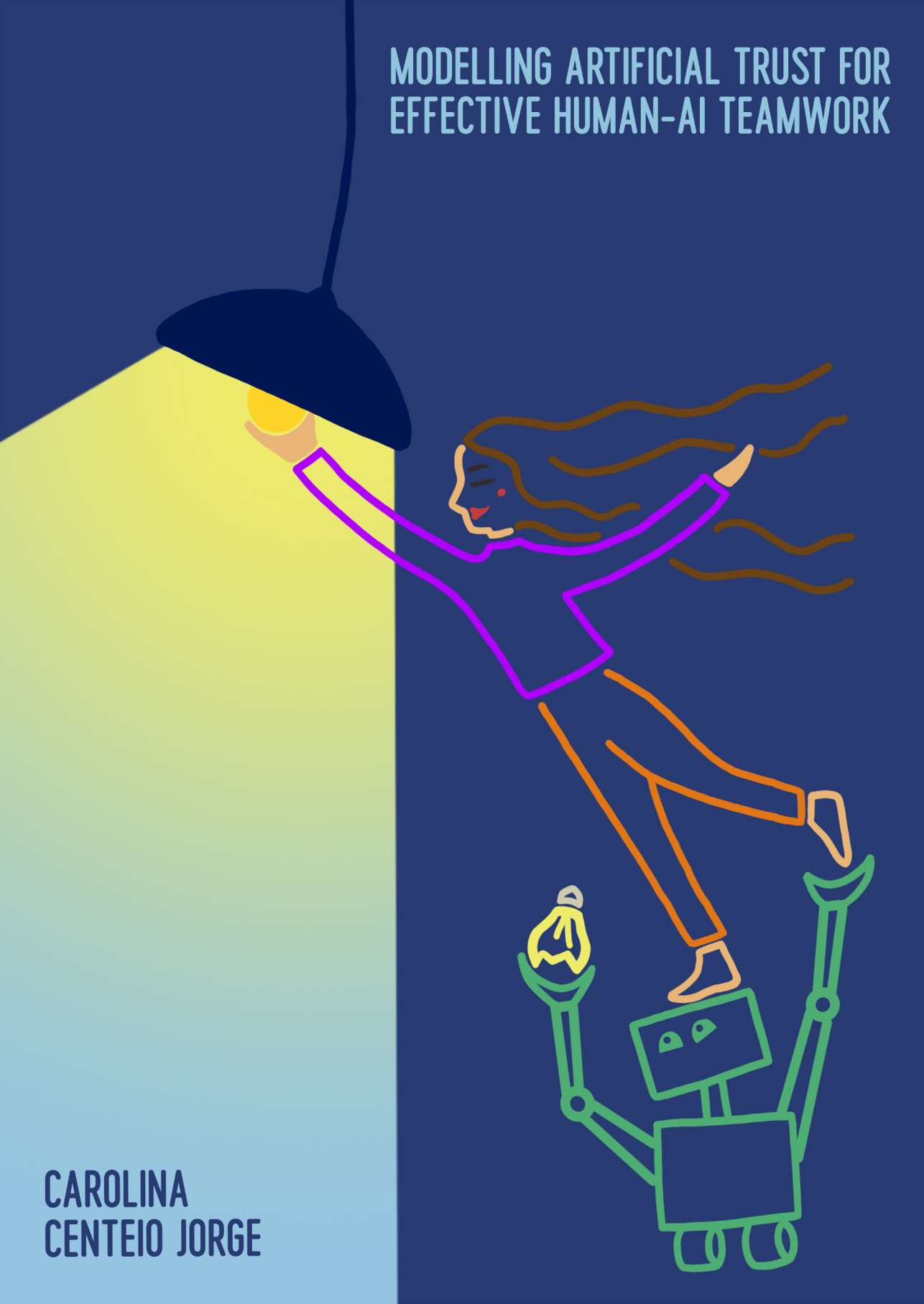
MODELLING ARTIFICIAL TRUST FOR EFFECTIVE HUMAN-AI TEAMWORK



CAROLINA
CENTEIO JORGE

Propositions

accompanying the dissertation

Modelling Artificial Trust for Effective Human-AI Teamwork

by

Carolina Ferreira Gomes Centeio Jorge

🎓 1. Human trustworthiness in human-machine collaboration depends on task criticality and duration.

🎓 2. The artificial agent's trust in the human affects the human's trust in the artificial agent.

🎓 3. Prioritising performance in human-machine teams is not sustainable.

🎓 4. Increasing human oversight of AI systems involves increasing the user's effort.

5. High mutual trust only contributes to human-machine team effectiveness when appropriate.

6. There is no ground truth for human behaviour models.

7. The benefit Artificial Intelligence can bring to society is hampered by the commercial benefit Artificial Intelligence can generate for corporations.

