oz. Another finding is that the artificial agents themselves, when employed, might be implemented by diverse tools from the specialized AI literature, for instance, classifiers, regressors, reinforcement learning agents, computer vision tools, LLMs and chatbots. Also, the trust models proposed stem from various disciplines and there is no exhaustive list of influencing factors, which might explain the diversity of qualitative trust models. However, it seems clear that anthropomorphic characteristics play a major role in fostering trust in AI, while the use of xAI to that end still needs exploration. Finally, a key observation from this review is the lack of dynamic calibration methods, especially proposals sustained by underlying mathematical models.

7. Conclusion

The scientific literature on trust in automation goes far back in the past, but has recently evolved to investigate trust in AI. With the growing inclusion of intelligent autonomous agents in modern society, it is necessary to understand when AI based systems should be trusted and when to question their recommendations and override their actions. Understanding trust involves proposing models based on factors that influence human trust in general. Therefore, this review brings studies with different proposals. And most of the existing proposals are qualitative and not readily deployable or testable.

The literature review presented here used exclusively quantitative techniques, mitigating biases in the selection of the papers to be further analyzed. We followed established techniques to retrieve the most influential papers on trust in AI, along with a proposal for the main path of this literature. With the compilation of selected works, we structured the review within the key points of the papers, drawing insights from the diverse approaches and their conclusions. Then, we were able to provide a general overview of the research on the theme, highlight trends and suggest opportunities for future research.

Previous research on trust in AI advanced the understanding of human behavior when interacting with recommender tools, in the context of decision support systems, and autonomous agents, such as human-machine teams (HMT). Some questions remain without definitive answers, such as if xAI can always improve trust. On the other hand, it seems clear that anthropomorphism is a reliable tool to foster trust and acceptance. However, how those factors can be combined in a quantitative model that can measure trust and determine optimal levels is still an open research question.

Although the methods used in our research are solid and objective, they are all based on the quality of the initial compilation of documents. In Section 3, we did our best to come up with the most important papers on trust in AI, but one can never be certain that absolutely all prominent papers are included in a database search result. Thus, this is always going to be a limitation of a literature review, with results varying especially with regards to the keywords selected. Although we presented the rationale for our choices, future researchers can apply bibliometric methods with their own choices.

In conclusion, AI is a fast paced research field, with fascinating applications and results, but there is still much to cover, especially regarding quantifiable measures of trust and mathematical models of its dynamics. Another area with much to be explored is trust calibration in AI, that is, how to predict and achieve correct levels of trust in order to ensure the optimal results for a human-machine team. That aspect of AI technology is as important as performance or accuracy, since they can be lowered if the human mistrusts the artificial agent or over-relies on it. On this point, as seen before, the AI literature has some recommendations, but it is still lacking mathematical models and measures of trust calibration.

CRediT authorship contribution statement

Bruno Miranda Henrique: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Eugene Santos:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

None.

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Acknowledgements

This work was supported by Fulbright-CAPES, grant number 88881.625406/2021-01, and in part by Air Force Office of Scientific Research Grant Nos. FA9550-22-1-0022 and FA9550-20-1-0032 and Office of Naval Research Grant No. N00014-19-1-2211.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Albrecht, S. V., & Stone, P. (2018). Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence, 258*, 66–95.

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 6618–6626.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183–189.

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society, 35*, 611–623.

Batagelj, V. (2003). *Efficient algorithms for citation network analysis*. arXiv preprint cs/0309023.

Bernardo, E., & Seva, R. (2023). Affective design analysis of explainable artificial intelligence (XAI): A user-centric perspective. *Informatics, 10*, 32.

Brush, A. B., Lee, B., Mahajan, R., Agarwal, S., Saroiu, S., & Dixon, C. (2011). Home automation in the wild: Challenges and Opportunities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011*, 2115–2124.

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *AI Magazine, 26*, 53.

Buehler, M. C., & Weisswange, T. H. (2020). Theory of mind based communication for human agent cooperation. In *2020 IEEE international conference on human-machine systems (ICHMS)*. IEEE.

Candrian, C., & Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior, 134*, Article 107308.

Castelfranchi, C. (1998). Modelling social action for AI agents. *Artificial Intelligence, 103*, 157–182.

Castelfranchi, C., & Falcone, R. (1998). Towards a theory of delegation for agent-based systems. *Robotics and Autonomous Systems, 24*, 141–157.

Castelfranchi, C., & Falcone, R. (2000). Trust and control: A dialectic link. *Applied Artificial Intelligence, 14*, 799–823.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons.

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior, 127*, 1–10.

Cohen, M. C., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). The dynamics of trust and verbal anthropomorphism in human-autonomy teaming. In *2021 IEEE 2nd international conference on human-machine systems (ICHMS)*. IEEE.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6*, 94–98.

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence, 299*, Article 103525.

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). A design methodology for trust cue calibration in cognitive agents. In *Lecture notes in computer science* (pp. 251–262). Springer International Publishing.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*, 331–349.

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2019). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics, 12*, 459–478.

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human Relations, 13*, 123–139.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*, 114–126.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*, 697–718.

Evans, G. (1991). Solving home automation problems using artificial intelligence techniques. *IEEE Transactions on Consumer Electronics, 37*, 395–400.

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). The influence of agent reliability on trust in humanagent collaboration. In *Proceedings of the 15th European conference on cognitive ergonomics: The ergonomics of cool interaction* (pp. 1–8). ACM.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals, 14*, 627–660.

