Computational models of artificial and natural trust in robotics: A systematic review and operational guide.

VINANZI, Samuele <http://orcid.org/0000-0003-0241-9983>, ROMEO, Marta <http://orcid.org/0000-0003-4438-0255>, CANGELOSI, Angelo <http://orcid.org/0000-0002-4709-2243> and SEMERARO, Francesco <http://orcid.org/0000-0002-8812-0968>

Original article

The International Journal of Robotics Research
2026, Vol. 0(0) 1–39

DOI: 10.1177/02783649261459131
journals.sagepub.com/home/ijr

Computational models of artificial and natural trust in robotics: A systematic review and operational guide

Samuele Vinanzi[1,*], Marta Romeo[2,*], Angelo Cangelosi[3] and Francesco Semeraro[3,4,*]

Abstract

Trust forms the bedrock of successful human interactions, and its integration into human–robot collaboration remains a critical challenge. Contemporary research predominantly explores human trust in robotic systems, focusing on refining the appearance and behavior of artificial agents to foster their acceptability in social settings. However, this systematic review centers on the less-explored dimension of trust mechanisms within autonomous robotic systems. Our aim is to survey what we define as Computational Trust models, which encompass a robot's capability to both assess the trustworthiness of other agents ("Artificial Trust") and to predict their levels of trust towards itself ("Natural Trust"). To achieve this objective, an initial set of 1916 papers, ranging from 2013 to 2023, was collected from IEEE Xplore, Scopus, and ISI Web of Science. Eligibility criteria were then applied to this set to select works that designed a Computational Trust model for a robotics application, which was validated through an experiment. These criteria were agreed upon by all authors to ensure unanimous decisions on whether to retain or remove results. At the end of this process, 101 key papers were identified. Following the selection process, we conducted thorough analyses to cluster these works based on the type of Computational Trust model used, the application domain, the robotic platforms employed in the validation, the experimental design, and the evaluation metrics. Finally, we identify common trends in this emerging branch of Human–Robot Interaction and provide guidelines for scholars wishing to contribute to this field.

Keywords

trust, robotics, artificial intelligence, human–robot interaction, quantitative modeling

1. Introduction

Rapid advancements in robotics and Artificial Intelligence (AI) are leading to the increasing integration of these technologies into our daily lives. AI systems already surround us and have emerged as a significant driving force in the global economy, with robotics expected to follow. As these autonomous systems become more pervasive, the necessity for them to be able to seamlessly interact with humans becomes even more critical. Effective interaction necessitates these machines to possess not only technical proficiency but also the nuanced social abilities inherent in human interaction. In other words, to integrate robots into our daily lives, we should equip them with similar social and cognitive skills as the ones that enable us to function as members of society.

This requirement is driving growing interest within both industry and academia. Researchers and scholars are focusing their efforts on developing computational models that are able to replicate some of our mental capabilities, in an effort to make these artificial entities a little more human-like (Vinanzi, 2021). One of these abilities is the one that allows us to reason about trust: a fundamental and inevitable component of social interactions.

[1]Centre of Excellence in AI and Robotics (CEAIR), School of Computing and Digital Technologies, Sheffield Hallam University, Sheffield, UK
[2]School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK
[3]Manchester Centre for Robotics and AI, The University of Manchester, Manchester, UK
[4]Human-Robot Interfaces and Interaction (HRI2) Research Unit, Italian Institute of Technology, Genova, Italy

*These authors contributed equally to this work.

Corresponding author:
Samuele Vinanzi, Centre of Excellence in AI and Robotics (CEAIR), School of Computing and Digital Technologies, Sheffield Hallam University, Howard St, Sheffield City Centre, Sheffield S1 1WB, UK.
Email: s.vinanzi@shu.ac.uk

Trust has proven difficult to define, because of the sheer range of situations in which it can be applied. For instance, it has profoundly different meaning in a financial relationship than a sentimental one. A broad definition, which is able to generalize across situations, comes from Mayer et al. (1995): trust is the willingness of one party (the trustor) to rely on the actions of another party (the trustee), with the former having no control over the latter. In essence, it represents the trustor's willingness to delegate responsibility for a task to the trustee and to accept the associated risk. Trust permeates all types of social interactions and plays a pivotal role in fostering successful relationships, ensuring personal safety (Das and Teng, 2004), and facilitating team cooperation (Jones and George, 1998). The consequences of misplaced trust, stemming from an incorrect estimation of someone's trustworthiness, can lead to severe damage, either economic, emotional, or even physical. Therefore, proficiency in this skill is a survival requirement for humans.

Given how important trust is for us (Khavas et al., 2020), researchers in the field of Human-Robot Interaction (HRI) have carried out a significant amount of research regarding this topic. Consequently, there have been attempts in providing standards to incorporate a trust component in HRI designs. For instance, Hancock et al. (2021) gave guidelines about the factors that affect trust during an interaction. In their work, they mainly explored a trust relationship in which the human is the trustor and the robot is the trustee. They listed and discussed these factors, dividing them into three categories: human-related factors (e.g., attitude towards robots and engagement), robot-related factors (e.g., dependability and adaptability), and environmental factors (e.g., team communication and physical environment). For many years, the core of research on trust in HRI has focused on identifying the parameters that empirically influence different levels of human trust.

Recently, researchers have started exploring the problem of mathematically modeling the trust of humans towards robots (Wang et al., 2023c). They identified two possible pathways, which make use of, respectively, deterministic and probabilistic methods. The former uses performance variables measured during the task, auto-regressive models, moving average ones or combinations of both. Probabilistic approaches model trust as a probability distribution influenced by other variables in a graph model. An example of the latter is OPTIMo (Xu and Dudek, 2015a), a probabilistic trust inference model based on Dynamic Bayesian networks. Other ways of modeling trust are inherited from other disciplines, like Machine Learning, formal methods and Game Theory (Kok and Soh, 2020). Although these models are seldom employed to influence a robot's behavior during interactions (Wang et al., 2023c), researchers have recently started embedding trust awareness into robotic behavior. More recent research explores the possibility of enabling robots to trust humans. Indeed, some researchers have developed trust models that allow a robot to gauge trust towards users during interactions or collaborations (Sanders and Nam, 2021). A subset of this research even argues that trust models should be bilateral to capture the full spectrum of trust dynamics between humans and robots (Sanders and Nam, 2021; Zonca and Sciutti, 2021).

In this paper, we propose the use of the term **"Computational Trust" (CT)** to define the mathematical models that can be used by a robot or a non-embodied artificial agent to perform trust evaluations on other agents. With this term, we are encompassing both the cases in which the agent acts as the trustor, a domain known as **"Artificial Trust" (AT)** (Azevedo-Sa et al., 2021; Jorge et al., 2022), and the ones where the agent is assessing the trust levels of another entity towards itself, which we refer to as **"Natural Trust" (NT)**. Figure 1 summarizes this definition.

This fertile area of research is still in its early development but holds significant benefits, especially for collaborative or socially assistive service robots. For example, in the context of providing care to the elderly, robots could assess trust and use it in various aspects, such as determining when the user needs assistance (Wilson et al., 2023). Likewise, robots tasked with security and surveillance in a shop could employ trust estimation mechanisms to identify potentially suspicious behaviors, thereby aiding in threat detection and risk assessment (Kousar Nikhath et al., 2023). Trust can also be a critical factor in joint tasks where humans and robots depend on each other's efforts to achieve a shared goal: whereas a robot can fail, so can a person, and for an artificial agent to know when to trust or distrust somebody and adapt its plans to this prediction can make all the difference in the success or failure of the task. Consider, for instance, a service robot assisting someone in setting up a table (Vinanzi and Cangelosi, 2024). If the robot learns that the individual frequently drops dishes while handling them, it may begin to distrust the person's ability to perform that specific task. Consequently, the robot might offer to take charge of this role, suggesting the individual to take care of other tasks for which they are trusted, such as fetching and placing cutlery.

To the best of our knowledge, there is no previous attempt of producing a review with the purpose of collecting and formalizing the knowledge in this newborn field of study. Our main contribution is to summarize and synthesize existing research on CT, to provide an overview of the current state of knowledge, to identify trends and to offer insights for future research.

Although this review focuses on models of CT in robotics, it is essential to recognize that these models are ultimately grounded in human social cognition. Trust, as conceptualized in interpersonal and organizational psychology, is not merely a function of observed behavior or performance metrics: it is a relational, context-sensitive phenomenon. One of the most influential frameworks in

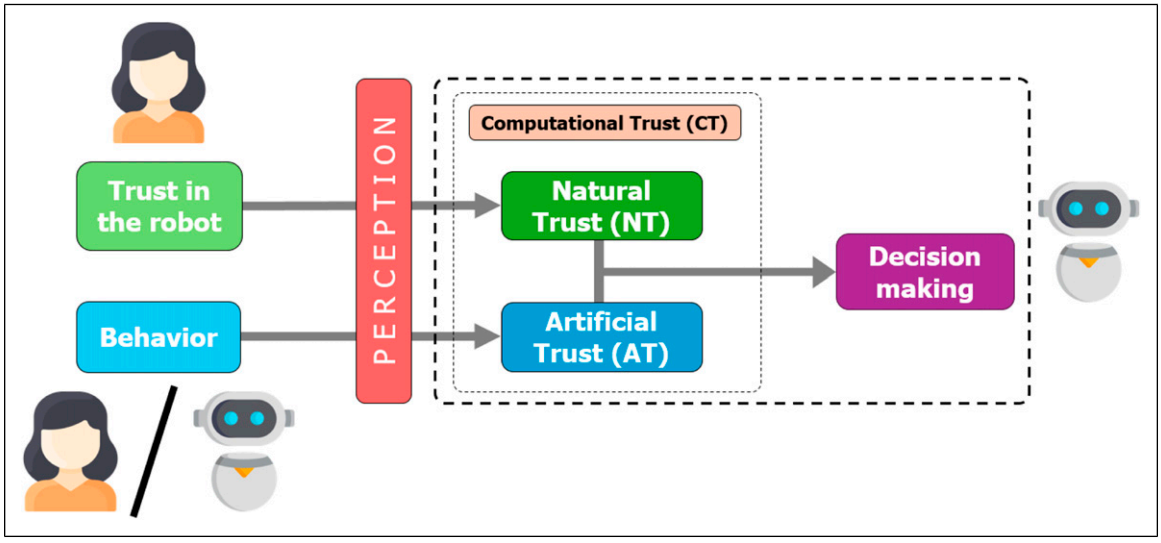

Figure 1. Our definition of Computational Trust (CT) includes both Natural Trust (NT), where the robot estimates the trust of another agent toward itself, and Artificial Trust (AT), where the robot acts as a trustor toward another agent. The agent may be either a human or another robot.

interpersonal trust, proposed by Mayer et al. (1995), defines trust as a willingness to accept vulnerability based on perceived ability, benevolence, and integrity. These dimensions remain highly pertinent in HRI, where users frequently anthropomorphize artificial agents and rely on social heuristics to assess their trustworthiness. Lee and See (2004) extended this understanding to human–automation interaction, emphasizing trust calibration, how trust should dynamically align with system capabilities, and drawing directly from interpersonal trust theory. Bainbridge et al. (2011) further demonstrated that physical embodiment significantly shapes trust-related behaviors, echoing interpersonal dynamics such as personal space and compliance. More recently, Schaefer et al. (2016) synthesized decades of empirical work in a meta-analysis, identifying key antecedents of trust in automation, including transparency, reliability, and feedback. Building on this foundation, Schäfer et al. (2024) proposed a relational, trustor-centered definition of trust in robots, emphasizing intentionality and the asymmetry of trust as a directed relation. Together, these works underscore that trust in automation is best understood not as a novel phenomenon but as an extension of well-established interpersonal trust processes, adapted to the unique affordances and constraints of human–machine interaction. Consequently, CT models, particularly those that aim to predict or influence human behavior, must move beyond static, performance-based metrics and incorporate mechanisms for modeling reciprocity, repair, and social signaling. Embedding these human-centered principles into CT architectures is essential for developing robotic systems that are not only functionally competent but also socially intelligible and ethically aligned.

This paper is structured as follows. Section 2 reports the main criteria adopted to achieve a selection of paper suitable to highlight trends and challenges regarding the described topic. Section 3 discusses the domains of application of the experiments performed by the selected papers. Section 4 briefly introduces a taxonomy of robots that have been utilized in CT experiments. Section 5 provides insights in the computational models that have been used to endow robots with trust capabilities. Section 6 analyzes the different experimental settings that have been adopted by the community. Section 7 delves into the analysis of the most common evaluation metrics and the kind of results presented by the selected papers. Section 8 proposes an in-depth discussion about the findings and provides some research guidelines for scholars wishing to contribute to this field. Section 9 wraps up the main achievements of this paper.

2. Research methodology and overview

This section begins by describing the process of gathering the set of works that constitute the main references of this paper to provide trends and challenges of CT in robotics and autonomous agents. The selection process was performed by referring to the PRISMA 2020 guidelines (Page et al., 2021). High-level observations regarding the selection are then reported, followed by a more in-depth analysis in the subsequent sections.

2.1. Selection criteria

The term "trust" has multiple meanings according to the context in which it is used. Besides, many publications tend

to strategically employ this term as a buzzword (Morgner, 2013). When searching for works in which trust is a key element of the query, a database is very likely to generate thousands of results, most of which are not pertinent to the actual focus of the investigation. Therefore, at the beginning of this work, several brainstorming sessions took place to identify a set of search hyper-parameters best suited for the purpose. This choice aimed to initially filter out unrelated works while preserving those related to the concept of CT in robotics.

After having identified the initial set of papers in January 2023, the authors agreed on a selection process, described below. At each stage of the selection process, every author individually reviewed each paper in the current selection, applying the criteria agreed upon for that stage and voting on whether to remove it or not. If a paper was unanimously voted for removal by all reviewers, it was immediately discarded. Otherwise, the authors gathered to discuss the specific contested entries until a unanimous decision was reached to either remove or retain them for the next phase. This search and selection procedure was repeated in December 2023 to include works released during that year. For simplicity, the selection process is reported with these two instances merged together. EndNote™20 (Clarivate™, Philadelphia, USA) was utilized to organize the outcomes of every stage of the selection.

The following set of keywords was used:

"*trust**"

AND

"*human*"

AND

("*robot**" OR "*autonomous vehicle**" OR "*intelligent agent**" OR "*cognitive agent**" OR "*virtual agent**" OR "*autonomous agent**")

AND

("*architecture*" OR "*model**")

This combination was searched in the title, abstract, or keywords of works available on IEEE Xplore® (IEEE®, New York City, USA), ISI Web of Science™ (Clarivate™, Philadelphia, USA) and Scopus® (Elsevier, Netherlands). From the results of these queries, works written in English and published from 2013 to 2023, inclusive, were retained. This initial set was composed of 1,916 papers: 526 from IEEE Xplore, 447 from ISI Web of Science, and 943 from Scopus. After merging the results of the three databases, duplicates, reviews, and early access papers were removed, reducing the set to 1,429 items.

At this point, the papers underwent a first screening phase. The works were skimmed by reading their titles and abstracts. They were included if they were accessible conference proceedings or journal papers, with an experimental validation involving trust with the presence of an artificial agent. They were instead discarded if trust was treated as a component of a bigger phenomenon, for example, acceptability (Semeraro et al., 2024a), or they were related to the fields of social networks, cybersecurity, trustworthy systems or explainable AI (XAI). The exclusion of the fields of social networks and cybersecurity is self-explanatory, as they fall outside the scope of robotics research. The rationale for excluding works on trustworthy systems and XAI is less immediate but still consistent with the focus of our investigation. Our review centers on the mathematical modeling of the trust variable, from both natural and artificial perspectives (see Section 1). Research on trustworthy systems primarily examines the factors (Hancock et al., 2021) that lead to the establishment of human trust in a system, which is not the objective of this review. Instead, we aim to explore studies that propose a priori mathematical models of trust and validate them through real-world deployment. For the same reason, XAI was excluded, as it primarily addresses how experts and end users perceive model explanations (Xu et al., 2019), rather than offering insights into the design of trust models themselves. This screening shrunk the set to 438 items.

A second screening then took place, in which the remaining set of papers was inspected with further depth. Papers were included in the set of studies to analyze if they provided a quantitative, mathematical CT model assessed through experimental validation. If multiple papers by the same authors presented methods and validations that were too similar, indicating a lack of novel contribution from one to another, only the most recent paper was retained and the other ones were discarded. Additionally, cases in which the CT model was obtained solely through regression analysis of trust measurements collected from a previous user study were excluded. This choice was again based on the fact that, in these works, the models were not conceived prior to interaction with another agent and therefore did not assist a robotic system in making more informed decisions. In contrast, if a model was developed following an initial user study and then validated in a second study using the newly derived trust model, the work was included in the final pool. As a result of this second screening, 101 papers were selected to constitute the final set upon which this review bases its discussions. The whole selection process is summarized in Figure 2.

2.2. Result overview

The main features of the selected set of papers are reported in Table 1. The next sections delve into thorough analysis of each of the reported features, as well as a cross-referencing of the most relevant of them. As a broad introspection of the results, Figure 3 reports the diagram of the co-occurrences

of the keywords of the final set over time (Semeraro et al., 2024b). First, it is possible to appreciate that the main keywords were "*trust*" and "*human–robot interaction*," which indicate that the selection process was able to gather contributions to the main fields of interest of this work. Interestingly, there is a consistent link between "*trust*" and "*human–robot collaboration*," more recent than the previous one. This shows that researchers are being inclined to consider trust differently from a merit parameter to assess the acceptability of a robotic design. Rather, they are moving towards considering the trust dynamics in the specific case of human–robot collaboration, in which users and robots share a goal and try to accomplish it together. This is also consequence of having looked for an experimental validation as selection criterion.

Please note that, despite the clear focus found on human–robot collaboration, we did not want to limit our work to well-defined embodiments. Hence our inclusion of search keywords such as "*intelligent agent**" or "*autonomous agent**." CT can be produced and studied for non-embodied agents, for example in simulation, with the aim to develop generalizable models to be then embedded in different robotics platforms.

A result of interest is the upper branch of the diagram. It contains the very recent keyword "*robot trust*," which is closely related to CT. This term can be used to indicate research works that embed trust models into the robot's behavioral model. In these works, trust is used as a variable that can affect the outcome of the interaction during the deployment of the robotic solution. The influence of trust on the robot's behavior can also be seen by its connections to "*decision making*" and "*cognitive model*," which indicate systems capable of gathering information from the environment and extrapolating high-level information, which is then used to tune the behavior of the robot according to the user's behavior. This is evidence of the increasing importance given to incorporating trust dynamics in the behavior of robots, which justifies our attempt at providing standards in the design of trust models for robotics applications.

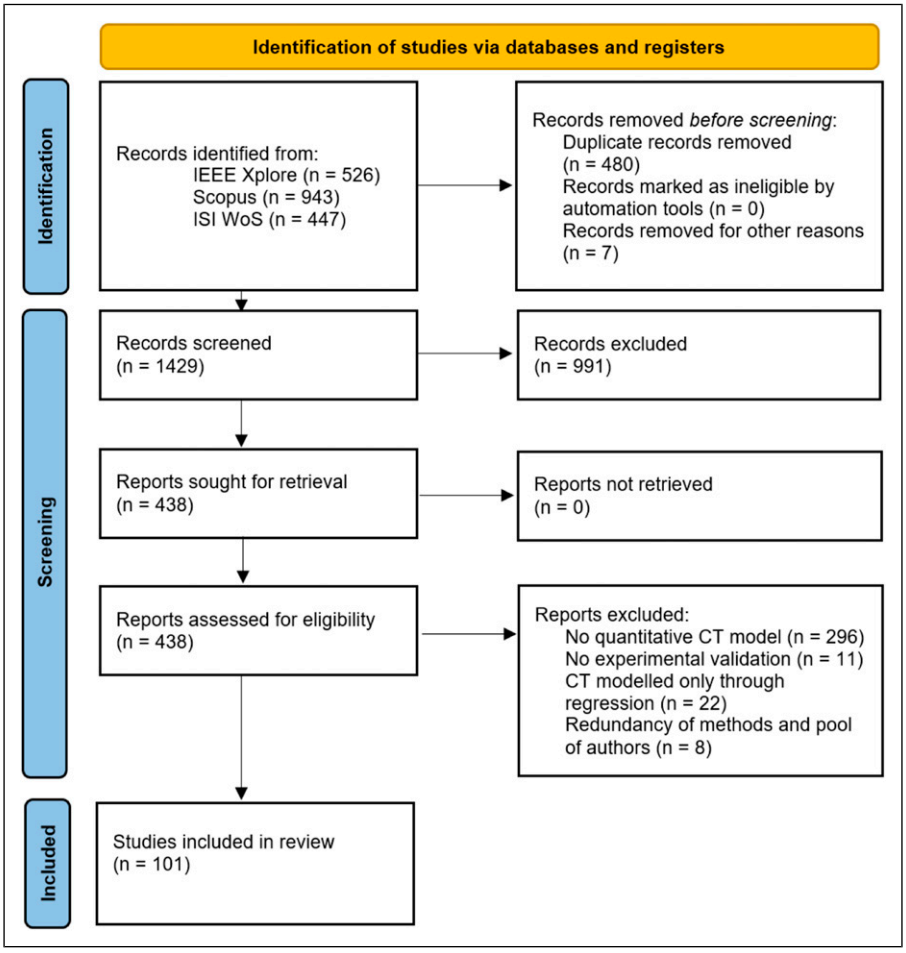

Figure 2. Selection process of the systematic review, according to the PRISMA guidelines (Page et al., 2021).