8. Academia rewards deep expertise within single fields better than multidisciplinary expertise.

9. Managing PhD candidates' expectations shapes their academic achievements and mental health.

10. Detachment from institutional governance contributes to one's vulnerability.

These propositions are regarded as opposable and defendable, and have been approved as such by promotor Prof. dr. C. M. Jonker, promotor Prof dr. M. A. Neerincx and copromotor dr. M.L. Tielman

🎓 Pertains to this dissertation.

Modelling Artificial Trust for Effective Human-AI Teamwork

Modelling Artificial Trust for Effective Human-AI Teamwork

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus prof. dr. ir. H. Bijl,
chair of the Board for Doctorates,
to be defended publicly on April 1st 2026, 12h30

by

Carolina FERREIRA GOMES CENTEIO JORGE

This dissertation has been approved by the promotors and the copromotor.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. C. M. Jonker	Delft University of Technology *(promotor)*
Prof. dr. M. A. Neerincx	Delft University of Technology *(promotor)*
Dr. M. L. Tielman	Delft University of Technolog, *(copromotor)*

Independent members:

Prof. dr. ir. M. Mulder,	Delft University of Technology
Prof. dr. J. M. P. Gevers,	Eindhoven University of Technology
Dr. R. Falcone,	ISTC-CNR, Italy
Prof. dr. L. C. Verbugge,	University of Groningen
Prof. dr. ir. W. P. Brinkman	Delft University of Technology *(reserve member)*

SIKS Dissertation Series No. 2026-16. The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: human-AI, teamwork, collaboration, artificial trust, task allocation

Printed by: ProefschriftMaken

Cover by: Carolina Ferreira Gomes Centeio Jorge

Style: TU Delft House Style, with modifications by Moritz Beller
https://github.com/Inventitech/phd-thesis-template

The author set this thesis in LaTeX using the Libertinus and Inconsolata fonts. Cynzia Bold on the cover.

ISBN 978-94-6518-260-5

An electronic version of this dissertation is available at http://repository.tudelft.nl/.

You only are free when you realize you belong no place,
you belong every place, no place at all.
The price is high. The reward is great...

Maya Angelou

Contents

Summary

As machines take on more complex tasks, we move from asking how well they can perform those tasks to asking how well they can collaborate with us. After all, the goal of building technology should be to improve our lives, not make them harder, but that requires mutual understanding, coordination, and trust. This dissertation looks at the role of trust in decision-making within teams of humans and semi-autonomous machines, including AI systems, agents and robots. In particular, we look at the concept of *artificial trust*, that is when an artificial agent reasons about someone's trustworthiness.

Trust is central to human decision-making. When we work with others, we constantly judge who is reliable and who is not, and we delegate tasks based on how trustworthy we think our teammates are and what risks those choices pose to us individually and to the team as a whole. When we see someone is not very trustworthy for a task they are expected to perform, and that poses risks to them or us, we can also offer help. The same logic can extend to artificial agents. When humans and intelligent artificial agents work together, artificial agents must not only be trusted by humans but also develop ways of assessing how trustworthy their human partners are for different tasks. In other words, artificial agents can use *artificial trust* to make decisions. This requires defining, modelling and using trustworthiness for decision-making in human–agent teamwork. We go over all of those steps in this dissertation.

This research argues that human trustworthiness is not only about a few internal traits such as ability, benevolence or integrity. In fact, what counts as trustworthiness can vary depending on the task and team characteristics. For example, if success in a task depends only on being somewhere on time, then punctuality may be the only relevant trait. Furthermore, to perform a task successfully, a person not only needs to be able to do it but also needs to choose to do it. Our research shows that in human–agent collaborative scenarios, task choices can often be explained by contextual cost–benefit reasoning. People consider a task by weighing its potential benefits, such as reward, against its potential costs, such as effort and time. This translates into a person's willingness to do a task. At the end of the day, it is not enough that someone has the skills to succeed in a certain task, but it is also important that they are willing to do it.

Although it is challenging to infer someone's willingness for different tasks, both for humans and machines, we can try to find ways around it. For example, asking directly about teammates' competence and willingness can give machines better information to work with, helping them to make fairer, more transparent and more efficient decisions. One of our studies found that people want artificial teammates, such as robots, to consider their preferences and willingness, but only in non-critical situations. In urgent or high-stakes work, efficiency mattered most. However, over time, recognising willingness may help make collaboration more sustainable and engaging.

