You Haven't Changed a Bit! Initial Findings from a Bibliometric Analysis of Two Decades of Empirical Trust in AI Research

MICHAELA BENK, Mobiliar Lab for Analytics, ETH Zurich, Switzerland
SOPHIE KERSTAN, ETH Zurich, Switzerland
ANDREA FERRARIO, Mobiliar Lab for Analytics, ETH Zurich, Switzerland

Trust is widely regarded as a critical component in interactions between humans and artificial intelligence (AI), yet its measurement remains a critical challenge for the human-computer interaction (HCI) research community. In this work, we present preliminary results of an ongoing large-scale bibliometric analysis of two decades of empirical research measuring trust in AI, comprising 538 core articles and 15'551 cited articles across multiple disciplines. Our aim is to provide insights into trends and trajectories in empirical trust in AI research that may inform future research. Initial findings reveal that technological foci have changed over the years, while seminal works from the trust in automation literature remain influential. We discuss implications, as well as next steps in our ongoing work.

Additional Key Words and Phrases: Trust, Artificial Intelligence, Methodology, Bibliometric Analysis

ACM Reference Format:
Michaela Benk, Sophie Kerstan, and Andrea Ferrario. 2023. You Haven't Changed a Bit! Initial Findings from a Bibliometric Analysis of Two Decades of Empirical Trust in AI Research. In *CHI TRAIT '23: Workshop on Trust and Reliance in AI-Human Teams, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXX

1 INTRODUCTION

The landscape of user studies that measure trust in AI has grown significantly over the past years. Although trust between humans has been studied for decades within various disciplines [5, 6], the scientific discussions on how to define and formalize trust in the context of AI is ongoing [10, 14]. As a consequence of theoretical disagreements, the operationalization and measurement of trust in AI remains a critical challenge for the research community [25]. The lack of consensus on how trust in AI is formalized, and thus, how it is measured in different contexts, has led to a variety of measurement approaches, including psychometric instruments [15, 19], task-dependent behavioral metrics [25], or ad-hoc questionnaires [27], resulting in a vast landscape of measurement instruments to choose from and little guidance on their appropriate use.

To this end, researchers across disciplines have started to address the question of how to measure trust in AI in a coherent and valid way. In the field of Human-Computer Interaction (HCI), premier academic conferences such as the ACM Conference on Human Factors in Computing Systems (CHI) or the Conference on Mobile Human-Computer Interaction (MobileHCI) have organized workshops and tutorials in which trust and its measurement are central foci (see for instance the CHI Workshop on Trust and Reliance in AI-Human Teams [1] or the MobileHCI tutorial on Integrating Trust Measurements into Experimental Designs [30]). Furthermore, various works have discussed specific

Manuscript submitted to ACM

issues surrounding the measurement of trust, such as the lack of risk in empirical works measuring trust [14], the dangers of 'proxy tasks' [4], the inconsistencies between different types of scales [26] or their utility [3], or the need for context-specific trust in AI scales [7]. Moreover, a number of review articles have appeared in recent years, aimed at structuring the discussion on how to evaluate trust in AI [11, 29].

While these efforts have provided valuable insights into various applications of existing measurement techniques and their challenges, it remains unclear if and how research on trust has evolved over time, if different approaches to its definition and measurement exist across disciplines, and what emerging or burgeoning trends can inform future research. These circumstances highlight the need for a mapping of the vast scientific landscape, including trends and trajectories of the research activities of empirical trust in AI research. In this way, researchers may better understand the roots, trends, and developments in empirical works measuring trust in AI and identify new directions and opportunities to move forward.

Motivated by this background, we conducted a large-scale bibliometric analysis of two decades of empirical research measuring trust in AI, comprising 538 core articles and 15'551 cited articles across multiple disciplines. We further enhanced our quantitative findings with a qualitative analysis of the most influential articles per discipline. In this work, we present initial results of the ongoing analysis. We thereby aim to answer the following research question:

RQ: What are emerging or burgeoning trends with respect to the main research streams and influential works in empirical research measuring trust in AI?

Focusing on trends allows us to (a) explore types of AI technologies that have been placed in the trustee role when measuring trust, and (b) explore which works have been influential in shaping the evolution of the measurement (and its methods) of the construct "trust in AI" across disciplines.