Table 1. Features extracted from the final set of papers. For the "CT" column, the legend is: AT = Artificial Trust, NT = Natural Trust, BT = Bilateral trust model (see Subsection 8.2). For the "Robot" column, the legend is: A = Humanoid, B = Unmanned Aerial Vehicle (UAV)/Drone, C = Autonomous Vehicles, D = Mobile robot, E = Manipulator, F = Autonomous Agents. The presence of a "," means that the two agents were deployed separately in different experiments, while an "and" shows that they were present in the same experiment. For the "Trust model" column, A = Deterministic, B = Probabilistic, C = Machine learning, D = Game theory. For the "Domain" column, A = Human–robot collaboration (HRC), B = Reconnaissance, C = Navigation, D = Multi-agent systems, E = Autonomous vehicles, F = Human–robot interaction (HRI), G = Telerobotics, H = Miscellanea. For the "Experiment" column, A = Computer simulation, B = Real-world setting, C = Online. Regarding other abbreviations used in the table, ANN = Artificial Neural Network, ANOVA = ANalysis Of VAriance, HMM = Hidden Markov Model, MDP = Markov Decision Process, POMDP = Partially Observable Markov Decision Process, RF = Random Forest, RL = Reinforcement Learning, SVM = Support Vector Machine, UGV = Unmanned Grounded Vehicle.

Ref.	CT	Robot	Trust model		Domain		Experiment		Evaluation methods
Ab Aziz et al. (2017)	NT	F	Linear combination of perception, reliability, transparency, deception, competency, trust and distrust	A	Robot-based therapy	H	To test different trustee characteristics to evaluate effect on trustor's trust levels	A	Mathematical analysis and logical verification
Abdulhussain and Aziz (2022)	NT	F	Network-oriented, with use of differential equations to describe causal and temporal links in the model	A	Military	F	One robot and one human are negotiating with a terrorist to convince them to disarm a bomb and surrender	A	Mathematical analysis and verification through equilibria and temporal trace analysis
Akash et al. (2019)	NT	F	POMDP	B	Reconnaissance	B	Humans have to decide whether a building is safe following indications received from a robot	C	ANOVA on 5 different metrics using pairwise t-test to analyze the differences
Alhaji et al. (2021)	AT	F	Non-linear combination of predictability in MDP	A	Close-contact collaboration	C	To navigate towards the target, avoiding collisions with the human by estimating their predictability	A	Analysis based on number of steps and number of collisions
Ali et al. (2022)	AT	F and F	Probabilistic model that takes into account agent capability and task requirements	B	Task allocation in HRC	A	The model, based on its beliefs, assigns a tasks to either the human or the robotic agent, both present in the same setting	A	Team performance, individual performance and team total reward; comparison with other models
Almohamade et al. (2021)	AT	E	Piecewise definition of trust score influenced by probability of detecting the expected user	A	Manufacturing	A	To jointly navigate through a maze	B	Accuracy, precision, recall, F measure, average number of genuine actions and average number of imposter actions
Aydoğan et al. (2015)	AT	F	Differential equations in a belief network, depending on confidence and evaluated information variables	A	Logistics	G	To handle an airline decision process	A	Trust values of information source
Aziz and Abdulhussain (2022)	NT	F	Linear combination of physical embodiments, social and reliable behaviors and perceived controllability	A	Robot-based therapy	H	To test different trustee characteristics to evaluate affect on trustor's trust levels	A	Mathematical analysis and logical verification
Bhat et al. (2022)	NT	D	Beta distribution in MDP	B	Reconnaissance	B	To clear an area from threats	C	Trust score and workload questionnaires

(continued)

Table 1. (continued)

Ref.	CT	Robot	Trust model		Domain		Experiment		Evaluation methods
Boer et al. (2013)	NT	F	Linear combination of previous trust and experience	A	Reconnaissance	B	To understand which source of information is reliable in a quiz game	C	Task performance and repeated measures ANOVA
Carbo and Molina (2023)	NT	F	Piecewise increases of trust score	A	Logistics	C	In the GAMA platform, the agents have to carry different objects from their locations to other ones; objects and places have different privacy issues	A	Average reward; comparison with model w/o emotions
Carneiro et al. (2019)	AT	F, F	Non-linear combination of positive and negative interactions	A	Games	H	To play the game "Settlers of Catan," against humans or other agents, in separate experiments	A, C	Win rate and questionnaires; comparison with baseline
Chen et al. (2018)	NT	E	Pre-trained Gaussian variable in a POMDP	B	HRC	A	To clear objects off a table in a turn-taking fashion	B	Trust score and intervention rate; comparison with baseline
Chen et al. (2020)	NT	F	Pre-trained Gaussian variable in a POMDP	B	HRC	A	To clear objects off a table in a turn-taking fashion	A	Mean accumulated reward
Cheng et al. (2021)	AT	F and F	Piecewise combinations of tuples of belief, disbelief, uncertainty and base rate	A	Driver management	E	To cross intersections regulated by the intersection management system	A	Collision rate and throughput; comparison with case without trust
Cheng et al. (2023)	NT	C	Linear combination of questionnaires responses	A	User profiling	E	Users were surveyed regarding their trust towards autonomous vehicles under different driving scenarios	A	Mean absolute error
Dorbala et al. (2021)	AT	F, D	2D Gaussian distribution with mean on the task goal	B	Human-guided robot navigation	C	The robot, unembodied in one setting and embodied in the other one, finds the optimal path towards a goal location collaborating with the human	A, B	Reward per episode and robot confidence; comparison with random strategy
Dubey and Kumar (2019)	AT	D, D	Non-linear combination of capability and intention variables	A	Robot navigation	C	The robot plans a path towards a target in an efficient way in collaboration with another robot	B	Number of best, fair and unsuccessful plans achieved in the interaction cycles
Floyd et al. (2014)	NT	F, F	Linear combination of trust value of a command in case-based reasoning	A	Reconnaissance	B	To navigate from an area to the next one/to find possible threats in an area	A	Number of cases evaluated before reaching a trustworthy behavior
Floyd et al. (2015)	NT	F	Linear combination of trust value of a command in case-based reasoning	A	Reconnaissance	B	To find possible threats in an area	A	Number of cases evaluated before reaching a trustworthy behavior
Guo and Yang (2021)	NT	F	Beta distribution depending on tasks succeeded and failed, with maximum likelihood	B	Reconnaissance	B	To search for potential threats in an extended area	C	Self-reported trust value; comparison with baseline

(continued)

Table 1. (continued)

Ref.	CT	Robot	Trust model		Domain		Experiment		Evaluation methods
Guo et al. (2023a)	NT	F	Beta distribution depending on tasks succeeded and failed, in a MDP	B	Reconnaissance	B	In a simulated environment, a human and a robot have to intervene in a dangerous area	A	Value function
Guo et al. (2021)	NT	F	Beta distribution in a POMDP	B	Reconnaissance	B	To search for potential threats in an extended area	A	Cumulative reward during an episode
Guo et al. (2023b)	NT	B	Beta distribution with maximum likelihood	B	Reconnaissance	B	In a simulated environment, teams of two humans, one trustor and one trustee, reconnoiter the area with two drones	A	RMSE; comparison with baseline
Hale et al. (2019)	NT	F	Linear combination of previous trust value and gradient of a cost function of the knowledge of the human user	A	HRI	F	To analyze stability of the trust framework	A	Trust and privacy values
Hannum et al. (2020)	AT	E	Non-linear combination of previous trust value and current intention of the user	A	HRC	A	To co-carry a wheeled cart from one place to another	B	Trust value
Hoogendoorn et al. (2014)	NT	B	Linear combination of previous trust values and current experience	A	Reconnaissance	B	To detect, in collaboration with another human trustee, threats on a wide area visualized through generated images of footages of unmanned aerial vehicles	C	Prediction accuracy; comparison between absolute and relative trust model
Hsieh et al. (2022)	NT	C	ANN, SVM, RF	C	Autonomous vehicles	E	To collect trust measurements from user to use for later training and validation of the trust model	B	Classification accuracy; comparison between model that use different sources of information
Hu et al. (2022)	NT	C	Linear combination of piecewise variable depending on error between desired and actual vehicle driving behavior	A	Autonomous vehicles	E	To drive in four different scenarios	A	Trust value, trust tracking error
Hu and Wang (2022)	NT	F	Piecewise non-linear combination of trust of knowledge	A	HRI	F	Teams perform tasks with different starting trust levels	A	Trust value; comparison with model without trust
Kang (2018)	AT	F	Non-linear combination of ratings	A	Recommendation systems	H	To train the trust systems with rating datasets available in literature	A	Mean absolute error; comparison with baseline
Khattar and Eskandarian (2022)	AT	F	Non-linear combination of accuracy of predicted state in HMM	B	Driver management	E	To train the trust systems with rating available from the UAH-DriveSet dataset	A	Trust ratio
Kirtay et al. (2023)	AT	A and A	Q matrix	C	HRI	F	To recall visual patterns previously experienced	B	Accumulated cumulative reward; comparison with model without trust

(continued)

Table 1. (continued)

Ref.	CT	Robot	Trust model		Domain		Experiment		Evaluation methods
Kraus et al. (2021)	AT	A	SVM, boosting, gated recurrent unit	C	Proactive dialogue corpus	H	To train the trust system with a corpus of dialogues	A	F1, unweighted average recall, Cohen's k and Spearman's correlation coefficient
Kumar and Dubey (2017)	AT	F and E	Linear combination	A	HRC	A	A robotic arm has to perform path planning to reach an objective position, avoiding obstacles	A	Trust value
Lang et al. (2023a)	AT	E	Non-linear combination of the probabilities of the hidden variables of a HMM	B	HRC	A	To accomplish different actions in a collaborative way with the robot	B	Custom metrics on intention recognition; comparison with baseline
Lang et al. (2023b)	AT	E	Collaborative filtering	B	Intention reading	F	A human acts in front of the agent, which has to infer their intentions	B	Performance comparison with similar models
Lee et al. (2013)	NT	F	SVM trained from videos annotated with trust-related social signals, supported by a HMM	B, C	HRI	F	Humans first familiarize with a new partner, then play a cooperative game with them	B	Mean prediction error
Lee et al. (2021)	AT	F	Fuzzy inference system	B	HRI	F	Simulation on the model's response with different sets of inputs	A	Evaluation of the outputs from specified input combinations
Li et al. (2023)	AT	F	Ensemble of machine learning models using audio and textual features	C	HRI	F	Decision-making tasks about a space station management	B	RMSE and adjusted $R\hat{2}$
Lin et al. (2023)	AT	F and F	Beta reputation system to calculate agent state, impact of decision, and agent capability; Q-learning to fuse these values	A, C	Multi-agent teaming	D	To collect balls, cooperating with other agents (which might be faulty or functional)	A	Comparison to baseline
Losey and Sadigh (2019)	NT	E	MDP	B	HRC	A	To balance an inverted pendulum collaboratively	B	Paired t-tests
Ma et al. (2022)	AT	D	Fusion of performance metrics through RL	C	HRC	A	A team of a human and a robot compete against another robot in a ball collection task within an environment with obstacles	A	Comparison with a baseline
Mahani and Wang (2017)	AT	D	Linear combination of variables including robot performance and human performance	A	Robot navigation	C	A human and a robot each have to design a path to reach a goal; based on the level of trust, the robot can decide to follow either of them	A	Empirical evaluation
Mahani et al. (2021)	NT	B	Dynamic Bayesian network	B	Reconnaissance	B	A human operator has to supervise a team of UAVs in a search and rescue operation	A	Prediction accuracy
Maithani et al. (2019)	AT	E	Linear combination based on the users performance	A	HRC	A	Trajectory tracking	B	Task completion time, average force applied

(continued)

Table 1. (continued)

Ref.	CT	Robot	Trust model		Domain		Experiment		Evaluation methods
Mangalindan et al. (2023)	NT	E	POMDP	B	Human-supervised object collection	A, F	Object collection task; the human can decide to intervene or to let the robot operate autonomously	A	Comparison to baseline
Mansoor et al. (2013)	NT	F	Linear combination tuned from a user study	A	Autonomous vehicles	E	Autonomous vehicle guided by 3 navigation systems	A	Model accuracy
Nam et al. (2017)	NT	B	MDP and inverse RL	B, C	Swarm robotics	D	Swarm locates targets on a map	A	RMSE, comparative analysis
Nam et al. (2020)	NT	B	MDP and inverse RL	B, C	Swarm robotics	D	Swarm locates targets and receives feedback from human	A	RMSE, comparative analysis
Pang et al. (2021)	NT	B	Deep neural network	C	Surveillance	B	UAVs need to reach target destinations in urban environment	A	Ablation study
Patacchiola and Cangelosi (2016)	AT	F	Bayesian network	B	HRI	F	Sticker finding game	A	Comparison with psychology experiment
Patacchiola and Cangelosi (2022)	AT	A	Cognitive architecture with Bayesian networks and SOMs	B, C	HRI	F	Object name learning	B	Comparison with psychology experiment
Ponnambalam et al. (2021)	AT	D and D	TD learning + beta reputation system	B	Multi-agent systems	D	Robots collaborate to identify red flags	B	RMSD analysis and performance metrics
Rabby et al. (2020)	NT	E	Linear combination of performance-related factors	A	HRC	A	Conveyor belt object sorting task	A	Scenario analysis
Rahman (2019b)	BT	A	Linear combination of performance-related factors	A	HRC	A	Object search via mixed-initiative role negotiation	B	Trust values and questionnaires
Rahman (2019c)	NT	F	Linear combination of performance-related factors	A	HRC	A	Power assistance in lifting task	B	Evaluation of displacement, velocity, questionnaires
Rahman et al. (2016a)	BT	E	Linear combination of performance-related factors	A	HRC	A	Collaborative assembly of LEGO bricks	B	Quantitative performance comparison with and without the trust-aware subtask allocation; user study questionnaires
Rahman et al. (2016b)	AT	E	Linear combination of performance-related factors	A	HRC	A	Collaborative assembly of industrial components	B	Comparison of performance with and without trust considerations, questionnaire
Rahman (2019a)	BT	E	Linear combination of performance-related factors	A	HRC	A	Collaborative assembly of LEGO bricks	B	Comparison of performance with and without trust considerations, questionnaire

(continued)

Table 1. (continued)

Ref.	CT	Robot	Trust model		Domain		Experiment		Evaluation methods
Ranasinghe et al. (2015)	AT	E	Virtual inertial damped stiffness system	A	Robotic guiding systems	A	A robot would guide a human using a hard rein in several different paths, characterized by the steepness of their turns	B	User questionnaires to assess trust levels, compared with virtual damping, mass and stiffness coefficients to identify trusting patterns
Razin and Feigh (2021)	NT	F	Game theory strategies	D	Games	H	A collection of trust-based tasks	A	Mean squared error for trust and trust fulfillment, performance measures
Rishwaraj et al. (2017)	AT	F and F	Beta reputation system and RL	B, C	Multi-agent systems	D	A swarm of robots has to locate some targets on the map; each of them can communicate information on their findings to the others, but some of them have sensor malfunctions	A	Comparison with baseline
Rishwaraj and Ponnambalam (2017)	AT	D and D	MDP and RL	B, C	Multi-agent systems	D	A swarm of robots has to locate some areas on the map; each of them can communicate information on their findings to the others, but some of them have sensor malfunctions	A, B	Average time steps to achieve the objectives; comparison with the literature
Rjoub et al. (2023)	AT	F and C	Interpretable model-agnostic explanations (LIME)	C	Autonomous vehicle evaluation	E	Vehicles are evaluated by LIME to estimate trustworthiness; deep Q-learning is used to select suitable candidates	A	Comparison to baselines
Rutard et al. (2020)	AT	F	Exponentiated negative Bhattacharyya distance between actor and teacher state-action value distributions	A	Navigation	C	Maze solving	A	Performance comparison between agents trained with teachers possessing decreasing levels of ability
Sadrfaridpour et al. (2016)	NT	E	Time-series model as a function of prior trust, change of robot performance, change of human performance, and fault occurrence	A	HRC	A	Joint assembly line	A, B	Performance comparison between only manual allocations, only autonomous and collaborative
Sadrfaridpour and Wang (2018)	NT	E	Time-series model as a function of prior trust, change of robot performance, change of human performance, and fault occurrence	A	HRC	A	Joint assembly line	B	Performance comparison of 4 control models and user questionnaires

(continued)

Table 1. (continued)

Ref.	CT	Robot	Trust model		Domain		Experiment		Evaluation methods
Sadrfaridpour et al. (2018)	NT	E	History-based probabilistic Bayesian network	B	HRC	A	Simulation of human-robot jointly carry an object from a starting position to a goal position	A	Case-study evaluation
Saeidi et al. (2017b)	NT	B	Linear combination of performance variables	A	Telerobotics	G	To control a UAV to track the path of an UGV	B	Performance comparison of different control strategies
Saeidi and Wang (2015)	NT	D	Quantitative model as a function of the difference in human-robot trust and self-confidence	A	Tele-autonomous operations	G	To keep moving the robot as close to the desired trajectory as possible between the guide posts	A, B	Performance metrics evaluation
Saeidi et al. (2016)	BT	B	Linear combination of robot and human performance	A	Telerobotics	G	To control a UAV to track the path of a UGV	A	Passivity analysis and users questionnaires
Saeidi et al. (2017a)	BT	B	Linear combination of robot and human performance	A	Multi-robots teleoperation	D, G	To control formation and motion of robots while visiting checkpoints	A	Performance metrics evaluation
Saeidi and Wang (2019)	BT	B	Linear combination of robot, human performance and human self-confidence	A	Telerobotics	G	To control a UAV to track the path of a UGV	A	Performance metrics evaluation and users questionnaire
Sapienza and Falcone (2023)	AT	F	Linear combination of performance variables	A	Smart devices and Internet of Things	H	Each device has to determine how to execute its task, based on the user's trust in it	A	Task performance in different operational settings
Scherf et al. (2022)	AT	E	Conservative linear combination of factors linked to the history of human performance and behavioral cues	A	HRC	A	Gridworld task with simulated human users/Sorting objects into boxes based on their weights	A, B	Comparison against baselines and analysis of how task-dependent uncertainties influence response time and behavioral cues of participants
Setter et al. (2017)	AT	F and F	Gradient descent framework	C	Multi-agents systems/ swarm robotics	D	Swarm aggregation task	A	Study on model convergence
Soh et al. (2020)	NT	C, D	Task-dependent latent dynamic function	A	HRC	A	Autonomous fetching/Autonomous driving	A	Average negative log-likelihood and mean absolute error
Spencer et al. (2016)	NT	D	Time-series model as a function of robot performance, human performance and faults made in the joint human–robot system	A	Multi-agent systems	D	The robot team has to successfully reach different goals in the grid environment avoiding 16 obstacles and each other along the way	A	Matlab model checking using NuSMV
Sun et al. (2023)	NT	C	Linear combination of the drivers' personal and contextual attributes, fuzzy logic	A, B	Autonomous driving in mixed traffic flow	E	Online questionnaire on drivers behaviors	C	Statistical evaluation on questionnaires' answers
Tilloo et al. (2022)	AT	E	POMDP	B	HRC	A	Handover task	B	Qualitative analysis of the behavior of the robot

(continued)

Table 1. (continued)

Ref.	CT	Robot	Trust model		Domain		Experiment		Evaluation methods
Tjøstheim et al. (2019)	AT	F	Brain inspired architecture	A	HRI	F	Training and testing of a model simulating trust building based on familiarity in facial features and positive or negative interactions	A	Evaluation of how trust changes based on facial features and gentle touch
Vinanzi et al. (2019)	AT	A	Bayesian networks	B	HRC	A	The robot has to identify the correct location of pre-defined markers with the help of the human	B	Test of the ability of the system to recognize trickster using a Theory of Mind model
Vinanzi et al. (2021)	AT	E	Bayesian networks	B	HRC	A	Human and robot cooperate to use blocks to form a pre-defined line of colored blocks	A	Success rate, artificial opinion
Wagner et al. (2018)	NT	D,A	Framework using game theory's computational representations	D	Miscellanea	H	Participants read about 12 different scenarios, including the investment scenario/simulated maze game/investor-trustee game with NAO	B,C	Multiple human subject assessments
Wang et al. (2023b)	NT	C	Numerical value between [0,1] representing trust in the connected vehicles information	A	Autonomous driving in mixed traffic flow	E	Numerical simulations where trust follows a normal distribution with a mean of 0.5	A	Correlation and sensitivity analysis
Wang et al. (2014)	BT	D	Time-series model that takes into account robot performance, human performance and overall system fault rates	A	Underwater robotics	B, G	Station keeping task	A	Performance investigation on the outcomes of the task
Wang et al. (2015)	BT	F	Time-series model taking into account agent performance and agent fault rate	A	Multi-agent systems	D	Simulation with pre-defined parameter values	A	Numeric evaluation
Wang et al. (2018)	NT	D	Dynamic Bayesian network	B	Multi-agent systems	D	Multiple robots need to reach a set of destinations while avoiding obstacles and collisions, potentially taking shorter but riskier paths with human oversight	A	Trust evaluation from human inputs; performance comparison between manual autonomous and switching control strategies
Wang et al. (2022)	AT	E	Linear combination of co-workers performance factors including safety, robot singularity, smoothness, physical performance and cognitive performance	A	HRC	A	Trajectory tracking	B	Parameter investigation of the proposed linear combination