Gomez, C., Unberath, M., & Huang, C. M. (2023). Mitigating knowledge imbalance in AI-advised decision-making through collaborative user involvement. *International Journal of Human-Computer Studies, 172*, Article 102977.

Guevarra, M., Das, S., Wayllace, C., Demmans Epp, C., Taylor, M., & Tay, A. (2023). Augmenting flight training with AI to efficiently train pilots. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*, 16437–16439.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 53*, 517–527.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change, 105*, 105–120.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Building direct citation networks. *Scientometrics, 115*, 817–832.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature Review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications, 124*, 226–251.

Hoff, K. A., & Bashir, M. (2014). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*, 407–434.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11*, 39–63.

Israelsen, B. W., & Ahmed, N. R. (2019). "Dave...I can assure you ...that it's going to be all right" A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Computing Surveys, 51*, 1–37.

Jha, K., Doshi, A., Patel, P., & Shah, M. (2019). A comprehensive review on automation in agriculture using artificial intelligence. *Artificial Intelligence in Agriculture, 2*, 1–12.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399.

Karran, A. J., Demazure, T., Hudon, A., Senecal, S., & Léger, P. M. (2022). Designing for confidence: The impact of visualizing artificial intelligence decisions. *Frontiers in Neuroscience, 16*.

Kelly, S., Kaye, S. A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics, 77*, Article 101925.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *Journal of the Association for Information Science and Technology, 14*, 10–25.

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics, 61*, Article 101595.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*, 1243–1270.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*, 50–80.

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior, 139*, Article 107539.

Li, B.h., Hou, B.c., Yu, W.t., Lu, X.b., & Yang, C.w. (2017). Applications of artificial intelligence in intelligent manufacturing: A review. *Frontiers of Information Technology amp; Electronic Engineering, 18*, 86–96.

Liu, X., Han, Y., Bai, S., Ge, Y., Wang, T., Han, X., Li, S., You, J., & Lu, J. (2020). Importance-aware semantic segmentation in self-driving with discrete wasserstein training. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 11629–11636.

Liu, J. S., Lu, L. Y., Lu, W. M., & Lin, B. J. (2013). Data envelopment analysis 1978–2010: A citation-based literature survey. *Omega, 41*, 3–15.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*, 629–650.

Lu, Y., Xu, X., & Wang, L. (2020). Smart manufacturing process and system automation–A critical review of the standards and envisioned scenarios. *Journal of Manufacturing Systems, 56*, 312–325.

Lukyanenko, R., Maass, W., & Storey, V. C. (2022). Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. *Electronic Markets, 32*, 1993–2020.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). *An integrative model of organizational trust* (Vol. 20, pp. 709–734). The Academy of Management Review.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology. *ACM Transactions on Management Information Systems, 2*, 1–25.

Muir, B. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*, 1905–1922.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies, 169*, Article 102941.

Naiseh, M., Cemiloglu, D., Thani, D. A., Jiang, N., & Ali, R. (2021). Explainable recommendations and calibrated trust: Two systematic user errors. *Computer, 54*, 28–37.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*, 81–103.

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology, 11*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics, 55*, 1059–1072.

Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., & Gombolay, M. (2021). The utility of explainable AI in Ad Hoc HumanMachine teaming. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 610–623). Curran Associates, Inc.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*, 230–253.

Parasuraman, R., Sheridan, T., & Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30*, 286–297.

Pinyol, I., & Sabater-Mir, J. (2011). Computational trust and reputation models for open multi-agent systems: A review. *Artificial Intelligence Review, 40*, 1–25.

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies, 151*, Article 102630.

Rieger, T., Roesler, E., & Manzey, D. (2022). Challenging presumed technological superiority when working with (artificial) colleagues. *Scientific Reports, 12*.

Saam, N., & Reiter, L. (1999). Lotka's law reconsidered: The evolution of publication and citation distributions in scientific fields. *Scientometrics, 44*, 135–155.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot?. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. ACM.

Saßmannshausen, T., Burggräf, P., Wagner, J., Hassenzahl, M., Heupel, T., & Steinberg, F. (2021). Trust in artificial intelligence within production management – an exploration of antecedents. *Ergonomics, 64*, 1333–1350.

Schelble, B. G., Flathmann, C., McNeese, N. J., Freeman, G., & Mallick, R. (2022). Let's think together! Assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction, 6*, 1–29.

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2022). Towards ethical AI: Empirically investigating dimensions of AI ethics, trust repair, and performance in human-AI teaming. Human factors. *The Journal of the Human Factors and Ergonomics Society*, 1–19.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems, 29*, 260–278.

Small, H. (1973). Co-Citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology, 24*, 265–269.

Sun, N., & Botev, J. (2021). Intelligent autonomous agents and trust in virtual reality. *Computers in Human Behavior Reports, 4*, Article 100146.

Turing, A. M. (1950). *Mind LIX*.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117.

Westby, S., & Riedl, C. (2023). Collective intelligence in human-AI teams: A bayesian theory of mind approach. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*, 6119–6127.

Wilder, B., Horvitz, E., & Kamar, E. (2020). Learning to complement humans. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence* (pp. 1526–1533). International Joint Conferences on Artificial Intelligence Organization.

Wollschlaeger, M., Sauter, T., & Jasperneite, J. (2017). The future of industrial communication: Automation networks in the Era of the internet of things and industry 4.0. *IEEE Industrial Electronics Magazine, 11*, 17–27.

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns, 3*, Article 100455.

Zhang, G., Chong, L., Kotovsky, K., & Cagan, J. (2023). Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Computers in Human Behavior, 139*, Article 107536.