Our results indicate that different technological foci of AI have emerged over the years, including trust in recommender systems, robots, and autonomous vehicles. Nevertheless, the most influential articles (i.e., the most cited works by our dataset over the time period under consideration) have largely remained the same over the years. They include the seminal work by Lee and See [18] and Jian et al. [15], which focus on trust in automation, as well as Mayer et al. [20], which focus on human trust in the organizational context. We discuss implications and conclude with an outlook of this research.

2 METHODOLOGY

We conducted a comprehensive search of peer-reviewed literature available from 2000 to 2021, using the electronic databases Scopus, Web of Science, ACM Digital Library, IEEE Xplore Digital Library, PubMed, and APA PsycNet. The databases were chosen to broadly cover different interdisciplinary lenses on the measurement of trust in AI. The search used a combination of keywords and controlled vocabulary for the concepts of AI, trust, and measurement. The search was piloted in Scopus and ACM Library by two investigators independently and subsequently adapted to each database. The search query can be found in section A of the appendix.

We used the following criteria for study selection. Articles needed to be (a) published between the years 2000 and 2021 in English language in a journal, conference proceeding, or book chapter; (b) contain quantitative or qualitative empirical research; (c) measure trust between human subjects and AI. Studies were excluded if (a) they measured trust in something other than AI (e.g., the organization providing the AI); (b) measured other constructs than trust, such as reliance or propensity to trust; (c) did not include actual participants (i.e., a simulated study or proposal); (d) did not

adequately report the employed methodology.

After the full-text review, we extracted the following meta-information from the included articles, using the Scopus API: (a) citation information, including author names, document and source titles, and citation counts; (b) bibliographical information, including affiliations and publisher; (c) abstract & keywords, including author keywords; (d) all references. A total of 538 articles and 15'551 citing articles were included in the dataset of curated articles and used for further analysis.

2.1 Bibliometric Techniques

The quantitative techniques used in this work are based on bibliometrics, a commonly used statistical method for literature reviews. Bibliometrics uses meta-data, such as number of citations, references, author information, and keywords, to characterize the intellectual structure and development of a research field [8, 21]. Unlike traditional literature reviews, bibliometrics benefits from a variety of statistical approaches. The most commonly used bibliometric techniques, which we employed in this work include (a) (co-)citation analyses and bibliometric coupling, using citation data to construct measures of influence and similarity; (b) co-author analysis, using co-authorship data to measure collaboration; and (c) co-occurence networks of textual data to find connections among concepts that co-occur in document titles, keywords, or abstracts. Another main benefit of bibliometric methods is their potential to objectively evaluate the developments of research streams, topics, or authorships over time. Using citation counts as its primary unit of measurement, bibliometrics is based on the assumption that citations can be used as a measure of utility and as a means to identify the intellectual structure of a research field, where highly cited papers are considered influential or impactful [8, 21].

Two main limitations of bibliometrics are: (a) its success can be hampered by the quality of the available data, and (b) interpretations solely based on citation and occurrence counts cannot offer insights into the thematic content of the articles [24]. We took several measures to overcome these challenges. First, by following the PRISMA protocol and carefully selecting each article, we created a curated, high-quality database of core articles that fit our criteria. Secondly, we employed a combination of automatic pre-processing, using a Python string-matching library[1] and manual processing to clean and harmonize attributes such as author keywords and references. Lastly, we supplemented the insights derived from the statistical analyses with qualitative assessments of the most influential articles in our database. To do so, three researchers independently and manually coded selected articles according to a pre-determined list of classification items.

3 PRELIMINARY RESULTS

As can be seen in Table 1, our dataset covers a wide range of disciplines, including the physical, social, and life sciences. Purely interdisciplinary journals (e.g., PLOS one) and conferences were categorized as "Interdisciplinary." Disciplines with 3 occurrences or less (e.g., Tourism) were grouped into the category "Other." The majority of articles (60%) originate in technology-focused domains, such as human-robot interaction, computer science, or robotics.

The number of articles per year show an increasing trend of empirical works measuring trust, with an exponential increase over the last 3 years. The majority of articles (56%) are published in conference proceedings, followed by journal publications (43%) and book chapters (1%).