(continued)

Table 1. (continued)

Ref.	CT	Robot	Trust model		Domain		Experiment		Evaluation methods
Wang et al. (2023a)	AT	E	Linear combination of objective robot and human performance measures including safety, robot singularity, smoothness, physical performance and cognitive performance	A	HRC	A	Moving target tracking	B	Metric-based evaluation plus NASA-TLX
Wu et al. (2017)	NT	E	MDP	B	HRC	A	Assembly process involving a robot and human working on a shared workspace	A	Case-study evaluation of different allocation strategies
Xu and Dudek (2015b)	NT	B	Dynamic Bayesian network	B	HRC	A	Patrolling task of a simulated environment	A	Performance comparison with other models in the literature
Xu and Howard (2020)	NT	A	HMM and logistic regression	B	HRC	A	Counting toothpicks, self-driving car	C	Performance comparison against 2 baselines
Xu and Song (2021)	BT	B	Time-series model based on past agent performance, fault rate and past trust towards it	A	Multi-agent systems/ swarm robotics	D	Simulated map surveillance mission	A	Accomplishment rate, human performance, mission performance
Yan et al. (2022)	NT	F	Time-series model based on surgeons' operation performance, robot performance and the number of undesired operations	A	Robot-assisted vascular interventional surgery	H	To insert a guidewire into a bifurcation	B	Accuracy and success rate
Zahedi et al. (2023)	NT	D	Numerical value between [0,1]	A	Navigation	C	Ablation study on the framework/ Gridworld scenario simulating a Mars Rover navigation mission	A, B	Statistical analysis on the results; comparison against baselines
Zheng et al. (2023b)	NT	E	Weighted combination of human's and robot's planned actions, adapted based on the deviation between the robot's estimated human actions and the real-time actual human inputs	A	HRC	A	Collaborative transportation task in presence of obstacles	B	Comparison to a baseline on different metrics (i.e., number of collisions)
Zheng et al. (2018)	NT	D	Dynamic Bayesian network	B	Multi-agent systems	D	Navigation in a simulated gridworld	A	Evaluation of the paths used by the team of robots
Zheng et al. (2023a)	NT	D	Linear state-space equation, Bayesian inference	A, B	Multi-agent teaming	D	Simulation and user study regarding collaborative bounding overwatch task	A, B	Human self-reported measures of trust every time the autonomous subteam makes a decision, task completion metrics (i.e., number of collisions)
Zhou et al. (2021)	BT	D	Linear combination integrating the status of the operator, performance of the robot, and environment information	A	Reconnaissance	B	Navigation task	B	Comparison of 3 navigation control strategies on the generated paths and their smoothness

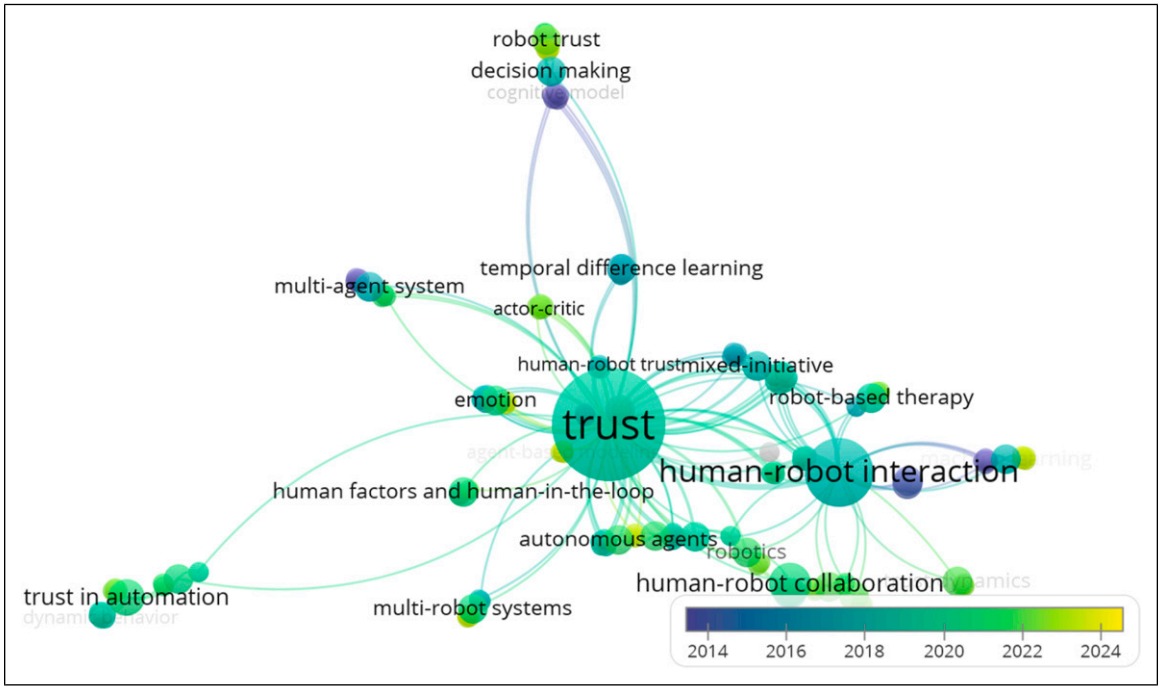

Figure 3. Keyword network of the selected set of papers (generated through VOSviewer (Van Eck and Waltman, 2010)).

However, we argue that it is preferable to use the term CT to refer to this research sector, instead of robot trust, because the latter could be easily misinterpreted as the trust that users place into robots. While such conferred meaning to robot trust is a part of the research, it does not necessarily encompass all the possible ways in which trust dynamics can be modeled. Indeed, trust is a bidirectional relationship (Zonca and Sciutti, 2021); therefore, it is possible to model the trust of the user towards the robot, the trust the robot could have towards the users, or both simultaneously.

3. Domains of application

Understanding the importance of CT requires investigating the application domains in which it is used and how it could contribute to the field. In this section, we outline the different domains of application that could benefit from CT, based on the papers we reviewed. We have identified eight categories to classify these domains: (A) Human–Robot Collaboration; (B) Reconnaissance; (C) Navigation; (D) Multi-Agent Systems; (E) Autonomous Vehicles; (F) Human–Robot Interaction; (G) Telerobotics, plus a Miscellaneous category (H) for works that do not fit into any of the other categories and are not numerous enough to form one on their own. Inevitably, some works could fall into more than one category. In such cases, we chose to affiliate them with the domain in which they have the potential to be most impactful. Figure 4 summarizes our findings, which are discussed in detail in the rest of this Section.

3.1. Human–robot collaboration

Human–Robot Collaboration (HRC) aims to understand what enables humans and robots to efficiently work together to complete a given task. Of the 101 papers in our review, 32 fall within this category, making it the most represented in this work. This high representation is due to the diverse nature of HRC research, encompassing various types of tasks that humans and robots can complete together. For example, one of the main areas linked to HRC is manufacturing (Almohamade et al., 2021). In industrial settings, the use of collaborative robots, or cobots, that physically interact with humans in a shared workspace is being explored for a variety of tasks (Wang et al., 2022; 2023a). Assembly tasks (Rabby et al., 2020; Rahman, 2019a; Rahman et al., 2016a; Sadrfaridpour et al., 2016; Sadrfaridpour and Wang, 2018) and object handover tasks (Rahman et al., 2016b; Tilloo et al., 2022) involve robots and their human partners sharing a static workspace. Examples of human and robot teams completing tasks in a shared static workspace include studies where the team coordinates to clear objects from a table (Chen et al., 2018, 2020). However, human–robot collaboration is not limited to static tasks, where the robot sits statically in one location. For example, in studies like Hannum et al. (2020) Sadrfaridpour et al. (2018) and Zheng et al. (2023b) the human–robot team needs to carry objects from point A to point B while avoiding obstacles in their path. In another study (Rahman, 2019c), they work together to lift an object to the desired vertical position.

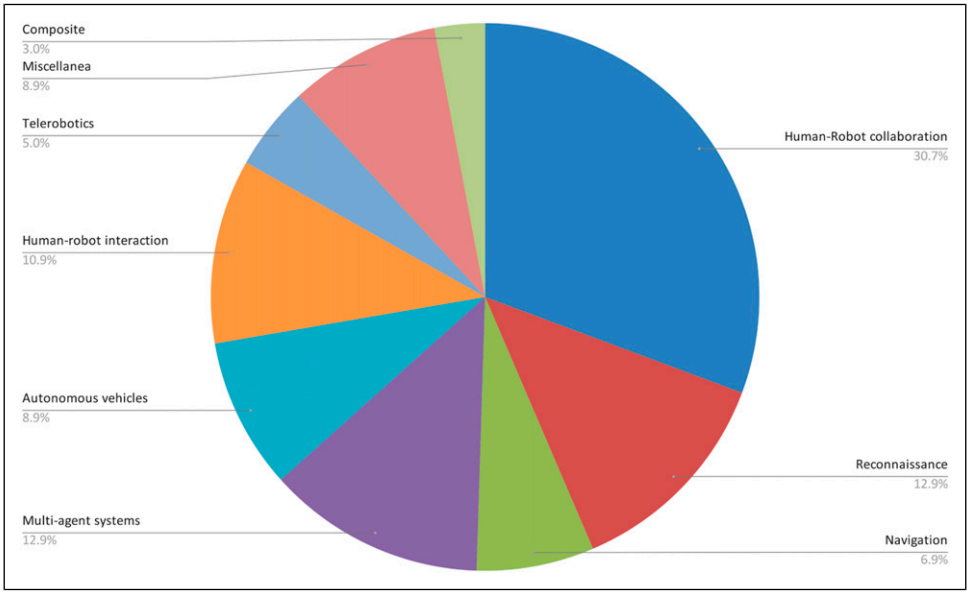

Figure 4. Distribution of the domains of application used by the selected papers. Computational Trust models appears to be used mostly in Human–Robot collaboration scenarios. Papers that span multiple domains have been classified as Composite.

Another common task in the HRC domain is trajectory tracking (Maithani et al., 2019; Xu and Dudek, 2015b). It can involve defining better trajectories for robot joints (Kumar and Dubey, 2017; Maithani et al., 2019) or fulfilling a patrolling task (Xu and Dudek, 2015a). Object collection is another prevalent problem in HRC (Mangalindan et al., 2023; Rahman, 2019b; Ranasinghe et al., 2015; Soh et al., 2020). For instance, in Rahman (2019b), the human and robot search a set of locations to retrieve objects of interest; in Soh et al. (2020), the human monitors the robot's autonomous fetching capabilities; and in Ranasinghe et al. (2015), they work together to successfully retrieve objects from a shelf.

Within collaborative scenarios, task allocation is crucial, and CT could be fundamental in effectively dividing tasks among team members (Ali et al., 2022; Wu et al., 2017) and executing a plan towards a given goal (Scherf et al., 2022). To achieve the final goal, robots need to identify trustworthy and untrustworthy confederates who might potentially hijack the team's mission. CT is used to enhance the robot's identification capabilities in these scenarios (Lang et al., 2023a; Vinanzi et al., 2019, 2021).

Robots could use CT to understand their human partners and either predict their future decisions (Xu and Howard, 2020) or influence them. For instance, Losey and Sadigh (2019) specifically investigate whether a robot can leverage the trust of a human to directly influence them in a collaborative task. Finally, CT can also play an important role in scenarios where a human-robot team competes against another agent. For example, Ma et al. (2022) use a CT model to establish tactical coordination within a team competing against another robot in an object collection task.

Over the past decade, academic interest in CT for HRC has followed a clear upward trajectory. Beginning with just 3 publications in 2013, the number of relevant papers has increased steadily, reaching 19 in 2019. This consistent growth in publication volume is an indication of a sustained and expanding engagement with the topic across the research community, reflecting its growing importance and relevance. The observed pattern suggests that this trend is not due to isolated spikes in attention but reflects a wider and more lasting scholarly focus.

3.2. Reconnaissance

Reconnaissance scenarios, from search and rescue (Mahani et al., 2021; Zhou et al., 2021) to threats detection and clearance (Bhat et al., 2022; Floyd et al., 2014, 2015; Guo et al., 2021; Hoogendoorn et al., 2014), provide many opportunities for research in CT to test the need and efficacy of proposed models. Of the 101 papers included in our review, 14 are framed within the reconnaissance domain. Search and rescue scenarios include coal mine rescue robots (Zhou et al., 2021) and human-multi-robot UAV teams for search missions (Mahani et al., 2021). Reconnaissance missions can also involve the control and maintenance of specific areas (Guo et al., 2023a, 2023b; Wang et al., 2014; Guo and Yang, 2021; Akash et al., 2019) or missions where the goal is to reach predetermined destinations following precise trajectories (Pang et al., 2021). An example of area monitoring is given by Akash et al. (2019), where operators decide whether a building is safe based on information gathered by a robot. Similarly, Boer et al. (2013) use a CT model to estimate if a source of information is reliable.

3.3. Navigation

The category of Navigation includes works that focus on path planning and navigation missions not necessarily related to classic reconnaissance scenarios. Only 7 papers fall into this category. Examples of works discussing the use of CT for more efficient path planning include Alhaji et al. (2021), Dorbala et al. (2021), Dubey and Kumar (2019), and Mahani and Wang (2017), where humans and robots collaborate to find the best path to reach a target location. Carbo and Molina (2023) go a step further, framing their work within the logistics domain and examining agents that have to carry objects with different privacy concerns across two locations. Rutard et al. (2020) consider navigation within a maze-solving task, while Zahedi et al. (2023) incorporate CT in a simulated Mars rover navigation scenario.

3.4. Multi-agent systems

Multi-agent systems (MAS) refer to teams of more than two agents cooperating toward a common goal, where agents can be robots or humans. In this review, 14 papers fall within the MAS field. A common application domain for MAS is swarm robotics (Nam et al., 2017, 2020; Rishwaraj et al., 2017; Rishwaraj and Ponnambalam, 2017; Setter et al., 2017). In Nam et al. (2017, 2020), a swarm of UAVs need to locate a set of targets on a map, guided by a human operator. Similarly, Rishwaraj et al. (2017); Rishwaraj and Ponnambalam (2017) use a CT model for a swarm of mobile robots to locate targets on a map by communicating among themselves and dealing with sensor malfunctions during the task. Setter et al. (2017) use CT to help with the classic problem of swarm aggregation.

Outside the swarm robotics field, we see works on multi-agent teaming overlapping with navigation and surveillance tasks (Ponnambalam et al., 2021; Saeidi et al., 2017a; Spencer et al., 2016; Wang et al., 2018; Xu and Song., 2021; Zheng et al., 2018, 2023a). In particular, Ponnambalam et al. (2021) integrate a CT model into two robots collaborating to identify the positions of flags in an environment. Saeidi et al. (2017a) use CT to control the formation and motion of robots visiting different checkpoints. Spencer et al. (2016) and Wang et al. (2018) apply CT in navigation problems with obstacle avoidance. In Lin et al. (2023), a CT model is used by virtual agents to estimate trust from lexical and acoustic data. Similarly to what some works in HRC have done (see Subsection 3.1), CT in MAS can also help with improving real-time scheduling of tasks (Wang et al., 2015).

3.5. Autonomous vehicles

Autonomous Vehicles are attracting increasing effort from researchers in CT models, as their introduction in public spaces leads to a range of open problems that this domain might help solve. For example, Cheng et al. (2021) explore the possibility of improving an autonomous intersection management system, and Wang et al. (2023b) propose a generalized car-following model for mixed traffic flow. Work is still needed to fully understand how drivers' trust can be predicted and used in autonomous vehicles. For this reason, studies such as Sun et al. (2023) and Cheng et al. (2023) use questionnaires to better understand which features drivers' trust levels depend on. Additionally, works like Khattar and Eskandarian (2022), Hu et al. (2022), and Hsieh et al. (2022) assess how well CT models estimate the driver's trust, even in the face of an abrupt decrease in performance (Hu et al., 2022). Finally, studies such as Mansoor et al. (2013) and Rjoub et al. (2023) try different models to find the best navigation system for autonomous vehicles. Overall, not many works fall into this category. Indeed, only 9 papers are directly related to Autonomous Vehicles.

3.6. Human–robot interaction

In this category, we included works that examine CT in the interaction between humans and robots without being particularly linked to a specific and generalizable application. 12 papers fall into this category. We can distinguish between works that try to estimate human trust in robots from their intentions and internal states (Abdulhussain and Aziz, 2022; Lang et al., 2023b) and works that try to estimate the trustworthiness of the human partner from facial features (Tjøstheim et al., 2019), behaviors (Kirtay et al., 2023; Patacchiola and Cangelosi, 2016), or conversational inputs (Li et al., 2023).

Another important aspect of trust and trust-building is the history of interaction between the human and the robot. For this reason, Lee et al. (2021) investigate human antecedents and familiarity as clues for CT, while Lee et al. (2013) explore whether a robot can successfully identify a trustworthy human it had not previously engaged with. The notion of history is also used to discuss and compare trust-based dynamics in teams carrying out tasks starting from different levels of trust (Hu and Wang, 2022).

Two works look at trust from a more theoretical perspective: Hale et al. (2019) attempt to formalize a mathematical model for trust-driven levels of privacy, while Patacchiola and Cangelosi (2022) focus on developing a biologically inspired CT model.

3.7. Telerobotics

Telerobotics concerns itself not with autonomous robots but with platforms that are controlled by a human operator. Out of the total 101 papers, 7 refer to teleoperated robots. In particular, works in this area focus on creating more efficient control strategies that yield or retain autonomy of the

operated robot (Saeidi et al., 2016, 2017b; Saeidi and Wang, 2015, 2019; Wang et al., 2014). Specifically, Saeidi et al. (2016); Saeidi and Wang (2019) examine mixed-initiative control scaling and autonomy allocation strategies for a group of UAVs tracking the path of a mobile robot. A potential domain where telerobotics could improve processes is logistics. For instance, Aydoğan et al. (2015) explore how a CT model could enhance an airline decision-making process.

3.8. Miscellanea

This category includes all those application domains that were not represented enough to form a category of their own. These include robot-assisted therapy (Ab Aziz et al., 2017; Aziz and Abdulhussain, 2022), gaming (Carneiro et al., 2019; Razin and Feigh, 2021), recommender systems (Kang, 2018), smart devices and the Internet of Things (Sapienza and Falcone, 2023), robot-assisted vascular surgery (Yan et al., 2022), works on conceptual frameworks from Game Theory (Wagner et al., 2018), and CT-oriented dialogue systems (Kraus et al., 2021). A total of 9 papers fall into the Miscellanea category.

3.9. Insights

Although the domains in which CT contributes to shaping human-agent interactions are undoubtedly diverse, we can confidently identify the motif behind the decision to develop and study CT: collaboration. It is not only in line with what we found in this review, namely that the majority of the papers included fit within the HRC domain, but also with the definition of trust given by Mayer et al. (1995) which hints to the existence of a task to be completed for trust to even manifest. Whether it is in HRC scenarios involving cobots and manufacturing, surveillance or telepresence, CT is clearly central when a human-agent team needs to solve a shared problem. We have seen that such problems can span from identifying positions of targets in an environment (Ponnambalam et al., 2021) to object collection (Rahman, 2019b) and object handover tasks (Rahman et al., 2016b).

From the analysis of the domains, it is clear that CT is fundamental because it helps both the agents and the humans involved in the collaboration to understand whether their partner is trustworthy or not (not necessarily due to bad intentions) and accordingly make decisions to successfully complete the task they are assigned. To this end, we have found examples of works looking at reliability of information (Boer et al., 2013), at human characteristics that could identify a trustworthy or untrustworthy partner (Tjøstheim et al., 2019), and looking at agents characteristics that could be deemed trustworthy or not by their human partners (Sun et al., 2023). Ultimately, a way in which CT could be successfully employed is by using these results to develop more efficient control strategies that can rely on CT to decide whether to yield or retain the autonomy of the agents (Saeidi et al., 2017b).

Therefore, the important message that this scrutiny points to is that CT is mostly situated in those domains where there is a common goal to be reached and where the human-agent team need to find the best strategy to do so. Future work directions see CT more and more integrated in the decision-making modules of autonomous agents.

4. Robots

When discussing CT, it is important to analyze the different robotic platforms used over the years and across application domains to identify trends in the current literature. For this reason, we provide an overview of the types of robots used in the papers we have analyzed. We have identified the following categories: (A) Humanoids; (B) Aerial Robots; (C) Autonomous Vehicles; (D) Mobile Robots; (E) Manipulators; (F) Autonomous Agents. Some papers make use of different types of robots and are therefore included in multiple categories. The discussion highlights whether the robot in question acts as a trustor or a trustee. Figure 5 summarizes the distribution across these categories for both cases.

4.1. Humanoids

While there is no formal definition for humanoid robots, they are generally designed to resemble and mimic human form and behavior, often featuring a head, torso, arms, and legs. Their human-like appearance and capabilities make them particularly valuable for tasks that require social interaction and collaboration with humans, and for this reason they are mainly used for applications that span across HRI and HRC domains. Examples of humanoid robots are Pepper and Nao from Aldebaran, formerly known as SoftBank Robotics (Kirtay et al., 2023).

Humanoids have been used as trustors in 5 papers (Kirtay et al., 2023; Kraus et al., 2021; Patacchiola and Cangelosi, 2022; Rahman, 2019b; Vinanzi et al., 2019) spanning from 2019 (Rahman, 2019b; Vinanzi et al., 2019) to 2023 (Kirtay et al., 2023). Of this total, 2 fall into the HRC domain (Rahman, 2019b; Vinanzi et al., 2019), other 2 in the HRI domain (Kirtay et al., 2023; Patacchiola and Cangelosi, 2022) and one in the Miscellanea category (Kraus et al., 2021). Humanoids have been used as trustees in 4 papers (Kirtay et al., 2023; Rahman, 2019b; Wagner et al., 2018; Xu and Howard, 2020), the first of which dates to 2018 (Wagner et al., 2018). Also in this case, application domains are HRC (Rahman, 2019b; Xu and Howard, 2020), HRI (Kirtay et al., 2023), and Miscellanea (Wagner et al., 2018).

It is interesting to note that in 4 out of the 5 papers where humanoids take on the role of trustors, the authors present

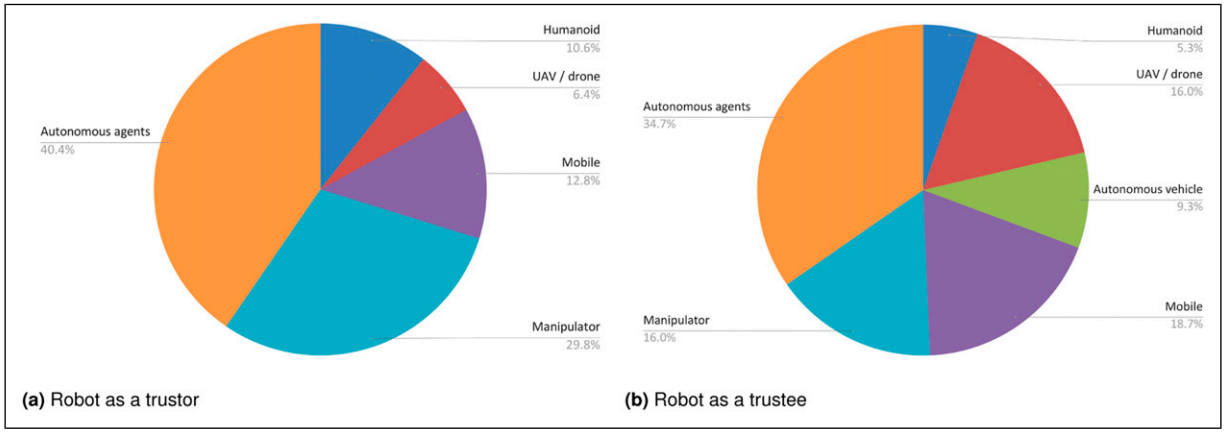

Figure 5. Distribution of the types of robotic platforms used in the experiments covered in this review. The most common class is Autonomous Agents, which aligns with the fact that most experiments have been performed in simulation (see Section 6). (a) Robots that have been used as trustors, that is, those enabled with Artificial Trust capabilities. (b) Robots used as trustees, endowed with Natural Trust prediction skills.

real-world experiments in which the multi-modal communication capabilities of the platforms are exploited to indicate their trust levels towards the trustees (Kirtay et al., 2023; Patacchiola and Cangelosi, 2022; Rahman, 2019b; Vinanzi et al., 2019). A similar trend can be observed for the papers in which humanoids play the trustee role, as 3 out of the 4 papers also present real-world experiments (Kirtay et al., 2023; Rahman, 2019b; Wagner et al., 2018). This use of multi-modal communication suggests that, possibly due to the nature of their embodiment and the application domains in which they are primarily used, where the other interacting agent plays a pivotal role, these robotic platforms are typically utilized for in-person evaluations.

4.2. Aerial robots

Aerial robots include drones and Unmanned Aerial Vehicles (UAVs): platforms capable of sustained flight with or without human intervention. These devices are most frequently used in human-in-the-loop scenarios, where a person modulates their degree of control over the robot based on how much they trust it. For this reason, there are only 3 papers, authored by Saeidi et al., that use aerial robots as trustors (Saeidi et al., 2016, 2017a; Saeidi and Wang, 2019). The latter present experiments in the domains of MAS and Teleoperation, proposing bilateral trust models. The majority of papers discussing CT in the context of aerial robots place them at the receiving side of the trust relationship (Guo et al., 2023b; Hoogendoorn et al., 2014; Mahani et al., 2021; Nam et al., 2017, 2020; Pang et al., 2021; Saeidi et al., 2017b; Xu and Dudek, 2015a). There are 12 papers that use aerial robots as trustees, the oldest of which is from 2014 (Hoogendoorn et al., 2014), and the latest from 2023 (Guo et al., 2023b). Four of these papers fall within the Reconnaissance application domain (Guo et al., 2023b; Hoogendoorn et al., 2014; Mahani et al., 2021; Pang et al., 2021), 4 within MAS (Nam et al., 2017, 2020; Saeidi et al., 2017a; Xu and Song., 2021), 4 within Telerobotics (Saeidi et al., 2016, 2017a, 2017b; Saeidi and Wang, 2019), and one within HRC (Xu and Dudek, 2015b).

Unlike the works outlined in Section 4.1, those that use aerial robots mainly present results from simulation-based studies, with 10 out of 12 papers falling into this category (Guo et al., 2023b; Mahani et al., 2021; Nam et al., 2017, 2020; Pang et al., 2021; Saeidi et al., 2016, 2017a; Saeidi and Wang, 2019; Xu and Dudek, 2015a, Xu and Song, 2021). This shift into simulated environments could be due to the fact that much of these works deal with teams (Guo et al., 2023b; Mahani et al., 2021; Pang et al., 2021; Saeidi et al., 2017a) or swarms (Nam et al., 2017, 2020; Xu and Song., 2021) of aerial robots, making real-life experiments more complex and resource-consuming. The tasks often involve patrolling (Guo et al., 2023b; Mahani et al., 2021; Nam et al., 2017, 2020; Xu and Dudek, 2015b, Xu and Song., 2021), path planning (Pang et al., 2021), and control allocation (Saeidi et al., 2016, 2017a; Saeidi and Wang, 2019), which are easily simulated without deploying aerial robots in real-world settings.

4.3. Autonomous vehicles

An autonomous vehicle is a self-driving vehicle capable of navigating and operating without human intervention. All the 7 papers that cover their use in the context of CT modeling assign them the role of trustees (Cheng et al., 2023; Hsieh et al., 2022; Hu et al., 2022; Rjoub et al., 2023; Soh et al., 2020; Sun et al., 2023; Wang et al., 2023b). As a very recent area of research, the oldest paper is from 2020 (Soh et al., 2020), while 4 are from 2023 (Cheng et al., 2023; Rjoub et al., 2023; Sun et al., 2023; Wang et al., 2023b). All of them fall into the Autonomous Vehicles application domain category, with Soh et al. (2020) also

addressing HRC. The absence of CT models that use autonomous vehicle as trustors is notable and represents a gap in the current literature.

Similarly to what has been observed in the context of aerial robots in Section 4.2, most papers of this class (5 out of 7) present simulation studies (Cheng et al., 2023; Hu et al., 2022; Rjoub et al., 2023; Soh et al., 2020; Wang et al., 2023b). The exceptions are one paper from Hsieh et al. (2022), who collect trust measurements from the human during a real-world drive, and Sun et al. (2023) who run an online experiment, collecting data through a questionnaire to build a CT model.

4.4. Mobile robots

Mobile robots encompass all robotic platforms that are designed to move around an environment but do not classify as vehicles. Only 6 papers use mobile robots as trustors, with the oldest being (Mahani and Wang, 2017) from 2016 and the most recent (Ma et al., 2022) from 2022. The most common application domain, used by 3 papers, is Navigation (Dorbala et al., 2021; Dubey and Kumar, 2019; Mahani and Wang, 2017). The remaining 3 papers cover experiments that have been conducted in the realms of HRC (Ma et al., 2022) and MAS (Rishwaraj and Ponnambalam, 2017), plus one work classified as Miscellanea (Wagner et al., 2018).

Thirteen of the reviewed papers use mobile robots as trustees. The oldest is (Wang et al., 2014) from 2014, and the most recent are (Zahedi et al., 2023; Zheng et al., 2023a) from 2023. Application domains for mobile robots as trustees include: HRC (Soh et al., 2020; Zahedi et al., 2023), Reconnaissance (Bhat et al., 2022; Wang et al., 2014; Zhou et al., 2021), Navigation (Dubey and Kumar, 2019; Zahedi et al., 2023), MAS (the largest category, represented by 6 papers) (Ponnambalam et al., 2021; Rishwaraj and Ponnambalam, 2017; Spencer et al., 2016; Wang et al., 2018; Zheng et al., 2018, 2023a), and Miscellanea (Wagner et al., 2018).

Mobile robots have been used in robotics and AI research for a very long time. Modular platforms like TurtleBot make it easy for researchers to carry out real-world experiments. For this reason, we see an almost equal split of simulation-based validations and real-world experiments being presented in works using mobile platforms. Specifically, 6 works test their models through simulation studies (Ma et al., 2022; Mahani and Wang, 2017; Soh et al., 2020; Spencer et al., 2016; Wang et al., 2014, 2018), 3 use real-world experiments (Dubey and Kumar, 2019; Ponnambalam et al., 2021; Zhou et al., 2021), and 5 use a two-phase validation involving both simulations and real-world experiments (Dorbala et al., 2021; Rishwaraj and Ponnambalam, 2017; Wagner et al., 2018; Zahedi et al., 2023; Zheng et al., 2023a). Works requiring simulation-based evaluations mostly deal with multi-agent systems and teleoperation (Rishwaraj and Ponnambalam, 2017; Saeidi and Wang, 2015; Spencer et al., 2016; Wang et al., 2014, 2018; Zheng et al., 2018, 2023a).

4.5. Manipulators

With the term manipulators, we refer to industrial robots as arm-like structures that can manipulate materials and objects within their workspace. From our pool of selected papers, 14 employ them as trustors. The oldest are (Rahman et al., 2016a; Ranasinghe et al., 2015) from 2015, while the newest are (Lang et al., 2023a, 2023b; Wang et al., 2023a) from 2023. The most common application domain for manipulators is HRC, which received contributions from 14 papers (Almohamade et al., 2021; Hannum et al., 2020; Lang et al., 2023a; Maithani et al., 2019; Rahman et al., 2016a, 2016b; Rahman, 2019a; Ranasinghe et al., 2015; Scherf et al., 2022; Tilloo et al., 2022; Vinanzi et al., 2021; Wang et al., 2022, 2023a; Chen et al., 2018), but there are also works that use them for HRI (Lang et al., 2023b).

Manipulators are used as trustees in 12 papers, the oldest of which (Rahman et al., 2016a) from 2015 and the newest being (Mangalindan et al., 2023; Zheng et al., 2023b) from 2023. All these papers belong to the HRC application domain (Kumar and Dubey, 2017; Losey and Sadigh, 2019; Mangalindan et al., 2023; Rabby et al., 2020; Rahman, 2019a; Rahman et al., 2016a; Sadrfaridpour et al., 2016, 2018; Sadrfaridpour and Wang, 2018; Wu et al., 2017; Zheng et al., 2023b), with one of them, authored by Mangalindan et al. (2023), also being classified as HRI.

The application domains for manipulators, much like those for humanoid robots (Section 4.1), require the physical co-presence of the robotic platform and the agent they are collaborating with. For this reason, 16 papers describe real-world experiments (Losey and Sadigh, 2019; Rahman et al., 2016a, 2016b; Rahman, 2019a; Sadrfaridpour and Wang, 2018; Zheng et al., 2023b; Almohamade et al., 2021; Hannum et al., 2020; Lang et al., 2023a, 2023b; Maithani et al., 2019; Ranasinghe et al., 2015; Tilloo et al., 2022; Wang et al., 2022, 2023a; Chen et al., 2018), while 6 rely on simulations (Kumar and Dubey, 2017; Mangalindan et al., 2023; Rabby et al., 2020; Sadrfaridpour et al., 2018; Vinanzi et al., 2021; Wu et al., 2017), and only 2 present double validation (Sadrfaridpour et al., 2016; Scherf et al., 2022).

4.6. Autonomous agents

Papers that use robots not falling into any of the aforementioned categories, as well as those that do not refer to a specific type of platform but simply to "robots" or "agents," fall into this category.

Our review has found 20 papers which use generic or unspecified robots as trustors, spanning from 2015 (Aydoğan et al., 2015) to 2023 (Li et al., 2023; Lin et al., 2023; Rjoub et al., 2023; Sapienza and Falcone, 2023). The

most represented application domains are MAS (Lin et al., 2023; Rishwaraj et al., 2017; Setter et al., 2017) and HRI (Lee et al., 2021; Li et al., 2023; Patacchiola and Cangelosi, 2016; Tjøstheim et al., 2019), each with 4 papers. Other works fall into the domains of HRC (Ali et al., 2022; Kumar and Dubey, 2017), Navigation (Alhaji et al., 2021; Dorbala et al., 2021; Rutard et al., 2020), Autonomous Vehicles (Cheng et al., 2021; Khattar and Eskandarian, 2022; Rjoub et al., 2023), Telerobotics (Aydoğan et al., 2015), and Miscellanea, including gaming applications (Carneiro et al., 2019), recommender systems (Kang, 2018), and smart devices or Internet of Things (Sapienza and Falcone, 2023).

Similarly, 28 papers place generic or unspecified robots in the role of trustees, spanning from 2013 (Lee et al., 2013; Mansoor et al., 2013) to 2023 (Carbo and Molina, 2023; Guo et al., 2023a; Hu and Wang, 2022; Lin et al., 2023). Of this total, 3 fall into the HRC domain (Ali et al., 2022; Chen et al., 2020; Rahman, 2019c), 7 in Reconnaissance (Boer et al., 2013; Floyd et al., 2014, 2015; Guo et al., 2021, 2023a; Guo and Yang, 2021), 4 in MAS (Lin et al., 2023; Rishwaraj et al., 2017; Setter et al., 2017; Wang et al., 2015), 4 in HRI (Abdulhussain and Aziz, 2022; Hale et al., 2019; Hu and Wang, 2022; Lee et al., 2013), 2 in Autonomous Vehicles (Cheng et al., 2021; Mansoor et al., 2013), one in Navigation (Carbo and Molina, 2023), and 5 in Miscellanea, including robot-based therapy (Ab Aziz et al., 2017; Aziz and Abdulhussain, 2022), games (Carneiro et al., 2019; Razin and Feigh, 2021), and surgery (Yan et al., 2022).

Since most of the works in this category do not specify a robotic platform on which they deploy their CT models, it is not surprising that 37 of them rely on simulations (Ab Aziz et al., 2017; Abdulhussain and Aziz, 2022; Alhaji et al., 2021; Ali et al., 2022; Aydoğan et al., 2015; Aziz and Abdulhussain, 2022; Carbo and Molina, 2023; Chen et al., 2020; Cheng et al., 2021; Floyd et al., 2014, 2015; Guo et al., 2021, 2023a; Hale et al., 2019; Hu and Wang, 2022; Kang, 2018; Khattar and Eskandarian, 2022; Kumar and Dubey, 2017; Lee et al., 2021; Lin et al., 2023; Mansoor et al., 2013; Patacchiola and Cangelosi, 2016; Razin and Feigh, 2021; Rishwaraj et al., 2017; Rjoub et al., 2023; Rutard et al., 2020; Sapienza and Falcone, 2023; Setter et al., 2017; Tjøstheim et al., 2019; Wang et al., 2015) or online studies (Akash et al., 2019; Boer et al., 2013; Carneiro et al., 2019; Guo and Yang, 2021). The 3 papers that provide real-world evaluations feature human participants interacting with robotic platforms or autonomous systems that cannot be classified as humanoids, aerial robots, mobile robots, manipulators, or autonomous vehicles. For example, Rahman (2019c) tests their CT model on an in-house built power assist system. Similarly, Yan et al. (2022) conduct experiments using a novel robotic system for vascular intervention surgery they developed, while Li et al. (2023) develop a generic conversational agent to interact with their human participants. Dorbala et al. (2021) classify both in the Autonomous Agents and the Mobile Robots categories because they start their validation with a simulation using an unspecified agent and complement it with a real-world experiment using a mobile robot.

4.7. Insights

The decision of which robot platform to use when developing and evaluating CT depends on factors such as the scenario intended for their use, the feasibility of deploying the robotic platform in the intended scenario in real-world settings, and the availability of the platform at the different institutions.

Mobile robots have a long-standing presence in robotics research, as evidenced by the large number of papers that deploy their systems on such platforms. Mobile robots have consistently offered researchers reliable simulators and a cost-effective means to conduct in-person experiments. This is clearly reflected in our review, where mobile robots were used in an equal number of studies involving real-world deployments (Dubey and Kumar, 2019) and simulations (Ma et al., 2022).

In our analysis, humanoids and manipulators emerged as the most commonly used robot platforms for in-person experiments, likely due to the demands of scenarios that require multi-modal communication (Kirtay et al., 2023) or physical co-presence to complete a task (Losey and Sadigh, 2019). Interestingly, manipulators are more frequently used than humanoids. This may be due to the higher cost of humanoid robots and their association with an only recent industrial push compared to the longer established use of manipulators.

Aerial robots and autonomous vehicles are also vastly used by the community, but mostly in simulated experimental setups. Aerial robots are mostly used in human-in-the-loop scenarios, with their role being most frequently that of trustees (Guo et al., 2023b). Similarly, autonomous vehicles are usually the trustees when they are involved in a study (Cheng et al., 2023). These differences with respect to humanoids and manipulators are mostly due to the nature of the task and environments in which these type of platforms are meant to be deployed, generally involving a higher degree of risk for the platforms and the people involved.

It is less surprising to see that autonomous agents without a specific embodiment have also been mostly used in simulation (Alhaji et al., 2021). It is important to acknowledge works that do not necessarily use pre-defined platforms as they have the potential to generalize to more than one platform. For this reason, future works are welcome to test these solutions for CT in specific embodiments.

In general, looking at the choices of platforms, where they have been deployed and how, we can identify a clear area of improvement for the field: more in-person experiments are needed to test generalizability and specificity of the models proposed.

5. Computational models of artificial and natural trust

This section delves into the computational models used in the selected papers, aiming to identify trends in the state-of-the-art. We have categorized techniques into 4 clusters present across our pool of scientific publications: (A) Deterministic, (B) Probabilistic, (C) Machine Learning (ML), and (D) Game Theory, each discussed in detail in the subsequent subsections.

Figure 6 summarizes the frequency of adoption of these classes of techniques within our selected sample. Among the selected 101 papers, 92 fall within a single category. The majority of works (54 papers) belong to the Deterministic class, followed by Probabilistic (28 papers). Despite the contemporary popularity of ML and Deep Learning solutions to develop AI systems, only 8 papers belong to that group. Finally, 2 papers adopt Game Theory approaches of CT. For the purpose of this analysis, we have separated the papers that use multiple classes of computational models in a separate category, which we name Composite. Nine papers belong to this class, of which 6 combine Probabilistic and ML methods. The latter is the most common hybridization, which is expected due to the increasing popularity of Reinforcement Learning (RL) methodologies and their natural pairing with Markovian models. Deterministic methods have been combined with Probabilistic techniques twice (Sun et al., 2023; Zheng et al., 2023a) and with ML only once (Lin et al., 2023).