[1] https://github.com/maxbachmann/RapidFuzz

Discipline	# Articles
Human-Robot Interaction	95
Computer Science	93
Human-Computer Interaction	70
Robotics & Engineering	70
Human Factors, Ergonomics, & Psychology	56
Interdisciplinary	57
Medicine & Health Informatics	24
Information Systems	20
Other	20
Transport	17
Business, Management, & Marketing	16

Table 1. Main disciplines in the dataset.

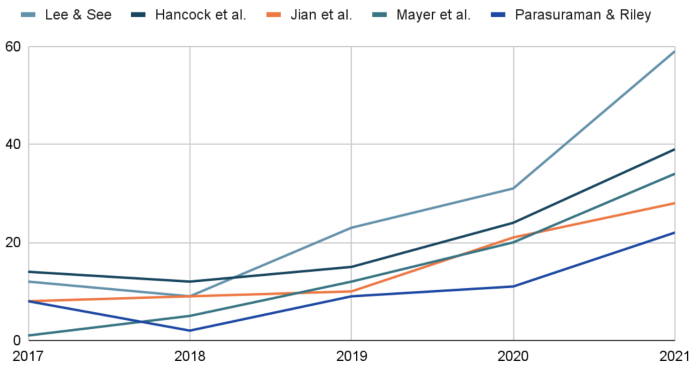


Fig. 1. Citation trend of the most cited articles in the last five years.

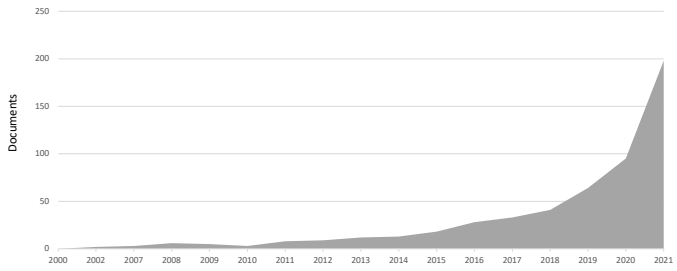


Fig. 2. Number of articles per year.

3.0.1 AI: Technological foci. Results of the co-occurrence analysis of titles and abstracts, using VOSViewer (Version 1.6.8) [28], revealed different major AI foci over time in our dataset. Figure 3a reveals main thematic clusters of that include several types of AI: (1) algorithmic decision-making and explainable AI (XAI), (2) robots, including social and humanoid robots, (3) autonomous vehicles, (4) recommender systems, (5) AI in service and consumer settings. Moreover, whereas empirical research on trust in recommender systems has seen a decline in publications over time, trust in algorithmic decision-making and explainability appeared later and has seen an increase in publications. Notably, trust in human-robot interactions has remained prevalent over time. In particular, Figure 3b shows changes over the past five years in more detail. While the technologies "recommender system" and "robot" appear early in the literature (before 2018), the research focus on "autonomous vehicle" (2019) and "artificial intelligence (systems)" (2020-2021) is more recent. Furthermore, meta-considerations on technology, e.g., "concern", "user acceptance" "intention", and "adoption" are also quite recent.

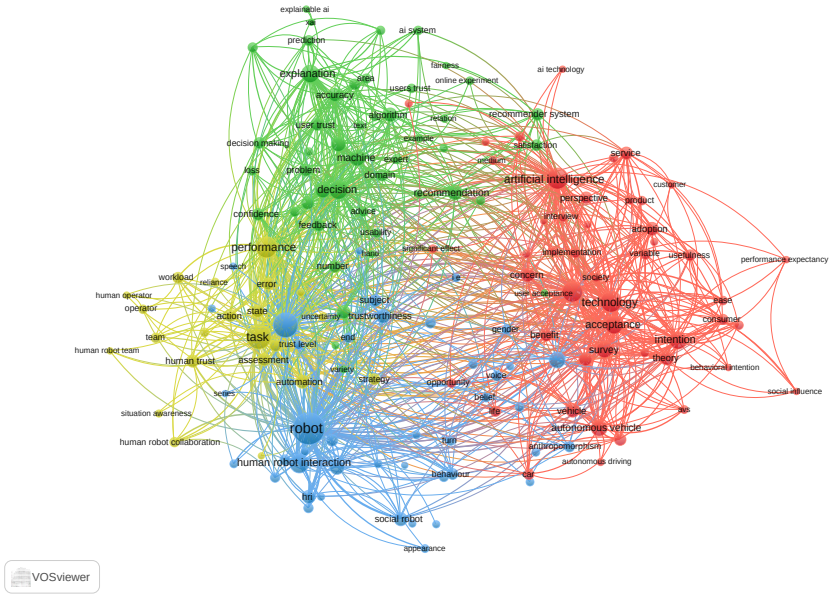
3.0.2 Influential works. Table 2 provides an overview over the ten most cited works by our dataset. One third of all articles in our dataset cite the seminal work by Lee and See [18] and a majority of influential articles originate in the human factors and ergonomics domain. An investigation of citation trends revealed that the seminal works by Lee and See [18], Hancock et al. [12] and Jian et al. [15] continuously remain the most cited over time. Interestingly, the work by Mayer et al. [20] was not widely cited before 2014 and became the third most cited in 2021, indicating a renewed interest in this work. Figure 2 shows the citation trends of the five overall most cited works within the last five years.