Figure 7 illustrates the trends in adoption of different classes of CT techniques during the years covered by this review, from 2013 to 2023. One notable observation from this graph is that Deterministic approaches have begun steadily, with moderate oscillations, indicating a consistent but not dominant preference within the field, and overall emerge as the most commonly adopted class of methods. It is closely followed by the Probabilistic category, which has gained momentum since 2018 and has managed to dominate the research landscape in CT for a few intermittent years. ML techniques had a spike in 2017 followed by a dip that lasted several years: only from 2021 they are seeing a slow but steady increase in popularity which seems to persist nowadays. Probabilistic and ML methodologies seem to exhibit comparable growth patterns, which align with the increasing popularity of RL solutions to complement Markovian models, as discussed previously. Overall, in the most recent years, Deterministic techniques seem to have re-emerged as prominent. CT models based on Game Theory appear sparingly, as do Composite models, which employ more than one class of techniques.

5.1. Deterministic models

Deterministic models are mathematical or computational models that predict the outcome of a system with certainty based on specific inputs or initial conditions. In deterministic models, there is no randomness or uncertainty involved in the prediction process, rather they are characterized by a cause-and-effect relationship which leads them to produce the same output every time they are run with the same set of input parameters. These models assume that the future behavior of the system is completely

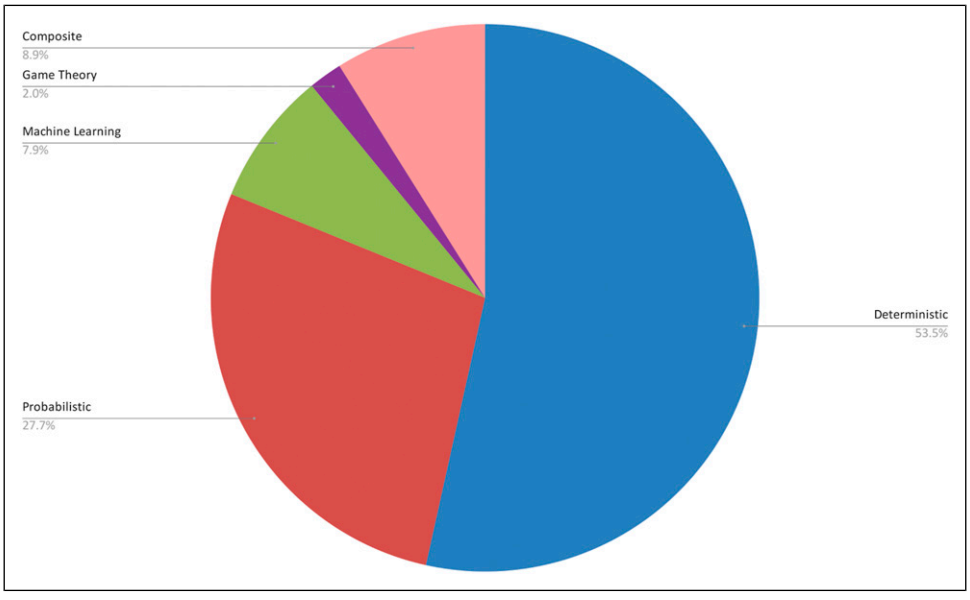

Figure 6. Distribution of the computational model grouping across the selected papers. Deterministic models appear as the most common choice, followed by Probabilistic. Papers that make use of multiple models have been classified as Composite.

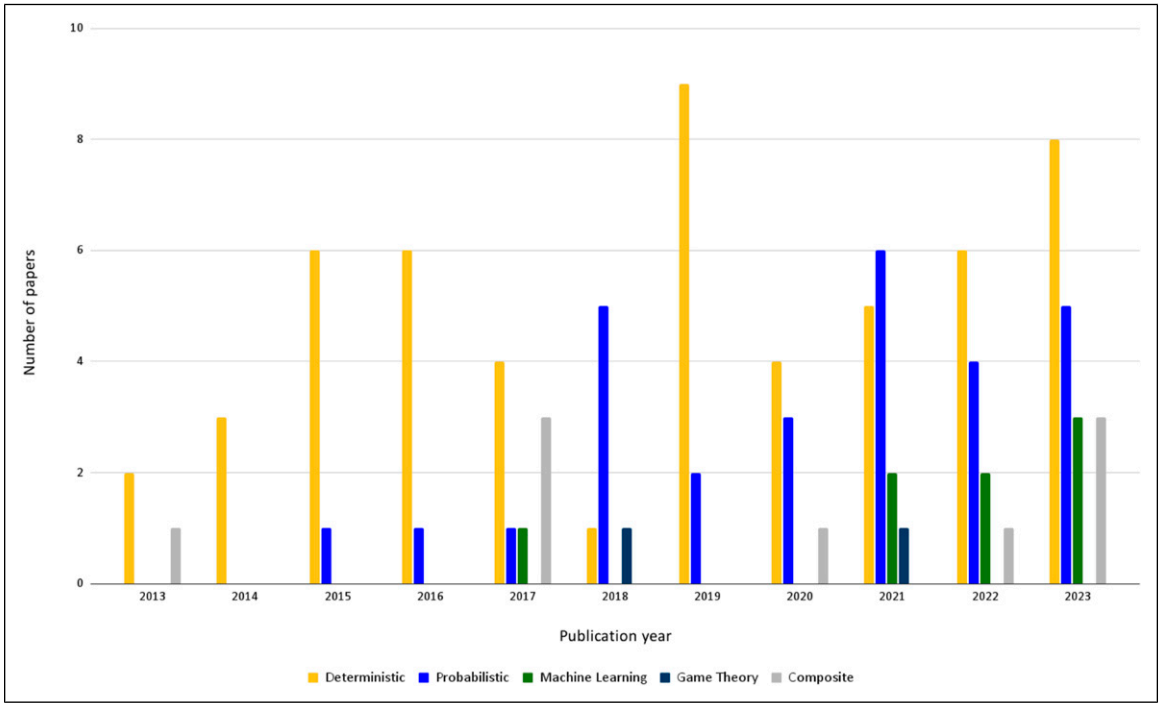

Figure 7. Yearly trends in the choice of techniques to model Computational Trust. Deterministic models appear to be the most popular, despite a close competition from Probabilistic ones that in recent years appears to be turning in favor of the former. Computational Trust models based on Machine Learning are on a growing trend since 2021.

determined by its current state and the rules governing its dynamics.

As discussed in Section 5, deterministic models have been the preferred choice in CT for robotic agents for the past 10 years, and they still dominate the research landscape today. Out of 57 papers using deterministic models, 35 of them employ some form of linear combination. The latter refers to a mathematical operation where two or more terms are added together, each term being multiplied by a constant coefficient. For example, Ab Aziz et al. (2017) measure short- and long-term trust as a combination of factors including transparency, risk perception, behavior, competency, deception, and experience, while Boer et al. (2013) use a measure of past trust, subject to temporal decay, in combination with a factor expressing the perceived experience level of the other agent in the context of a question-answering task. A similar approach is adopted by Hale et al. (2019), whose system calculates trust based on previous trust values and the gradient of a cost function over the human's privacy and cooperative attitudes. In general, the largest overall number of papers in the selected pool gravitates towards the use of a linear combination of factors that attempt to describe certain physical or mental characteristics of the trustee, such as performance-related factors (success rate, speed of execution, degree of attention, and so on) (Rabby et al., 2020; Rahman, 2019b; Sapienza and Falcone, 2023). These parameters are usually either measured during an experimental task (Floyd et al., 2015) or manually set as parameters for a simulation (Hu et al., 2022). A third option that has been utilized is to extract these values from user studies, that is conducting an experiment with human participants, collecting data from them through questionnaires, and using the latter to fine-tune a linear model (Mansoor et al., 2013).

Other works in the field make use of non-linear combinations. In contrast to linear ones, where terms are added together with constant coefficients, non-linear combinations involve operations such as multiplication, exponentiation, division, or other non-linear functions applied to the terms. For example, Alhaji et al. (2021) model the evolution of the robot's trust in the human user utilizing an initial trust value T_{init} and calculating the trust T_t at time t as a function of a weighting factor $\omega \geq 1$, a risk factor $r \in \{0, 1\}$, and a dynamic measure of the human's predictability.

Other techniques appear, albeit less commonly. Several papers make use of methodologies based on differential equations, which are used to describe trust in relation to derivative terms, like the change in an agent's confidence value (Aydoğan et al., 2015). Tjøstheim et al. (2019) propose a biologically inspired computational model of the brain that is able to regulate its levels of trust towards a user following painful and gentle touches. The model is composed of a series of interconnected brain nuclei together paired with larger cortical regions. The nuclei are governed by a set of differential equations that model excitatory and inhibitory stimuli. Ranasinghe et al. (2015)

make a robot guide a blindfolded user through a hard rein and model the follower's dynamics (in terms of trust or distrust) using a time varying virtual damped inertial system. Finally, a few works model trust as piecewise functions, defined by different expressions over different intervals of their domain (Almohamade et al., 2021; Carbo and Molina, 2023; Cheng et al., 2021; Hu et al., 2022; Hu and Wang, 2022).

5.2. Probabilistic models

Probabilistic models are mathematical techniques that incorporate uncertainty by representing random variables and their associated probabilities. These models use probability theory to describe the likelihood of different outcomes or events occurring within a given system. Unlike deterministic models, which produce the same output for a given set of inputs, probabilistic models account for variability and uncertainty in the data or parameters.

Two different methodologies appear to be the most commonly adopted to develop CT capabilities in artificial agents: Bayesian and Markovian (respectively, 10 and 19 papers). The former refers to those classes of statistical models that use Bayesian probability theory to represent uncertainty and update beliefs or probabilities based on observed data. These models start with prior beliefs about the parameters, which are then updated using Bayes' theorem to incorporate new evidence or observations. For example, Vinanzi et al. (2021, 2019) use a Bayesian Network to represent a robot's belief about the human's knowledge and intentions, which is then used to discriminate helpers from trickers in a sticker finding game. However, researchers make more often use of Dynamic Bayesian Networks, which extend the traditional framework to represent and analyze dynamic systems that evolve over time. Several papers (Mahani et al., 2021; Sadrfaridpour et al., 2018) formalize trust at time t as dependent on the performance of the robot at time t, other than trust and performance at the previous timestep $t-1$. Wang et al. (2018) expand on this approach by explicitly modeling robot performance, human performance, and joint human-robot system faults.

Where Bayesian models focus on updating beliefs or probabilities based on evidence, Markovian models focus on modeling sequential dependencies and transitions between states in a system. They are based on the concept of a Markov process, which describes a stochastic process where the future state depends only on the current state and is independent of past states, given the present. There are three classes of Markovian models that are commonly adopted by researchers: Markov Decision Processes (MDPs), Partially Observable Markov Decision Processes (POMDPs), and Hidden Markov Models (HMMs). While all three deal with uncertainty and sequential data, MDPs focus on decision-making in fully observable environments, POMDPs address decision-making in partially observable environments, and HMMs are used primarily for modeling hidden states in observed sequences without involving decision-making.

MDPs represent decision-making problems where an agent interacts with an environment over time, making decisions to maximize expected cumulative rewards. They have been used to model the way in which trustworthy and untrustworthy humans would solve a given task, giving the robot a way to perform appropriate decision-making based on observation (Losey and Sadigh, 2019; Wu et al., 2017).

POMDPs extend MDPs by considering situations where the agent lacks complete information about the environment's state, requiring the agent to maintain a belief state and make decisions based on partial observations. Akash et al. (2019) use them to describe the trust dynamics based on the observation of compliance and response time of the other agent and the adoption of a set of actions. Mangalindan et al. (2023) develop a framework that models the influence of trial complexity, experience gained, the robot's action and previous trust on the current level of trust.

HMMs model time-series data where the underlying state of the system is hidden and must be inferred from observed emissions. Lee et al. (2013) model trust from the observation of physical cues, which they refer to as high-trust and low-trust cues, and their temporal evolution over time. Xu and Howard (2020) adopt a similar approach: they have a trial in which a human takes an action, the robot gives a recommendation and a final action is decided, and train an HMM to determine the underlying trust.

5.3. Machine learning models

ML is a branch of AI focused on developing algorithms and techniques that enable computers to learn from data and improve their performance on specific tasks without being explicitly programmed.

The majority of reviewed papers which use ML to implement CT capabilities rely on RL: specifically, 10 out of 15. RL is based on the idea of agents learning to make sequential decisions by interacting with an environment to maximize cumulative rewards. Unlike supervised learning, where the agent is provided with labeled examples, and unsupervised learning, where the agent must discover hidden patterns in data, RL relies on trial and error. Several papers (Lin et al., 2023; Ma et al., 2022) aim to calculate the robot's trust in the human's supervision during an exploration task. They propose a metric based on the agent's state and capability, and the impact of decisions, then define a reward function based on the latter and finally fit a model through the Q-learning algorithm, through which the agent learns to make decisions by estimating the value of taking each action in a given state.

Where RL is designed to compute optimal policies, it often relies on Markovian models to provide a formal representation of the underlying dynamics of the environment, which it then exploits to efficiently explore and learn optimal strategies. For this reason, these two mathematical tools often work side-to-side. Rishwaraj and Ponnambalam (2017) develop a multi-agent system in which a robot is able to identify other trustworthy robots. To achieve this objective, they model the decision-making process of the agent to maximize the expected utility it would receive when moving to the next state after taking a specific action, and an optimal policy is computed through RL. Nam et al. (2017, 2020) use the same principle, but employ Inverse Reinforcement Learning (IRL): unlike traditional RL, where the goal is to learn a policy that maximizes cumulative rewards, in IRL the objective is to understand the reward function that best explains the observed behavior of an expert agent. Contextualized in the scope of their experiments with robot swarms, the computational model needs to infer the degree of trust of a human operator providing commands.

Despite RL appearing as the most common ML technique in the pool of selected papers, the presence of other techniques is also noteworthy. Pang et al. (2021) make use of an Artificial Neural Network (ANN): a computational model inspired by the structure and function of the human brain, consisting of interconnected nodes (neurons) organized in layers, capable of learning complex patterns from data. In particular, they use it to process flight and trajectory data from a drone and use it to predict the level of trust it would elicit in a human operator.

Hsieh et al. (2022) utilize an ensemble comprised of three models to quantify the trust of a human towards an autonomous vehicle: an ANN, a Support Vector Machine (SVM), and a Random Forest (RF). A SVM is a supervised machine learning algorithm used for classification and regression tasks, aiming to find the optimal hyperplane that separates different classes or fits the data with the maximum margin, while a RF is an ensemble learning method consisting of multiple decision trees: models consisting of a tree-like structure where each internal node represents a decision based on the value of a feature, and each leaf node represents the outcome or prediction. These three models were trained on vehicle data obtained during a user study with human participants. Kraus et al. (2021) also rely on three different ML models, but instead of using them as an ensemble, they train them separately on the same dataset and compare their ability to predict trust. In particular, they make use of a SVM, gradient boosting and a Gated Recurrent Unit (GRU) network, trained on an annotated data corpus of proactive dialogues. Gradient boosting is a technique that consists in sequentially combining the predictions of multiple weak learners, while GRU is a recurrent neural network, a type of ANN designed to process sequential data by maintaining a memory of previous inputs. Patacchiola and Cangelosi (2022) experiment on a biologically inspired cognitive architecture for CT that uses Bayesian networks and Self-Organizing Maps (SOM): unsupervised ANNs that reduce the dimensionality of input data while preserving its topological properties by organizing it into a lower-dimensional grid.

5.4. Game Theory models

A minority of papers (2 out of 101) are based on Game Theory: a mathematical framework used to study strategic interactions between rational decision-makers, known as players, in situations where the outcome of each player's action depends on the actions of others, aiming to predict and analyze the optimal strategies and outcomes of such interactions. Game Theory formalizes the strategic interaction between two or more players as a payoff matrix: a tabular representation of the possible outcomes. Each cell in the matrix corresponds to a combination of strategies chosen by the players, and it contains the payoffs or utilities associated with those strategies. Wagner et al. (2018) compute several payoff matrices that represent different trust or no-trust situations which an agent can use to explain another agent's behavior, and validate them through a user study. They then use this model on a robot to try and predict the levels of trust of the other party. Razin and Feigh (2021) also base their work on payoff matrices and use interdependence theory, which aims to determine which actor has power over which part of the total payoff structure. Through the latter, they define a set of variables which then proceed to compose in a "trust index" derived from Gottman's.

5.5. Insights

The landscape of CT modeling is marked by both methodological diversity and uneven adoption patterns. While the field has matured in some respects, our review reveals several areas where further development and refinement are needed.

A clear trend is the predominance of deterministic models, particularly those based on linear combinations of trust-related factors. Their continued popularity suggests a strong preference for models that are interpretable and easy to implement. However, this reliance may also reflect a certain conservatism in the field, where simplicity is favored over expressiveness. As trust modeling increasingly intersects with complex, real-world scenarios, researchers may need to move beyond these traditional approaches.

Probabilistic models offer a compelling alternative, especially in contexts where uncertainty and temporal dynamics are central. Their growing use indicates a shift toward more flexible and realistic representations of trust.

Despite the rising tendency, their adoption remains uneven, and their integration with other modeling paradigms is still limited. Probabilistic reasoning could then serve as a bridge between interpretable models and data-driven approaches, enabling systems that are both robust and adaptive.

ML, despite its transformative impact on broader AI, remains underutilized in CT. While promising, this approach is still relatively rare and often lacks standardization in terms of datasets, evaluation metrics, and experimental design. This suggests that the field has yet to fully embrace the potential of ML for modeling trust.

Game Theory models, though conceptually well suited to trust as a strategic phenomenon, are scarcely used. This underrepresentation is surprising, given their ability to formalize interdependence and rational behavior in MAS. Their limited presence may be due to the complexity of implementation, requiring precise formalization of agent strategies and payoffs or a lack of suitable experimental frameworks. Nonetheless, they remain a promising avenue for future exploration, particularly in competitive or cooperative scenarios.

One of the most significant gaps identified is the limited use of hybrid or composite models. While a few studies combine deterministic and probabilistic techniques with ML models, systematic efforts to integrate multiple modeling paradigms are rare. This is a missed opportunity: hybrid models could leverage the strengths of different approaches, combining the interpretability of deterministic models with the adaptability of ML or the uncertainty handling of probabilistic methods.

6. Experimental settings

This section is dedicated to exploring the different settings that researchers in the field of CT have utilized for deploying their models. We consider three different cases: (A) Simulation, (B) Real-world, and (C) Online, all of which are discussed in detail in the following subsections. Figure 8 shows the distribution of these categories across the selected pool of papers. Approximately half of them (55 out of 101) test their models in simulation. Thirty works deploy their computational architectures in the real world, while only a minority (8 papers) test them online. We have added the Composite category to group all those papers that utilize more than one experimental setting. There are 8 papers in this category, of which 7 adopt simulations that are then brought to the real world. Only one work uses both a physical and an online implementation (Wagner et al., 2018) and only one is performed in a mixture of simulated and online settings (Carneiro et al., 2019).

6.1. Computer simulations

This review highlights that the majority of CT research papers in literature tests their models in digital environments. This trend is unsurprising, given that digital settings are the most accessible, often not requiring specialized hardware such as robotic platforms. These environments are diverse, spanning from pure numerical simulations to fully developed graphical and physical environments.

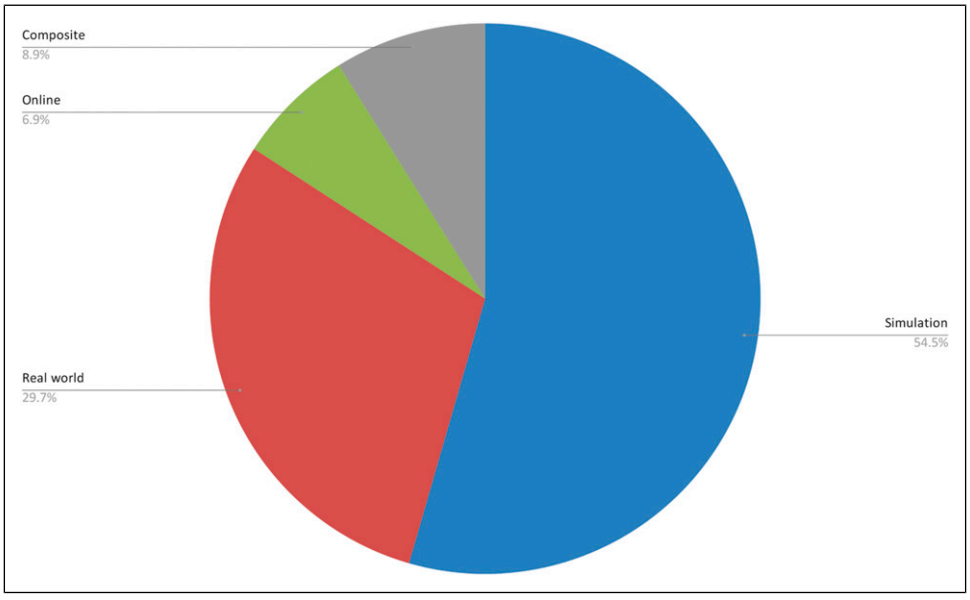

Figure 8. Distribution of the experimental environments used by the selected papers. Trials performed in computer simulations are the most commonly adopted, followed by real-world experiments with physical robots. Papers that make use of multiple settings have been classified as Composite.

We speak of numerical simulations when the authors do not leverage a fully sketched-out virtual world where an agent can move and interact, but rather a framework in which a model is tested against different combinations of parameters. Ab Aziz et al. (2017), for example, have developed a mathematical model that can predict trust based on the internal and external characteristics of the trustee, including but not limited to: reliability, transparency, and competency. They test different combinations of trustee characteristics, manually selected, to evaluate the output of their model, that is, the predicted level of trust towards that agent. The same approach is shared by several other works (Lee et al., 2021; Wang et al., 2015, 2023b).

Graphical simulations take one step further: in addition to emulating the model's mathematical framework, they also provide a simple environment for the agent to interact with. This simplicity often takes the shape of a gridworld: a grid of cells where agents can perform discrete actions according to certain rules (Carbo and Molina, 2023; Rutard et al., 2020; Spencer et al., 2016), for example, to perform path planning or maze solving guided by teachers that are more or less trustworthy.

Finally, there are complete physical simulations. These are complex environments, often three-dimensional, where agents can perform both perception and action in a continuous fashion. Hu and Wang (2022) make use of the CarSim-Simulink platform, a software framework that offers a complete rendering of a vehicle and the surrounding road, in which they test their trust-aware cruise control system. Another example comes from Lin et al. (2023), who use the popular robotics simulator Webots to build an environment in which a small swarm of robots can navigate and communicate in the search for a set of items scattered around the world. Another commonly used simulator is Gazebo, adopted for instance in an experiment authored by Dorbala et al. (2021), in which the robot had to navigate an environment by accepting or rejecting human guidance.

6.2. Real-world experiments

Experiments using real robotic platforms are the second most common found in the pool of reviewed papers. All works that deploy their computational models on an embodied machine that physically interacts with the real world are categorized under this heading. The majority of these experiments take place within a laboratory setting. The advantage of this approach is that it provides researchers with a more structured environment in which to test their software. For example, Vinanzi et al. (2021) utilize a Pepper robot placed in front of a table interacting with special markers, while Patacchiola and Cangelosi (2022) deploy their cognitive architecture on an iCub robot that had to interact with sets of objects shown to it by the experimenter. The latter cases involve commercially available robots, but not in every case: Rahman (2019c) built a robotic power assist system and had it perform some lifting tasks with the assistance of human participants.

Less commonly, these experiments take place outside the boundaries of the laboratory, in a more unstructured environment. Because of the complexity involved in deploying a robot in such settings, these experiments are more scattered across the literature. Hsieh et al. (2022), for instance, collected trust data from users onboard a Tesla autonomous vehicle driving on a real road.

6.3. Online experiments

We have categorized experiments as "Online" when they entail user studies where participants watch videos or engage with simulations. Such studies typically involve observing users as they complete tasks, gathering feedback through interviews or surveys, and analyzing the data to address research questions.

Not all the experiments in this category are strictly conducted via online platforms. For example, Bhat et al. (2022) develop a graphical simulation using the Unreal Engine 4 game engine and test with in-person participants. Similarly, Guo and Yang (2021) require physical presence for their trials due to the use of a specialized joystick, which make a purely online execution unfeasible.

Other studies utilize cloud-based solutions like online crowd-sourcing platforms or custom servers. Hoogendoorn et al. (2014) use a custom graphical interface to allow their participants to interact with the image of an area, within which they have to detect threats. Xu and Howard (2020) utilize Amazon Mechanical Turk to collect user data regarding different scenarios involving robot providing suggestions. This same tool is used by other works (Akash et al., 2019; Wagner et al., 2018).

6.4. Insights

As noted in Section 4, most of the works in our review use simulations as the primary means of testing their models. This is unsurprising, given the accessibility and flexibility that simulation environments offer. The complexity of these simulations varies widely, ranging from mathematical framework solving (Ab Aziz et al., 2017) to fully developed, game-like three-dimensional environments (Hu and Wang, 2022). Simulations are fundamental for continuous testing, especially in the early stages of development, and for studying scenarios that would be too risky to investigate in real-world settings. However, the AI and robotics community is well aware of the sim-to-real problem that sees performances decline when models are deployed on physical robots due to the discrepancies between simulated environments and the real world. In addition, most of the real-world experiments are still run in lab environments,

providing a structured and reproducible environment for researchers to test their solutions (Vinanzi et al., 2021) but shying away from the most complex and unstructured environments.

The research community now has a valuable opportunity to extend its work beyond simulations and laboratory settings, by pursuing more ecologically valid experimental designs. This shift will enable deeper insights into how results may change when environmental constraints are removed and high levels of uncertainty are introduced.

7. Evaluations and results

The analyses carried out and the type of results achieved in CT research within the selected set of papers are strictly dependent on the specific nature of the works. Therefore, a categorization like those pursued in the previous sections (see Sections 3, 4, 5, and 6) would produce overly granular clusters. However, it is still possible to gather useful qualitative insights on CT trends and challenges.

7.1. Evaluation metrics

Around a quarter of the works make use of trust-related measurements. More precisely, they use the trust computed by the CT model itself as a metric to assess the quality of their work. The use of the designed trust score specifically occurs when the robot assumes the role of trustor in the interaction. In these works, trust strongly affects the behavioral model of the robot, so it makes sense to directly use it as an interaction metric. For instance, Aydoğan et al. (2015) and Hannum et al. (2020) both use the trust value computed by their models to investigate how trust dynamics evolve throughout interactions with other agents. Slightly differently, Razin and Feigh (2021) use the mean squared error (MSE) of the computed trust value against a ground truth.

When the robot acts as a trustee, trust is most often measured through self-reported questionnaires. As the robot is the recipient of trust, the latter is evaluated from the other party, which in this case is the human user. In some studies, questionnaires are completed at the end of each condition (Sadrfaridpour and Wang, 2018), especially when statistical comparisons are made to assess which setting is perceived as more trustworthy. In other ones, particularly those aiming to validate a trust model, trust is measured at the end of each interaction iteration (Guo and Yang, 2021; Zheng et al., 2023a).

The majority of works use performance metrics that do not derive directly from the CT model estimate. In these cases, trust is used as a component of a more complex behavioral system. Most of these metrics are related to task performance, as the objective is often to prove that incorporating trust awareness improves efficiency in HRC (Xu and Song, 2021). When assessing human efficiency, a ratio of the number of correct actions to the total performed actions is chosen as the merit parameter (Boer et al., 2013; Vinanzi et al., 2021). When evaluating the efficiency of robotic agents (Xu and Song, 2021), the robot's capability to plan and perform actions is also assessed. For instance, Dubey and Kumar (2019) register the number of best, fair, and unsuccessful plans achieved by the robot in the interaction cycles with another robotic agent of the same nature.

For probabilistic CT models, the aim is often to maximize an internal reward of the framework, which then becomes the evaluation metric of the work (Kirtay et al., 2023). In some cases, the reward is considered at the end of every interaction with the system (Dorbala et al., 2021), while in others, the mean value is used (Chen et al., 2020). When trust is modeled through supervised ML, the evaluation metric is most often the accuracy of the classifier (Hsieh et al., 2022; Kraus et al., 2021) or the prediction error (Lee et al., 2013; Li et al., 2023). Pang et al. (2021) use these metrics while performing ablation studies on their ML models.

7.2. Analyses and results

As commonly performed in research, CT models are validated through comparison with other baseline cases. The vast majority of these works compare the performance of the robotic system under their own experimental settings with and without incorporating trust into the behavioral model (Saeidi et al., 2017b; Zheng et al., 2023b). For example, Saeidi et al. (2017b) compare their trust-based mixed-initiative model with the same system that did not use trust, under three different settings. Since research on CT is in its early stages and there are not yet many validation standards, researchers predominantly prefer to define their own baselines against which to compare their solution's performance.

In two cases, the authors compare their results with other frameworks in the literature. Rishwaraj and Ponnambalam (2017) test their trust evaluation control system against three other trust evaluation methods from different sources in the literature. All four methods were deployed in the same experimental settings and considered as independent variables of the experiment. In the other case, Xu and Dudek (2015a) compare their probabilistic trust model against regression techniques for trust estimation used in other studies.

Another example comes from Carbo and Molina (2023), who investigate how the introduction of different variables in the design affects the trust estimated by their CT model. More precisely, they studied its effect on task performance when it accounted for emotional variables present in the interaction design.

Looking at the results achieved in these papers, the majority primarily aims at validating their experimental

hypotheses. In most cases, either the CT model is validated (Ab Aziz et al., 2017; Aziz and Abdulhussain, 2022; Bhat et al., 2022) or the statistical reduction of the user's workload is demonstrated (Sadrfaridpour and Wang, 2018; Saeidi et al., 2016; Wang et al., 2023a). A few works take a step further by discussing the trade-offs associated with increased trust. For instance, Chen et al. (2020) demonstrate that in some instances of HRC, an increase in the user's trust can potentially lead to a loss of task performance when the robot makes mistakes. In Rahman (2019c), the relationship between trust and the level of assistance the robot provides to the user is investigated. Finally, on the human side of the interaction, Nam et al. (2020) and Pang et al. (2021) infer high-level human preferences for features of the robotic systems based on the estimates provided by their CT models.

7.3. Insights

It is important for this review to address common evaluation metrics and methodologies. As CT is still a relatively recent concept, there remains significant work to be done within the community to establish a consistent evaluation framework. Across the papers we reviewed, we observed a range of approaches used to assess CT: self-reported metrics via questionnaires (Sadrfaridpour and Wang, 2018); trust values computed by the CT models themselves (Aydoğan et al., 2015); and performance metrics that act as proxies for CT (Xu and Song, 2021). The choice of metrics varies considerably and is typically tailored to the specific task and experimental setup in which CT is tested. Currently, there is no consensus on the most appropriate metrics for evaluating CT models, highlighting the need for further work toward defining robust evaluation methodologies.

Regarding evaluation methodologies used to extract key insights from these metrics, the most common approach involves validating CT models by comparing them against baseline cases. The core issue, however, lies in defining what constitutes a valid baseline. Most of the works we analyzed compare system performance with and without the integration of trust into the agent's behavioral model, using their own experimental settings (Saeidi et al., 2017b). This practice largely stems from the absence of established validation standards. While no two tasks or scenarios are identical, this lack of standardization poses challenges to both the reproducibility and generalizability of findings related to CT. Future work could benefit from community-wide efforts to define shared baselines that can serve as reliable controls for evaluating CT models.

8. Discussions

This survey highlights a growing trend in the domain of CT, as evidenced by the increasing number of papers published on the topic. In 2013, only 3 papers were published, whereas in 2023 alone 19 papers were made available to the public. The surge in interest can be attributed to recent articles highlighting the bidirectional nature of trust relationships (Azevedo-Sa et al., 2021; Vinanzi et al., 2021), which encourages researchers to explore this venue further. Additionally, the establishment of workshops, such as the International Workshop on Multidisciplinary Perspectives on Human-AI Team Trust (MultiTTrust) (Brandizzi et al., 2023; Centeio Jorge and Ulfert-Blank, 2023; Tielman et al., 2024, 2025), is bringing together scholars focused on the topic of trust in AI and robotics, including the newborn field of AT (Semeraro et al., 2024b).

After analyzing the individual features of the selected papers throughout Sections 3 to 7, we now perform a cross-analysis that may offer valuable guidance for future research in this sector. Comparing all possible combinations of features would make the discussion unnecessarily lengthy, consequently only the most informative pairings are discussed. We then present a comparison between different approaches to CT modeling, and offer recommendations and guidelines for researchers seeking to contribute to this field, drawing on the lessons learned throughout this review. Finally, we address ethical considerations of robots performing their own trust-based decision-making.

8.1. Distribution of models across domains and experimental settings

Figure 9 presents a cross-analysis of three dimensions from this review: each paper is plotted according to the class of its computational model, its domain of application, and the experimental setting it proposes. The analysis is limited to these specific features, as they provide the most informative value and useful guidance to design CT frameworks. Notably, we excluded the Game Theory class of computational models from this analysis, as it only encompasses two data points (see Subsection 5.4), to maintain clarity.

Figure 9 clearly demonstrates the dominance of Deterministic methods and Simulation experiments across the board, as confirmed in Sections 2 and 6. A new insight provided by this graph is the roughly equal distribution of Deterministic models across the application domains. It is significant because it indicates that this methodology is robust enough to be applied to a wide range of situations and scenarios. Conversely, real-world experiments, often involving an embodied and situated robot, tend to rely less on Deterministic methodologies and more on Probabilistic ones. This shift can be explained by the increased complexity of real-world environments and the need for mathematical models that can account for uncertainty. Probabilistic models tend to be more commonly deployed to computationally model trust in both HRC and MAS applications performed in simulation. On the other hand, ML

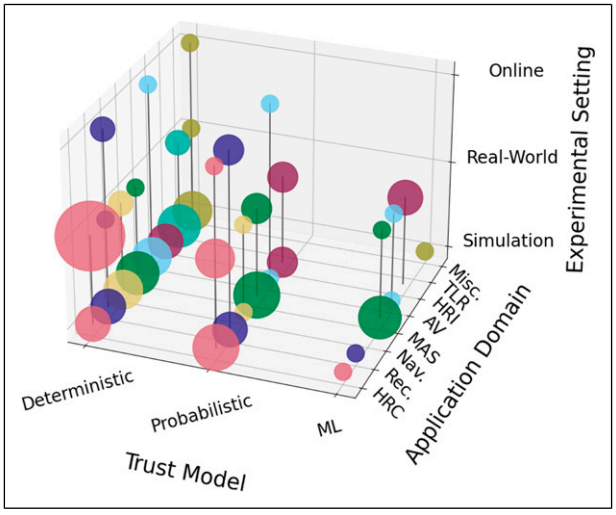

Figure 9. Scatterplot of the selected works according to Computational Trust technique, application domain, and experimental setting. A bigger dot represents a higher percentage of works sharing the same triplet of features. For better visualization purposes, dots with the same color have the same application domain. Due to their limited amount, the two works that used a model based on Game Theory were omitted for better visualization. Regarding the abbreviations, ML = Machine Learning, HRC = Human–Robot Collaboration, Rec. = Reconnaissance, Nav. = Navigation, MAS = Multi-Agent Systems, AV = Autonomous Vehicles, HRI = Human–Robot Interaction, TLR = Telerobotics, Misc. = Miscellanea.

models are primarily applied to AV experiments in simulators and HRI trials under laboratory conditions.

The most common point in the 3-dimensional space under consideration is the intersection of Deterministic models, real-world experiments, and HRC applications. It indicates a clear preference among scholars working in the field of CT. Finally, another insight from Figure 9 is that the field of Autonomous Vehicles still predominantly uses simulated environments, with a good distribution of all classes of methods.

8.2. Comparative analysis of trust modeling approaches

In this review, we have categorized CT models as NT or AT based on the directionality of the trust relationship. Each of these paradigms reflects a distinct epistemological stance on how trust is conceptualized and operationalized within robotic systems. NT models focus on the robot's estimation of the human's trust in it, typically inferred from behavioral cues, performance feedback, or self-reported measures. In contrast, AT models endow the robot with the capacity to evaluate the trustworthiness of its human partner, often to modulate its own behavior in collaborative or safety-critical contexts. While the majority of studies adopt a unidirectional perspective, some works have begun to explore scenarios in which both entities are engaged in a trust relationship. These bidirectional trust (BT) approaches attempt to capture the mutual and dynamic nature of trust, modeling both human-to-robot and robot-to-human trust simultaneously. Figure 10 shows the distribution of the papers selected for this survey across the three categories. The proportion of AT works is comparable to that of NT works. This result highlights the value of having conducted a systematic review that not only investigated computational models of trust, but also contributed to the conceptual framing of AT as a domain that now stands alongside NT in the broader discourse.

NT models are the most prevalent in the literature, appearing in approximately 50% of the surveyed papers. They are especially used in domains such as collaborative manipulation, autonomous driving, and reconnaissance. These models often rely on probabilistic frameworks such as Bayesian networks, dynamic Bayesian networks, or POMDPs, which allow the robot to update its belief about human trust over time based on observed interactions (Chen et al., 2018; Guo et al., 2021; Xu and Dudek, 2015b). Their popularity stems from their alignment with user-centered design principles and the relative ease of validation through user studies. However, NT models are inherently reactive: they estimate trust post hoc and may struggle to anticipate or mitigate trust breakdowns in real time.

AT models, while less common, offer a complementary perspective. Here, the robot actively assesses the human's reliability, predictability, or intent, often using non-linear combinations of performance metrics, fuzzy logic, or RL (Alhaji et al., 2021; Lang et al., 2023b; Rjoub et al., 2023). These models are particularly valuable in scenarios where the robot must make autonomous decisions about whether to follow, assist, or override human input. Their proactive nature enables adaptive behavior that can enhance safety and task efficiency. Interestingly, ML techniques appear

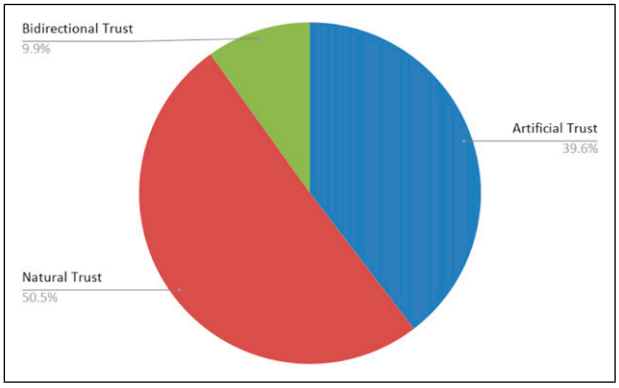

Figure 10. Distribution of the selected works between Natural Trust, Artificial Trust, and Bidirectional Trust models.

more frequently in AT models than in NT models, suggesting a growing reliance on data-driven approaches for trust estimation in autonomous systems. However, AT models face challenges in interpretability and generalizability, especially when deployed in unstructured environments or with diverse user populations.

This survey discovered only 10 papers that model BT, a more recent and ambitious direction in CT research. By simultaneously modeling both AT and NT, these approaches aim to support mixed-initiative interaction, dynamic role arbitration, and mutual adaptation (Rahman et al., 2016a; Xu and Song, 2021; Zhou et al., 2021). They are particularly well suited for multi-agent teaming and teleoperation, where trust must be continuously negotiated between agents. An interesting observation is that all bidirectional models identified in this review rely exclusively on deterministic methods. This co-occurrence may reflect the complexity of modeling mutual trust which might be hindering widespread adoption, but it also highlights a potential limitation in adaptability and expressiveness.

While NT models dominate current research due to their methodological maturity and alignment with human-centered evaluation, AT and bidirectional models offer critical capabilities for proactive and adaptive robot behavior. The choice among these approaches should be guided by the interaction context, the degree of autonomy required, and the desired balance between interpretability and adaptability. As the field progresses, hybrid models that allow bidirectional trust may offer the most robust and flexible solutions for trust-aware robotics.

8.3. Trust factors for AT design

Recent decades of research in human-robot trust have primarily focused on how people place trust in robots, something we have been referring to as NT. For a detailed analysis of the various factors, both human and beyond, that influence trust in robots, we refer to existing frameworks and reviews on the matter (Hancock et al., 2011, 2021; Hoff and Bashir, 2015). Having introduced AT, we now explore trust-related factors that researchers have identified as important for shaping a robot's trust in other agents. These are summarized in Figure 11. AT can follow two main pathways: modeling a robot's trust in a human or in another robot. Among the 50 AT studies reviewed (including both BT and pure AT), 40 focus on robots trusting humans, while 12 address trust between robots (two studies (Ali et al., 2022; Carneiro et al., 2019) investigate both pathways within the same research).

In the subset focused on robots trusting humans, a common approach is to model trust as a function of the human operator's task performance. Seventeen studies, nearly half of this group, use performance as an input to their CT models (Mahani and Wang, 2017; Maithani et al., 2019; Rahman, 2019a, 2019b; Rahman et al., 2016a, 2016b; Wang et al., 2014, 2015, 2022, 2023a; Saeidi et al., 2016, 2017a; Saeidi and Wang, 2019; Sapienza and Falcone, 2023; Scherf et al., 2022; Xu and Song., 2021; Zhou et al., 2021). This is consistent with the prevalence of HRC scenarios, where operator performance is closely tied to task efficiency. A smaller number of studies incorporate human behavioral cues (Hannum et al., 2020; Scherf et al., 2022). These variables are typically collected after an initial interaction with the user, without relying on prior trust history. In applications where trust can be built over time, HRI practitioners may consider modeling these cues for AT towards humans.