4 DISCUSSION

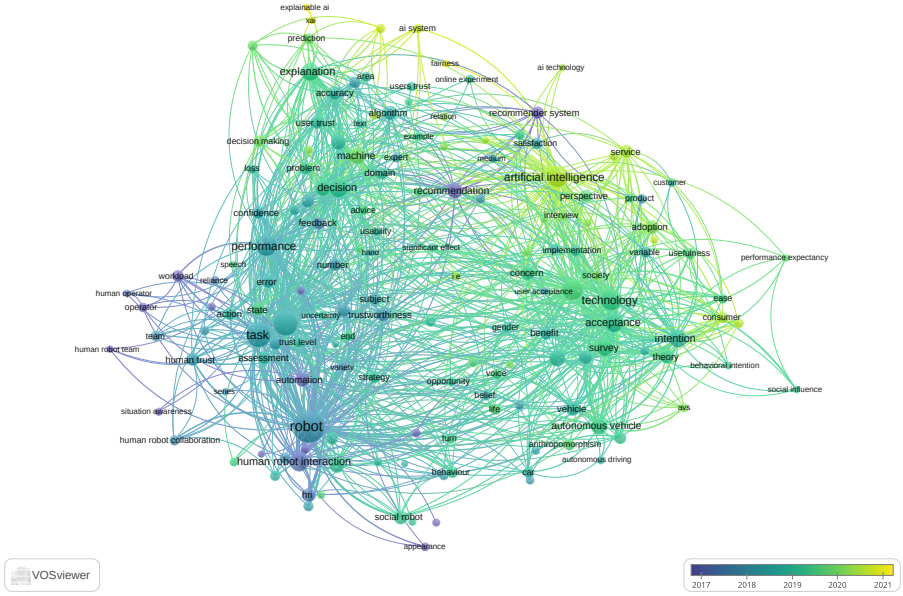
Our initial results reveal that, while there has been a shift in technological foci on AI over the years (e.g., burgeoning research trust in recommender systems and robots and emerging research on trust in artificial intelligence systems and AI explainability methods), seminal works in automation [15, 18] remain the most influential over time, and the majority of the most cited articles focus on trust in automation research. The seminal work by Mayer et al. [20] from the organizational behavior domain has seen a new-found interest. While the citation count does not reveal to what extent the models or metrics proposed by these works inform empirical works, they nevertheless show that trust in automation research continues to influence the trust in AI research landscape.

These findings warrant an important question for the validity and coherence of trust measurement: are the influencing factors and mechanisms of trust in automation the same as those of trust in various forms of artificial intelligence?[2] The technological focus of the works in automation show commonalities: for instance, they mainly focus on autopilots and automated navigation systems. However, while automation frequently refers to technology that follows programmed rules, AI has been defined as technology "imitating human behavior and decision-making capabilities" [16] (e.g., by deploying machine learning methods). These AI abilities, in turn, might trigger users' perceptions and trusting processes in a different way than automation. This may also explain the new-found interest in the work of human trust by Mayer et al. [20], which may better model certain human-AI trust interactions. In summary, our preliminary results show that more research is needed to address the differences between automation and AI and their characterization of the interactions of these technologies with human users, to determine whether existing theoretical models and measurement methods of trust in automation may translate to trust in AI.