Other variables are predictive, aiming to anticipate future human behavior. Several studies refer to predictability, defined as the likelihood of the human being in a particular state (Alhaji et al., 2021; Zhou et al., 2021). This concept is

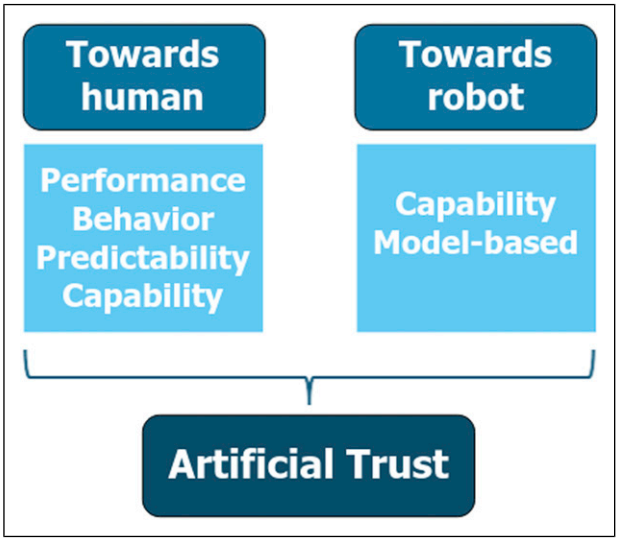

Figure 11. Trust factors in Artificial Trust, divided by trust towards the human and another robot.

interpreted in various ways, such as detecting the user's presence in the workspace (Almohamade et al., 2021) or using a HMM to infer user states (Khattar and Eskandarian, 2022). Capability is another predictive factor, representing the expected ability of the user to complete a task successfully (Ali et al., 2022; Saeidi and Wang, 2019). These variables can be modeled before interaction begins, making them suitable for scenarios where early performance in collaborative tasks is critical.

For AT towards robots, identifying consistent input patterns for CT models is more challenging due to the smaller number of studies, roughly four times fewer than those focused on trust towards humans. This suggests that robot-robot trust research is still in its early stages. Since the inputs now relate to robotic behavior, researchers often use variables specific to the trustee robot's decision-making system (Rishwaraj et al., 2017; Rishwaraj and Ponnambalam, 2017; Setter et al., 2017), rather than ones referring to abstract concepts. A minor exception includes two studies that explicitly incorporate capability into their model design (Ali et al., 2022; Lin et al., 2023).

8.4. Recommendations for research

In light of the patterns and gaps identified in this review, several directions emerge that could meaningfully advance the field of CT. First, there is a clear need to move beyond siloed methodologies and embrace a more integrative modeling approach. Hybrid models combining the interpretability of deterministic frameworks, the uncertainty handling of probabilistic reasoning and the adaptability of ML offer a promising path forward. Such combinations could better reflect the complexity of trust as it unfolds in real-world, dynamic environments.

ML, while increasingly central to AI research, remains under-exploited in trust modeling. Expanding its use could unlock new capabilities, especially in data-rich or real-time applications. To support broader adoption, future work should prioritize transparency and reproducibility, including the open sharing of datasets, model code, and evaluation pipelines.

Some modeling paradigms, such as Game Theory and biologically inspired architectures, have received limited attention despite their potential to capture strategic or embodied aspects of trust. Revisiting these approaches, especially in multi-agent and physically grounded contexts, could yield valuable insights and diversify the methodological toolkit available to researchers.

Large Language Models also present a promising opportunity. Their ability to process and generate human-like language makes them well suited for modeling trust in communication-heavy contexts, such as dialogue systems or collaborative agents. However, it is important to avoid treating language models as complete cognitive systems. While they can simulate aspects of trust-related reasoning, they do not possess grounded representations of belief, intention, or emotion (Jokinen, 2024). Future research should explore how language models can be integrated into broader trust architectures without substituting the full cognitive process.

Equally important is the development of shared resources and evaluation standards. The current lack of standardized datasets, metrics, and experimental protocols makes it difficult to compare models or replicate findings. Establishing common benchmarks would not only improve reproducibility but also foster more cumulative progress across the field.

Another area that deserves greater emphasis is the validation of CT models in real-world settings. As shown in this review, most models are tested in simulation, which, while useful for early-stage development, often fails to capture the complexity of human–robot interaction. Real-world experiments, particularly in domains like healthcare, autonomous driving, and collaborative manufacturing, are essential to ensure ecological validity and uncover practical challenges in deployment.

The review also reveals an uneven distribution of CT research across application domains. While HRC and MAS are well represented, areas such as autonomous vehicles, telerobotics, and smart environments remain underexplored. These domains present unique trust challenges and could benefit from tailored CT frameworks that reflect their specific constraints and user expectations.

Robot morphology and embodiment also play a significant role in shaping trust dynamics. Humanoids and manipulators are more frequently used in real-world experiments, while aerial and mobile robots are often confined to simulation. This observation suggests that physical form might influence the nature of trust interactions. Future work should investigate how embodiment affects trust modeling and whether CT architectures need to be adapted based on the robot's capabilities and appearance.

Finally, trust is not only a computational construct but also a deeply human one. Insights from psychology, neuroscience, and behavioral economics can enrich CT models by grounding them in empirically validated theories of human behavior. Interdisciplinary collaboration will be essential to develop models that are not only technically sound but also aligned with how trust operates in real-world human contexts.

8.5. Ethical considerations

The field of AT focuses on enabling robots, or artificial agents in general, to evaluate the trustworthiness of other agents, whether artificial or natural. The papers discussed in this systematic review propose various ideas and perspectives on how to model trust and how the latter can be quantified and reasoned upon. Fundamentally, they share the objective of enabling intelligent machines to evaluate

the behavior of others, which introduces a constellation of ethical challenges.

The emerging field of "intelligent disobedience" (Briggs and Scheutz, 2017) seeks to blend AI and ethics to create smart robots capable of discerning when it is unsafe or unethical to comply with instructions. The idea is to incorporate ethical frameworks into the agent's reasoning process in order to make it proactive, transparent and verifiable (Bremner et al., 2019). For example, should robots be allowed to say "no" to a human when they do not trust them? Should they blatantly refuse to obey commands, or should they limit themselves to strongly advising against a certain course of action? If an autonomous vehicle detects signs of impaired driving from its driver, should it be allowed to cut off manual control, potentially saving lives, or should it prioritize the driver's freedom? The domains of CT and intelligent disobedience could mutually benefit each other, as trust is a critical component in evaluating instructions provided by the other party.

Moreover, it is important to remember that AI models are subject to potential biases in their design or training. If we plan to allow robots to judge people's actions in order to make decisions, it is paramount that we ensure these systems are as transparent and unbiased as possible. This issue should be proactively addressed through fairness audits and diverse participant sampling. Privacy is another critical concern: robots equipped with the ability to evaluate trustworthiness may need to collect and process significant amounts of personal data. Ensuring that these data are handled in a way that respects individuals' privacy rights and complies with relevant regulations is vital. Finally, there is the question of consent. People interacting with trust-aware robots should be aware that their actions are being monitored and evaluated, especially when trust estimation is implicit or continuous. Designers must consider how to make trust assessments explainable and contestable, especially in cases where trust influences autonomy, access, or intervention. Informed consent is a fundamental principle that should be upheld to respect individuals' autonomy.

The integration of these considerations into the design, development, and deployment of autonomous trust systems is crucial. It is not enough to create technically proficient systems; they must also be ethically sound and aligned with societal values to support responsible and trustworthy human–robot collaboration. To that end, we recommend that future research on CT actively adopt ethical design frameworks such as Value-Sensitive Design (VSD) (Friedman et al., 2013) and Design-for-Values (DfV) (Van Den Hoven et al., 2015). These approaches offer concrete methodologies for embedding societal values, including transparency, fairness, privacy, and accountability, into the core of system development.

For example, VSD encourages early engagement with stakeholders to identify value tensions, which is particularly relevant for trust models that may influence behavior or decision-making in high-stakes contexts. A CT system that infers human reliability from task errors, for instance, should involve users in defining what constitutes a "meaningful" error and what information can be collected ethically.

DfV, by contrast, emphasizes aligning technical systems with publicly endorsed values or institutional norms. A CT system deployed in healthcare or public service contexts might draw from DfV principles by integrating constraints that reflect legal and cultural norms around dignity, consent, and non-discrimination. This coupling would ensure that trust evaluations are not only accurate but also socially acceptable and regulatable.

We therefore call for a shift in CT research: from a predominantly performance-focused paradigm to one that balances technical achievement with ethical accountability already from the design stage. This shift includes incorporating user perspectives, making trust models interpretable, and ensuring that design choices support users' rights and dignity.

9. Conclusion

In this paper, a systematic review of computational models of trust in robotics was performed. Through thorough analysis and selection criteria, 101 works were selected as representatives of this research field, ranging from 2013 to 2023. The increasing number of papers over time highlights the growing relevance of this topic within the robotics community.

After selection, these entries were categorized according to the type of Computational Trust (CT) model used, the application domain, the robotic platform employed and the nature of the experimental validation. Insights regarding the results, metrics, analyses, and evaluations were also provided. Among the main observations, it was noted that the majority of CT models are deterministic, followed by probabilistic models and machine learning models. Additionally, most of the models are validated in a simulated environment, with only a third of the selected works involving real-world validation.

After analyzing the individual features, these were cross-analyzed to identify more in-depth trends. It is worth mentioning that deterministic models are applied equally across almost every identified application domain, establishing a solid corpus of methods to be employed in robotics applications. Finally, discussions about the ethical implications of CT are encouraged, including the ties of the topic with intelligent disobedience.

Acknowledgments

We would like to express our sincere gratitude to Carolina Centeio Jorge for her invaluable early feedback on this paper. Her insightful comments and suggestions greatly contributed to the

enhancement of our work and its alignment with the Computational Trust community.

ORCID iDs

Samuele Vinanzi https://orcid.org/0000-0003-0241-9983
Marta Romeo https://orcid.org/0000-0003-4438-0255
Angelo Cangelosi https://orcid.org/0000-0002-4709-2243
Francesco Semeraro https://orcid.org/0000-0002-8812-0968

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Samuele Vinanzi's work was partially supported by Sheffield Hallam University's Early Career Research and Innovation Fellowship. This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Materiel Command, USA. Funder awards no. FA9550-19-1-7002 and FA8655-24-1-7046. For the purpose of open access, the authors has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission. Marta Romeo's work was supported by the UKRI Node on Trust (EP/V026682/1) https://trust.tas.ac.uk. Angelo Cangelosi's work was partially supported by the Horizon/UKRI PRIMI project (Ref. 101120727). This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Materiel Command, USA. Funder award no. FA8655-24-1-7047. Francesco Semeraro's work was supported by the UKRI DTP CASE-conversion "Human-Robot Collaboration for Flexible Manufacturing" (Ref. 2480772), sponsored by UKRI Engineering and Physical Sciences Research Council and BAE Systems plc., and the Horizon project MUSAE (Ref. 101070421).

Declaration of conflicting interests

The authors have no relevant financial or non-financial interests to disclose.

References

Ab Aziz A, Hussain WAA, Ahmad F, et al. (2017) An agent model for analysis of trust dynamics in short-term human-robot interaction. In: *Conference on Computational Intelligence in Information Systems (CIIS)*. CHAM: Springer International Publishing Ag, Vol. 532, 81–93. https://doi.org/10.1007/978-3-319-48517-1_8

Abdulhussain WA and Aziz AA (2022) A computational model of human-robot collaboration trust and its application in simulated operative domain. *Defence S&T Technical Bulletin* 15(2): 239–257.

Akash K, Reid T and Jain N (2019) Improving human-machine collaboration through transparency-based feedback - part ii: control design and synthesis. In: *2nd International-Federation-of-Auinternational-Federation-of-Automatic-Cber-Physical and Human-Systems (CPHS)*. Elsevier Science Bv, Vol. 51, pp. 322–328. https://doi.org/10.1016/j.ifacol.2019.01.026

Alhaji B, Prilla M and Rausch A (2021) Trust, but verify: autonomous robot trust modeling in human-robot collaboration. In: *HAI 2021 - Proceedings of the 9th International User Modeling, Adaptation and Personalization Human-Agent Interaction*, pp. 402–406. https://doi.org/10.1145/3472307.3484672

Ali A, Azevedo-Sa H, Tilbury DM, et al. (2022) Heterogeneous human-robot task allocation based on artificial trust. *Scientific Reports* 12(1): 15. https://doi.org/10.1038/s41598-022-19140-5

Almohamade SS, Clark JA and Law J (2021) Behaviour-based biometrics for continuous user authentication to industrial collaborative robots. In: *Lecture Notes in Computer Science, Volume 12596 LNCS*. Springer International Publishing, pp. 185–197. https://doi.org/10.1007/978-3-030-69255-1_12

Aydoğan R, Sharpanskykh A and Lo J (2015) A trust-based situation awareness model. In: *Lecture Notes in Computer Science*. Springer International Publishing, Vol. 8953, 19–34. https://doi.org/10.1007/978-3-319-17130-2_2

Azevedo-Sa H, Yang XJ, Robert LP, et al. (2021) A unified bi-directional model for natural and artificial trust in human-robot collaboration. *IEEE Robotics and Automation Letters* 6(3): 5913–5920. https://doi.org/10.1109/lra.2021.3088082

Aziz AA and Abdulhussain WA (2022) Computational analysis of a human–robot working alliance trust in robot-based therapy. In: *Lecture Notes in Networks and Systems*. Springer, Vol. 238, 431–442. https://doi.org/10.1007/978-981-16-2641-8_41

Bainbridge WA, Hart JW, Kim ES, et al. (2011) The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics* 3: 41–52. https://doi.org/10.1007/s12369-010-0082-7

Bhat S, Lyons JB, Shi C, et al. (2022) Clustering trust dynamics in a human-robot sequential decision-making task. *IEEE Robotics and Automation Letters* 7(4): 8815–8822. https://doi.org/10.1109/lra.2022.3188902

Boer M, PPv M and Vreeswijk G (2013) Supporting intelligence analysts with a trust-based question-answering system. In: *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Vol. 3, 183–186. https://doi.org/10.1109/WI-IAT.2013.179

Brandizzi N, Centeio Jorge C, Cipollone R, et al. (2023) Multittrust: 2nd workshop on multidisciplinary perspectives on human-ai team trust. In: *Proceedings of the 11th International Conference on Human-Agent Interaction*, pp. 496–497.

Bremner P, Dennis LA, Fisher M, et al. (2019) On proactive, transparent, and verifiable ethical reasoning for robots. *Proceedings of the IEEE* 107(3): 541–561. https://doi.org/10.1109/jproc.2019.2898267

Briggs G and Scheutz M (2017) The case for robot disobedience. *Scientific American* 316(1): 44–47. https://doi.org/10.1038/scientificamerican0117-44

Carbo J and Molina JM (2023) Trust model of privacy-concerned, emotionally aware agents in a cooperative logistics problem.

Applied Sciences 13(15): 8681. https://doi.org/10.3390/app13158681

Carneiro LR, Delgado CADM and Jcpd S (2019) Social analysis of game agents: how trust and reputation can improve player experience. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 485–490. https://doi.org/10.1109/BRACIS.2019.00091

Centeio Jorge C and Ulfert-Blank AS (2023) Multittrust-multidisciplinary perspectives on human-ai team trust. In: *CEUR Workshop Proceedings*, Vol. 3456.

Chen M, Nikolaidis S, Soh H, et al. (2018) Planning with trust for human-robot collaboration. In: *13th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Assoc Computing Machinery, pp. 307–315. https://doi.org/10.1145/3171221.3171264

Chen M, Nikolaidis S, Soh H, et al. (2020) Trust-aware decision making for human-robot collaboration: model learning and planning. *Acm Transactions on Human-Robot Interaction* 9(2): 23. https://doi.org/10.1145/3359616

Cheng M, Zhang J, Nazarian S, et al. (2021) Trust-aware control for intelligent transportation systems. In: *2021 IEEE Intelligent Vehicles Symposium (IV)*, pp. 377–384. https://doi.org/10.1109/IV48863.2021.9576045

Cheng Z, Kong D, Sun L, et al. (2023) Investigation, analysis, and quantification of drivers' trust level toward autonomous vehicles in human-machine mixed traffic flow. In: *CICTP 2023: Innovation-Empowered Technology for Sustainable, Intelligent, Decarbonized, and Connected Transportation - Proceedings of the 23rd COTA International Conference of Transportation Professionals*, pp. 1839–1850. https://doi.org/10.1061/9780784484869.174

Das T and Teng BS (2004) The risk-based view of trust: a conceptual framework. *Journal of Business and Psychology* 19(1): 85–116. https://doi.org/10.1023/b:jobu.0000040274.23551.1b

Dorbala VS, Srinivasan A and Bera A (2021) Can a robot trust you?: a drl-based approach to trust-driven human-guided navigation. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3538–3545. https://doi.org/10.1109/ICRA48506.2021.9561983

Dubey AD and Kumar BA (2019) A novel cognitive approach for measuring the trust in robots. *Journal of Information Technology Research* 12(3): 60–73. https://doi.org/10.4018/jitr.2019070104

Floyd MW, Drinkwater M and Aha DW (2014) How much do you trust me? Learning a case-based model of inverse trust. In: *Lecture Notes in Computer Science*. Springer International Publishing, Vol. 8765, 125–139. https://doi.org/10.1007/978-3-319-11209-1_10

Floyd MW, Drinkwater M and Aha DW (2015) Improving trust-guided behavior adaptation using operator feedback. In: *Lecture Notes in Computer Science*. Springer International Publishing, Vol. 9343, 134–148. https://doi.org/10.1007/978-3-319-24586-7_10

Friedman B, Kahn PH, Borning A, et al. (2013) Value sensitive design and information systems. In: *Early Engagement and New Technologies: Opening up the Laboratory*, 55–95.

Guo YH and Yang XJ (2021) Modeling and predicting trust dynamics in human-robot teaming: a bayesian inference approach. *International Journal of Social Robotics* 13(8): 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

Guo YH, Shi C and Yang XJ (2021) Reverse psychology in trust-aware human-robot interaction. *IEEE Robotics and Automation Letters* 6(3): 4851–4858. https://doi.org/10.1109/lra.2021.3067626

Guo Y, Yang XJ and Shi C (2023a) Reward shaping for building trustworthy robots in sequential human-robot interaction. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7999–8005. https://doi.org/10.1109/IROS55552.2023.10341904

Guo Y, Yang XJ and Shi C (2023b) Tip: a trust inference and propagation model in multi-human multi-robot teams. In: *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 639–643. https://doi.org/10.1145/3568294.3580164

Hale MT, Setter T and Fregene K (2019) Trust-driven privacy in human-robot interactions. In: *2019 American Control Conference (ACC)*, pp. 5234–5239. https://doi.org/10.23919/ACC.2019.8815004

Hancock PA, Billings DR, Schaefer KE, et al. (2011) A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53(5): 517–527. https://doi.org/10.1177/0018720811417254

Hancock PA, Kessler TT, Kaplan AD, et al. (2021) Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human Factors* 63: 1196–1229. https://doi.org/10.1177/0018720820922080/ASSET/IMAGES/LARGE/10.1177_0018720820922080-FIG7.JPEG

Hannum C, Li R and Wang W (2020) Trust or not? A computational robot-trusting-human model for human-robot collaborative tasks *2020 IEEE International Conference on Big Data (Big Data)*, pp. 5689–5691. https://doi.org/10.1109/BigData50022.2020.9378119

Hoff KA and Bashir M (2015) Trust in automation: integrating empirical evidence on factors that influence trust. *Human Factors* 57(3): 407–434. https://doi.org/10.1177/0018720814547570

Hoogendoorn M, Jaffry SW, Van Maanen PP, et al. (2014) Design and validation of a relative trust model. *Knowledge-Based Systems* 57: 81–94. https://doi.org/10.16/j.knosys.2013.12.012

Hsieh SJ, Wang AR, Madison A, et al. (2022) Adaptive driving assistant model (adam) for advising drivers of autonomous vehicles. *ACM Transactions on Interactive Intelligent Systems* 12(3): 28. https://doi.org/10.1145/3545994

Hu C and Wang JM (2022) Trust-based and individualizable adaptive cruise control using control barrier function approach with prescribed performance. *IEEE Transactions on*

Intelligent Transportation Systems 23(7): 6974–6984. https://doi.org/10.1109/tits.2021.3066154

Hu D, Dang Y and Yue X (2022) The effect of trust-based management strategy on performance of human-machine collaborative team: a dynamic computational model. *Procedia Computer Science* 221: 710–717. https://doi.org/10.1016/j.procs.2023.08.042

Jokinen K (2024) The need for grounding in LLM-based dialogue systems. In: Dong T, Hinrichs E, Han Z, et al. (eds) *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (Neusymbridge) @ LREC-COLING-2024*. ELRA and ICCL, pp. 45–52.

Jones GR and George JM (1998) The experience and evolution of trust: implications for cooperation and teamwork. *Academy of Management Review* 23(3): 531–546. https://doi.org/10.2307/259293

Jorge CC, Tielman ML and Jonker CM (2022) Artificial trust as a tool in human-ai teams. In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 1155–1157.

Kang A (2018) Collaborative filtering algorithm based on trust and information entropy. In: *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Vol. 3, 262–266. https://doi.org/10.1109/ICIIBMS.2018.8549962

Khattar V and Eskandarian A (2022) Stochastic reachable set threat assessment for autonomous vehicles using trust-based driver behavior prediction. *SAE International Journal of Connected and Automated Vehicles* 6(2): 123–137. https://doi.org/10.4271/12-06-02-0008

Khavas ZR, Ahmadzadeh SR and Robinette P (2020) Modeling trust in human-robot interaction: a survey. In: Wagner AR, Feil-Seifer D, Haring KS, et al. (eds) *Social Robotics*. Springer International Publishing, pp. 529–541.

Kirtay M, Hafner VV, Asada M, et al. (2023) Trust in robot-robot scaffolding. *IEEE Transactions on Cognitive and Developmental Systems* 15: 1. https://doi.org/10.1109/TCDS.2023.3235974

Kok BC and Soh H (2020) Trust in robots: challenges and opportunities. *Current Robotics Reports 2020* 1(4 1): 297–309. https://doi.org/10.1007/S43154-020-00029-Y

Kousar Nikhath A, Sandhya N, Khanum Pathan S, et al. (2023) Detection of suspicious human activities from surveillance camera using neural networks. In: Bhateja V, Carroll F, Tavares JMRS, et al. (eds) *Intelligent Data Engineering and Analytics*. Springer Nature Singapore, pp. 255–263.

Kraus M, Wagner N and Minker W (2021) Modelling and predicting trust for developing proactive dialogue strategies in mixed-initiative interaction. In: *ICMI 2021 - Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 131–140. https://doi.org/10.1145/3462244.3479906

Kumar B and Dubey AD (2017) Evaluation of trust in robots: a cognitive approach. In: *2017 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6. https://doi.org/10.1109/ICCCI.2017.8117701

Lang X, Feng Z, Yang X, et al. (2023a) Hmmcf:a human-computer collaboration algorithm based on multimodal intention of reverse active fusion. *International Journal of Human-Computer Studies* 169: 102916. https://doi.org/10.1016/j.ijhcs.2022.102916

Lang XJ, Feng ZQ, Yang XH, et al. (2023b) Hmmcf:a human-computer collaboration algorithm based on multimodal intention of reverse active fusion. *International Journal of Human-Computer Studies* 169: 14. https://doi.org/10.1016/j.ijhcs.2022.102916

Lee JD and See KA (2004) Trust in automation: designing for appropriate reliance. *Human Factors* 46(1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee JJ, Knox WB, Wormwood JB, et al. (2013) Computationally modeling interpersonal trust. *Frontiers in Psychology* 4(-DEC): 893. https://doi.org/10.3389/fpsyg.2013.00893

Lee SW, Hussain S, Issa GF, et al. (2021) Multi-dimensional trust quantification by artificial agents through evidential fuzzy multi-criteria decision making. *IEEE Access* 9: 159399–159412. https://doi.org/10.1109/access.2021.3131521

Li M, Erickson IM, Cross EV, et al. (2023) It's not only what you say, but also how you say it: machine learning approach to estimate trust from conversation. *Human Factors* 66: 1724–1741. https://doi.org/10.1177/00187208231166624

Lin C, Zhang H, Ou L, et al. (2023) Adaptive trust model for multi-agent teaming based on reinforcement-learning-based fusion. *IEEE Transactions on Emerging Topics in Computational Intelligence* 8: 1–11. https://doi.org/10.1109/TETCI.2023.3319253

Losey DP and Sadigh D (2019) Robots that take advantage of human trust. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7001–7008. https://doi.org/10.1109/IROS40897.2019.8968564

Ma W, Chang YC, Wang YK, et al. (2022) Human-autonomous teaming framework based on trust modelling. In: *Lecture Notes in Computer Science, Volume 13728 LNAI*. Springer International Publishing, pp. 707–718. https://doi.org/10.1007/978-3-031-22695-3_49

Mahani MF and Wang Y (2017) Runtime verification of trust-based symbolic robot motion planning with human-in-the-loop. In: *9th ASME Annual Dynamic Systems and Control Conference*. Amer Soc Mechanical Engineers.

Mahani MF, Jiang LS and Wang Y (2021) A bayesian trust inference model for human-multi-robot teams. *International Journal of Social Robotics* 13(8): 1951–1965. https://doi.org/10.1007/s12369-020-00705-1

Maithani H, Corrales-Ramon JA and Mezouar Y (2019) Trust-based variable impedance control for cooperative physical human-robot interaction. In: *IEEE International Conference on Mechatronics (ICM)*. IEEE, pp. 706–711.

Mangalindan DH, Rovira E and Srivastava V (2023) On trust-aware assistance-seeking in human-supervised autonomy. In: *2023 American Control Conference (ACC)*, pp. 3901–3906. https://doi.org/10.23919/ACC55779.2023.10156103

Mansoor F, Rohail M and Jaffry SW (2013) Empirical validation of model for human decision making. In: *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Vol. 3, 187–190. https://doi.org/10.1109/WI-IAT.2013.181

Mayer RC, Davis JH and Schoorman FD (1995) An integrative model of organizational trust. *Academy of Management Review* 20(3): 709–734. https://doi.org/10.2307/258792

Morgner C (2013) Trust and confidence: history, theory and sociopolitical implications. *Human Studies* 36(4): 509–532. https://doi.org/10.1007/s10746-013-9281-1

Nam C, Walker P, Lewis M, et al. (2017) Predicting trust in human control of swarms via inverse reinforcement learning. In: *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 528–533.

Nam C, Walker P, Li H, et al. (2020) Models of trust in human control of swarms with varied levels of autonomy. *IEEE Transactions on Human-Machine Systems* 50(3): 194–204. https://doi.org/10.1109/thms.2019.2896845

Page MJ, McKenzie JE, Bossuyt PM, et al. (2021) The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372: n71.

Pang YJ, Huang C and Liu R (2021) Synthesized trust learning from limited human feedback for human-load-reduced multi-robot deployments. In: *30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 778–783. https://doi.org/10.1109/ro-man50785.2021.9515509

Patacchiola M and Cangelosi A (2016) A developmental bayesian model of trust in artificial cognitive systems *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 117–123. https://doi.org/10.1109/DEVLRN.2016.7846801

Patacchiola M and Cangelosi A (2022) A developmental cognitive architecture for trust and theory of mind in humanoid robots. *IEEE Transactions on Cybernetics* 52(3): 1947–1959. https://doi.org/10.1109/TCYB.2020.3002892

Ponnambalam SG, Janardhanan MN and Rishwaraj G (2021) Trust-based decision-making framework for multiagent system. *Soft Computing* 25(11): 7559–7575. https://doi.org/10.1007/s00500-021-05715-3

Rabby MKM, Khan MA, Karimoddini A, et al. (2020) Modeling of trust within a human-robot collaboration framework. In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 4267–4272. https://doi.org/10.1109/SMC42975.2020.9283228

Rahman SMM (2019a) Cognitive cyber-physical system (c-cps) for human-robot collaborative manufacturing. In: *2019 14th Annual Conference System of Systems Engineering (SoSE)*, pp. 125–130. https://doi.org/10.1109/SYSOSE.2019.8753835

Rahman SMM (2019b) Mixed-initiative collaboration between a humanoid robot and a virtual human through a common platform for a real-world common task: evaluation and benchmarking. *Journal of Ambient Intelligence and Smart Environments* 11(5): 429–452. https://doi.org/10.3233/ais-190535

Rahman SMM (2019c) Trustworthy power assistance in object manipulation with a power assist robotic system. In: *IEEE SoutheastCon Conference*. IEEE.

Rahman SMM, Sadrfaridpour B, Wang Y, et al. (2016a) Trust-based optimal subtask allocation and model predictive control for human-robot collaborative assembly in manufacturing. In: *8th ASME Annual Dynamic Systems and Control Conference (DSCC 2015)*. Amer Soc Mechanical Engineers.

Rahman SMM, Wang Y, Walker ID, et al. (2016b) Trust-based compliant robot-human handovers of payloads in collaborative assembly in flexible manufacturing *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 355–360. https://doi.org/10.1109/COASE.2016.7743428

Ranasinghe A, Dasgupta P, Althoefer K, et al. (2015) Identification of haptic based guiding using hard reins. *PLoS One* 10(7): 22. https://doi.org/10.1371/journal.pone.0132020

Razin YS and Feigh KM (2021) Committing to interdependence: implications from game theory for human-robot trust. *Paladyn* 12(1): 481–502. https://doi.org/10.1515/pjbr-2021-0031

Rishwaraj G and Ponnambalam SG (2017) Integrated trust based control system for multirobot systems: development and experimentation in real environment. *Expert Systems with Applications* 86: 177–189. https://doi.org/10.1016/j.eswa.2017.05.074

Rishwaraj G, Ponnambalam SG and Loo CK (2017) Trust evaluation in a multi-robotics system through direct learning. In: *Lecture Notes in Electrical Engineering*. Springer, Vol. 398, 407–417. https://doi.org/10.1007/978-981-10-1721-6_44

Rjoub G, Bentahar J and Wahab OA (2023) Explainable trust-aware selection of autonomous vehicles using lime for one-shot federated learning. In: *2023 International Wireless Communications and Mobile Computing*. IWCMC 2023, pp. 524–529. https://doi.org/10.1109/IWCMC58020.2023.10182876

Rutard F, Sigaud O and Chetouani M (2020) Tirl: enriching actor-critic rl with non-expert human teachers and a trust model. In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 604–611. https://doi.org/10.1109/RO-MAN47096.2020.9223530

Sadrfaridpour B and Wang Y (2018) Collaborative assembly in hybrid manufacturing cells: an integrated framework for human–robot interaction. *IEEE Transactions on Automation Science and Engineering* 15(3): 1178–1192. https://doi.org/10.1109/TASE.2017.2748386

Sadrfaridpour B, Saeidi H and Wang Y (2016) An integrated framework for human-robot collaborative assembly in hybrid manufacturing cells. In: *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 462–467. https://doi.org/10.1109/COASE.2016.7743441

Sadrfaridpour B, Mahani MF, Liao ZR, et al. (2018) Trust-based impedance control strategy for human-robot cooperative manipulation. In: *11th Annual Dynamic Systems and Control Conference (DSCC 2018)*. Amer Soc Mechanical Engineers.

Saeidi H and Wang Y (2015) Trust and self-confidence based autonomy allocation for robotic systems. In: *54th IEEE Conference on Decision and Control (CDC)*. IEEE, pp. 6052–6057.

Saeidi H and Wang Y (2019) Incorporating trust and self-confidence analysis in the guidance and control of (semi) autonomous mobile robotic systems. *IEEE Robotics and Automation Letters* 4(2): 239–246. https://doi.org/10.1109/lra.2018.2886406

Saeidi H, McLane F, Sadrfaidpour B, et al. (2016) Trust-based mixed-initiative teleoperation of mobile robots. In: *American Control Conference (ACC)*. IEEE, pp. 6177–6182.

Saeidi H, Mikulski DG and Wang Y (2017a) Trust-based leader selection for bilateral haptic teleoperation of multi-robot systems. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6575–6581. https://doi.org/10.1109/IROS.2017.8206569

Saeidi H, Wagner JR and Wang Y (2017b) A mixed-initiative haptic teleoperation strategy for mobile robotic systems based on bidirectional computational trust analysis. *IEEE Transactions on Robotics* 33(6): 1500–1507. https://doi.org/10.1109/TRO.2017.2718549

Sanders NE and Nam CS (2021) Applied quantitative models of trust in human-robot interaction. *Trust in Human-Robot Interaction* 2021: 449–476. https://doi.org/10.1016/B978-0-12-819472-0.00019-8

Sapienza A and Falcone R (2023) Exploiting autonomy in a user-iot system collaborative trust model. *International Journal of Parallel, Emergent and Distributed Systems* 38(6): 477–489. https://doi.org/10.1080/17445760.2023.2234166

Schaefer KE, Chen JY, Szalma JL, et al. (2016) A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Human Factors* 58(3): 377–400. https://doi.org/10.1177/0018720816634228

Schäfer A, Esterbauer R and Kubicek B (2024) Trusting robots: a relational trust definition based on human intentionality. *Humanities and Social Sciences Communications* 11(1): 1–12. https://doi.org/10.1057/s41599-024-03897-3

Scherf L, Turan C and Koert D (2022) Learning from unreliable human action advice in interactive reinforcement learning. In: *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pp. 895–902. https://doi.org/10.1109/Humanoids53995.2022.10000078

Semeraro F, Carberry J, Leadbetter J, et al. (2024a) Good things come in threes: the impact of robot responsiveness on workload and trust in multi-user human-robot collaboration. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 2471–2478.

Semeraro F, Romeo M, Cangelosi A, et al. (2024b) Computational trust in robotics: preliminary investigations and evidence. In: *CEUR Workshop Proceedings*. CEUR-WS, Vol. 3825, pp. 176–179.

Setter T, Gasparri A and Egerstedt M (2017) Trust in multi-agent networks: from self-centered to team-oriented. In: *2017 American Control Conference (ACC)*, pp. 997–1002. https://doi.org/10.23919/ACC.2017.7963083

Soh H, Xie YQ, Chen M, et al. (2020) Multi-task trust transfer for human-robot interaction. *International Journal of Robotics Research* 39(2-3): 233–249. https://doi.org/10.1177/0278364919866905

Spencer DA, Wang Y and Humphrey LR (2016) Trust-based human-robot interaction for multi-robot symbolic motion planning *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 1443–1449.

Sun L, Cheng Z, Kong D, et al. (2023) Modeling and analysis of human-machine mixed traffic flow considering the influence of the trust level toward autonomous vehicles. *Simulation Modelling Practice and Theory* 125: 102741. https://doi.org/10.1016/j.simpat.2023.102741

Tielman ML, Meyer-Vitali A, Bailey ME, et al. (2024) Multi-disciplinary perspectives on human-ai team trust (preface). In: *HHAI Workshops*, pp. 164–166.

Tielman ML, Bailey M, Frattolillo F, et al. (2025) Multidisciplinary perspectives on human-ai team trust. *Interaction Studies* 26(2): 164–199. https://doi.org/10.1075/is.24048.tie

Tilloo P, Parron J, Obidat O, et al. (2022) A pomdp-based robot-human trust model for human-robot collaboration. In: *2022 12th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 1009–1014. https://doi.org/10.1109/CYBER55403.2022.9907660

Tjøstheim TA, Johansson B and Balkenius C (2019) A computational model of trust-, pupil-, and motivation dynamics. In: *HAI 2019 - Proceedings of the 7th International Conference on Human-Agent Interaction*, pp. 179–185. https://doi.org/10.1145/3349537.3351896

Van den Hoven J, Vermaas PE and Van de Poel I (2015) Design for values: an introduction. In: *Handbook of Ethics, Values, and Technological Design*. Springer, pp. 1–7.

Van Eck N and Waltman L (2010) Software survey: vosviewer, a computer program for bibliometric mapping. *Scientometrics* 84(2): 523–538. https://doi.org/10.1007/s11192-009-0146-3

Vinanzi S (2021) *Developmental Collaborative Intelligence for Embodied Robotic Agents. Phd Thesis*. The University of Manchester.

Vinanzi S and Cangelosi A (2024) Casper: cognitive architecture for social perception and engagement in robots. *International Journal of Social Robotics* 17: 1–19. https://doi.org/10.1007/s12369-024-01116-2

Vinanzi S, Patacchiola M, Chella A, et al. (2019) Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B* 374(1771): 20180032. https://doi.org/10.1098/rstb.2018.0032

Vinanzi S, Cangelosi A and Goerick C (2021) The collaborative mind: intention reading and trust in human-robot interaction.

iScience 24(2): 102130. https://doi.org/10.1016/j.isci.2021.102130

Wagner AR, Robinette P and Howard A (2018) Modeling the human-robot trust phenomenon: a conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems* 8(4): 24. https://doi.org/10.1145/3152890

Wang Y, Shi Z, Wang C, et al. (2014) Human-robot mutual trust in (semi)autonomous underwater robots. In: *Studies in Computational Intelligence*. Springer, Vol. 554, pp. 115–137. https://doi.org/10.1007/978-3-642-55029-4_6

Wang X, Shi Z, Zhang F, et al. (2015) Dynamic real-time scheduling for human-agent collaboration systems based on mutual trust. *Cyber-Physical Systems* 1(2-4): 76–90. https://doi.org/10.1080/23335777.2015.1056755

Wang Y, Humphrey LR, Liao ZR, et al. (2018) Trust-based multi-robot symbolic motion planning with a human-in-the-loop. *ACM Transactions on Interactive Intelligent Systems* 8(4): 33. https://doi.org/10.1145/3213013

Wang Q, Liu D, Carmichael MG, et al. (2022) Computational model of robot trust in human co-worker for physical human-robot collaboration. *IEEE Robotics and Automation Letters* 7(2): 3146–3153. https://doi.org/10.1109/LRA.2022.3145957

Wang Q, Liu D, Carmichael MG, et al. (2023a) Robot trust and self-confidence based role arbitration method for physical human-robot collaboration. In: *Proceedings - IEEE International Conference on Robotics and Automation*, Vol. 2023, pp. 9896–9902. https://doi.org/10.1109/ICRA48891.2023.10160711

Wang X, Zhang J, Li H, et al. (2023b) A mixed traffic car-following behavior model. *Physica A: Statistical Mechanics and Its Applications* 632: 129299. https://doi.org/10.1016/j.physa.2023.129299

Wang Y, Li F, Zheng H, et al. (2023c) Human trust in robots: a survey on trust models and their controls/robotics applications. *IEEE Open Journal of Control Systems* 3: 58–86. https://doi.org/10.1109/OJCSYS.2023.3345090

Wilson JR, Aung PT and Boucher I (2023) When to help? A multimodal architecture for recognizing when a user needs help from a social robot. In: *Social Robotics: 14Th International Conference, ICSR 2022, Florence, Italy, December 13–16, 2022, Proceedings, Part I*. Springer-Verlag, pp. 253–266. https://doi.org/10.1007/978-3-031-24667-8_23

Wu B, Hu B and Lin H (2017) Toward efficient manufacturing systems: a trust based human robot collaboration. In: *American Control Conference (ACC)*. IEEE, pp. 1536–1541.

Xu A and Dudek G (2015a) Towards efficient collaborations with trust-seeking adaptive robots. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts, HRI'15 Extended Abstracts*. Association for Computing Machinery, pp. 221–222. https://doi.org/10.1145/2701973.2702711

Xu AQ and Dudek G (2015b) Optimo: online probabilistic trust inference model for asymmetric human-robot collaborations. In: *10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Assoc Computing Machinery, pp. 221–228. https://doi.org/10.1145/2696454.2696492

Xu J and Howard A (2020) Would you take advice from a robot? Developing a framework for inferring human-robot trust in time-sensitive scenarios. In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 814–820. https://doi.org/10.1109/RO-MAN47096.2020.9223544

Xu F, Uszkoreit H, Du Y, et al. (2019) Explainable ai: a brief survey on history, research areas, approaches and challenges. In: *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, pp. 563–574.

Xu CT and Song HB (2021) Mixed initiative balance of human-swarm teaming in surveillance via reinforcement learning. In: *IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*. IEEE. https://doi.org/10.1109/dasc52595.2021.9594355

Yan YG, Wang HB, Yu HY, et al. (2022) Machine learning-based surgical state perception and collaborative control for a vascular interventional robot. *IEEE Sensors Journal* 22(7): 7106–7118. https://doi.org/10.1109/jsen.2022.3154921

Zahedi Z, Verma M, Sreedharan S, et al. (2023) Trust-aware planning: modeling trust evolution in iterated human-robot interaction. In: *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 281–289. https://doi.org/10.1145/3568162.3578628

Zheng HF, Liao ZR, Wang Y, et al. (2018) Human-robot trust integrated task allocation and symbolic motion planning for heterogeneous multi-robot systems. In: *11th Annual Dynamic Systems and Control Conference (DSCC 2018)*. Amer Soc Mechanical Engineers.

Zheng H, Smereka JM, Mikulski D, et al. (2023a) Bayesian optimization based trust model for human multi-robot collaborative motion tasks in offroad environments. *International Journal of Social Robotics* 15(7): 1181–1201. https://doi.org/10.1007/s12369-023-01011-2

Zheng T, Bujarbaruah M, Stürz YR, et al. (2023b) Safe human-robot collaborative transportation via trust-driven role adaptation. In: *2023 American Control Conference (ACC)*, pp. 22–27. https://doi.org/10.23919/ACC55779.2023.10156494

Zhou L, Dou Y, Liu H, et al. (2021) Shared control method for coal mine rescue robots. In: *2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, pp. 1–6. https://doi.org/10.1109/CVCI54083.2021.9661251

Zonca J and Sciutti A (2021) Does human-robot trust need reciprocity? In: *RO-MAN 2021 Workshop on Robot Behavior Adaptation to Human Social Norms (TSAR)*.