[2]Or, as the title of this work suggests, are we perhaps looking the other way to avoid addressing any possible changes that may have occurred over time in the technology that we empirically examine?



(a)



(b)

Fig. 3. Thematic mapping, using abstract and titles. Figure (a) shows the overall research clusters. Figure (b) shows their development in the last five years.

R	Title	Authors	Journal	Contribution	Trustee (examples)	TC
1	Trust in automation: Designing for appropriate reliance	Lee and See [18]	Human Factors	Theoretical model	Automation (autopilot, automated navigation systems)	164
2	A meta-analysis of factors affecting trust in human-robot interaction	Hancock et al. [12]	Human Factors	Survey of Influencing Factors	Robots (robotic home assistants, robotic devices)	126
3	Foundations for an empirically determined scale of trust in automated systems	Jian et al. [15]	Int. Journal of Cognitive Ergonomics	Measurement instrument development	Automation (aircrafts, cruise control)	92
4	An integrative model of organizational trust	Mayer et al. [20]	Academy of Management Review	Theoretical model	Humans (employees)	86
5	Humans and Automation: Use, Misuse, Disuse, Abuse	Parasuraman and Riley [22]	Human Factors	Survey of Influencing Factors	Automation (aviation, manufacturing, ground transportation)	71
6	Trust in automation: Integrating empirical evidence on factors that influence trust	Hoff and Bashir [13]	Human Factors	Survey & Model	Automation (flight management systems, GPS route planners, decision-support systems)	61
7	Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots	Bartneck et al. [2]	International Journal of Social Robotics	Review of measurement instruments	Robots (service, entertainment robots)	57
8	Trust, control strategies and allocation of function in human-machine systems	Lee and Moray [17]	Ergonomics	Experiment & Model	Automatic controllers (pasteurization plant)	47
9	Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust	Salem et al. [23]	HRI	Experimental study	Robots (Humanoid robotic assistants)	28
10	The role of trust in automation reliance	Dzindolet et al. [9]	International Journal of Human-Computer Studies	Experimental study	Automation (flight management systems)	45

Table 2. The 10 most cited articles in our dataset. TC = total citation count.

5 CONCLUSION AND FUTURE WORK

In this work, we presented initial results of a large-scale bibliometric analysis of two decades of trust in AI research. Particularly within the HCI community, the empirical measurement of trust continues to prove a critical challenge. Through our ongoing work, we aim to provide an interdisciplinary perspective of how empirical research of trust in AI has evolved over time. Our research will answer additional research questions, including whether there are different prevalent methodological approaches within different disciplines, as well as further trends and trajectories that can inform future empirical trust in AI research.

REFERENCES

[1] Gagan Bansal, Alison Marie Smith-Renner, Zana Buçinca, Tongshuang Wu, Kenneth Holstein, Jessica Hullman, and Simone Stumpf. 2022. Workshop on Trust and Reliance in AI-Human Teams (TRAIT) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 116, 6 pages. https://doi.org/10.1145/3491101.3503704

[2] Christoph Bartneck, Dana Kulić, Elizabeth A. Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1 (2009), 71–81.

[3] Michaela Benk, Suzanne Tolmeijer, Florian von Wangenheim, and Andrea Ferrario. 2022. The Value of Measuring Trust in AI - A Socio-Technical System Perspective. *CHI 2022 - Workshop on Trust and Reliance in AI-Human Teams (TRAIT)* (2022).

[4] Zana Buccinca, Phoebe Lin, Krzysztof Z Gajos, and Elena Leah Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI)* (2020).

[5] S. Castaldo, K. Premazzi, and F. Zerbini. 2010. The Meaning(s) of Trust. A Content Analysis on the Diverse Conceptualizations of Trust in Scholarly Research on Business Relationships. *Journal of Business Ethics* 96 (2010), 657–668.

[6] Christiano Castelfranchi and Rino Falcone. 2010. *Trust Theory: A Socio-Cognitive and Computational Model* (1st ed.). Wiley Publishing.

[7] Oscar Hengxuan Chi, Shizhen Jia, Yafang Li, and Dogan Gürsoy. 2021. Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Comput. Hum. Behav.* 118 (2021), 106700.

[8] Mary J. Culnan. 1986. The intellectual development of management information systems, 1972-1982: a co-citation analysis. *Management Science* 32 (1986), 156–172.

[9] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *Int. J. Hum. Comput. Stud.* 58 (2003), 697–718.

[10] Andrea Ferrario, Michele Loi, and Eleonora ViganÃ². 2019. In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philosophy & Technology* (Oct. 2019), 1–17. https://doi.org/10.1007/s13347-019-00378-3

[11] Ella Glikson and A. Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *The Academy of Management Annals* 14 (2020), 627–660.

[12] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of Human Factors and Ergonomics Society* 53 (2011), 517 – 527.

[13] Kevin A. Hoff and Masooda N. Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Hum. Factors* 57 (2015), 407–434.

[14] Alon Jacovi, Ana Marasović, Tim Miller, and Y. Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 624–635.

[15] Jiun-Yin Jian, A. Bisantz, C. Drury, and J. Llinas. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4 (2000), 53–71.

[16] Davinder Kaur, Suleyman Uslu, Arjan Durresi, Sunil Badve, and Murat Dundar. 2021. Trustworthy Explainability Acceptance: A New Metric to Measure the Trustworthiness of Interpretable AI Medical Diagnostic Systems. In *Conference on Complex, Intelligent, and Software Intensive Systems*. Springer, 35–46.

[17] J. Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35 10 (1992), 1243–70.

[18] John D Lee and Katrina A See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* (2004), 31.

[19] Maria Madsen and Shirley Gregor. 2000. Measuring Human-Computer Trust. In *Proceedings of the 11th Australasian Conference on Information Systems*. 6–8.

[20] Roger C. Mayer, James Herbert Davis, and F. David Schoorman. 1995. An Integrative Model Of Organizational Trust. *Academy of Management Review* 20 (1995), 709–734.

[21] Sridhar P. Nerur, Abdul A. Rasheed, and Vivek Natarajan. 2008. The intellectual structure of the strategic management field: an author co-citation analysis. *Southern Medical Journal* 29 (2008), 319–336.

[22] Raja Parasuraman and Victor A. Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of Human Factors and Ergonomics Society* 39 (1997), 230 – 253.

[23] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2015), 1–8.

[24] Silvia Salini. 2016. *An Introduction to Bibliometrics.* John Wiley Sons, Ltd, Chapter 14, 130–143. https://doi.org/10.1002/9781118763025.ch14 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118763025.ch14

[25] Philipp Schmidt and Felix Biessmann. [n. d.]. Quantifying Interpretability and Trust in Machine Learning Systems. *AAAI-19 Workshop on Network Interpretability for Deep Learning.*

[26] Jan Maarten C. Schraagen, Pia Elsasser, H.L.A. Fricke, Marleen Hof, and Fabyen Ragalmuto. 2020. Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64 (2020), 339 – 343.

[27] Shervin Shahrdar, Corey Park, and Mehrdad Nojoumian. 2019. Human Trust Measurement Using an Immersive Virtual Reality Autonomous Vehicle Simulator. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019).

[28] Nees Jan van Eck and Ludo Waltman. 2009. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 84 (2009), 523 – 538.

[29] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1 – 39.

[30] Philipp Wintersberger and Brittany E. Holthausen. 2020. Integrating Trust Measurements into Experimental Designs. *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (2020).

A SEARCH QUERY

We ran a search of publications on the databases Scopus, Web of Science, ACM Digital Library, IEEE Xplore Digital Library, PubMed, and APA PsycNet, using a combination of terms related to *AI* ("artificial intelligence" OR "intelligent agent" OR "neural network" OR "deep learning" OR "machine learning" OR "learning algorithm" OR robot* OR "autonomous car" OR "autonomous vehicle" OR "natural language processing" OR "recommender system" OR "ai" OR "xai" OR "expert system"), *trust* (trust*), and *measurement* (measure* OR experiment* OR empiric* OR assess* OR "questionnaire" OR ("survey" AND NOT "taxonomy") OR "user study" OR ("human evaluation" AND NOT ("trust network" OR "social trust" OR "taxonomy")). The search included English publications from the years 2000 until 2021.