This dissertation focusses on developing machines that can complement and even augment human teams, instead of replacing people. For that to happen, we need a solid

understanding of how people make decisions, what motivates them, and what they value in teamwork and in their artificial teammates. At the same time, giving machines the power to trust or distrust humans raises ethical risks. Used wrongly, it could harm individuals or undermine their autonomy. These concerns are especially pressing in areas such as defence, where collaborative technologies are already being explored, and can contribute to the escalation of armed conflicts. As such, the goal of this dissertation by building artificial trust is not to maximise efficiency at all costs. Instead, we hope to help design systems that support human well-being, safety, and dignity. This requires combining theoretical and technical advances from different disciplines, such as the social sciences and computer science, and carefully reflecting on the contexts where these systems are deployed.

Samenvatting

Naarmate machines steeds complexere taken uitvoeren, veranderd de vraag van hoe goed ze die taken kunnen uitvoeren naar hoe goed ze met ons kunnen samenwerken. Het doel van technologie moet immers zijn ons leven te verbeteren, niet te bemoeilijken, maar dat vereist wederzijds begrip, coördinatie en vertrouwen. Dit proefschrift onderzoekt de rol van vertrouwen in besluitvorming binnen teams van mensen en semi-autonome machines, waaronder AI-systemen, agenten en robots. We kijken met name naar het concept van *kunstmatig vertrouwen*, dat wil zeggen wanneer een kunstmatige agent redeneert over iemands betrouwbaarheid.

Vertrouwen is centraal in menselijke besluitvorming. Wanneer we samenwerken met anderen, beoordelen we voortdurend wie betrouwbaar is en wie niet, en wij delegeren taken op basis van hoe betrouwbaar we denken dat onze teamgenoten zijn en voor het hele team met zich mee kunnen brengen. Wanneer we zien dat iemand voor een taak die hij of zij moet uitvoeren niet erg betrouwbaar is, en dat risico's met zich meebrengt voor henzelf of voor ons, kunnen we ook hulp aanbieden. Diezelfde logica kan gelden voor kunstmatige agenten. Wanneer mensen en intelligente kunstmatige agenten samenwerken, moeten agenten niet alleen door mensen vertrouwd worden, maar ook manieren ontwikkelen om de betrouwbaarheid van hun menselijke partners voor verschillende taken te beoordelen. Met andere woorden, kunstmatige agenten kunnen kunstmatig vertrouwen gebruiken om beslissingen te nemen. Dit vereist het definiëren, modelleren en toepassen van betrouwbaarheid in de besluitvorming binnen mens–agent samenwerking. In dit proefschrift bespreken we al deze stappen.

Ons onderzoek stelt dat menselijke betrouwbaarheid niet alleen draait om een paar interne eigenschappen zoals vaardigheid, welwillendheid of integriteit. Wat als betrouwbaar wordt beschouwd, kan juist variëren afhankelijk van de taak en de kenmerken van het team. Als het succes van een taak bijvoorbeeld alleen afhangt van op tijd aanwezig zijn, kan punctualiteit de enige relevante eigenschap zijn.

Om een taak succesvol uit te voeren, moet een persoon niet alleen in staat zijn deze uit te voeren, maar ook ervoor kiezen om deze taak uit te voeren. Ons onderzoek laat zien dat keuzes in taken in mens–agent samenwerking vaak beter te verklaren zijn door contextueel kosten-batenafweging. Mensen wegen de potentiële voordelen van een taak, zoals een beloning, af tegen de potentiële kosten, zoals tijd en moeite. Dit vertaalt zich in de bereidheid van een persoon om een taak op te pakken. Uiteindelijk is het dus niet voldoende dat iemand de vaardigheden heeft; het is ook belangrijk dat hij of zij bereid is de taak uit te voeren. Het inschatten van iemands bereidheid voor verschillende taken is uitdagend, zowel voor mensen als voor machines. Door rechtstreeks te vragen naar de competentie en bereidheid van teamgenoten, kunnen machines betere informatie krijgen, waardoor ze eerlijkere, transparantere en efficiëntere beslissingen kunnen nemen.

Een van onze studies liet zien dat mensen willen dat kunstmatige teamgenoten, zoals robots, rekening houden met hun voorkeuren en bereidheid, maar alleen in niet-kritieke

situaties. In urgente of risicovolle situaties is efficiëntie het belangrijkst. Na verloop van tijd kan het herkennen van bereidheid echter bijdragen aan duurzamere en meer betrokken samenwerking.

Dit onderzoek richt zich op het ontwikkelen van machines die menselijke teams kunnen aanvullen en zelfs versterken, in plaats van mensen te vervangen. Hiervoor is het nodig dat we begrijpen hoe mensen beslissingen nemen, wat hen motiveert en wat zij waarderen in samenwerking en in hun kunstmatige teamgenoten. Tegelijkertijd brengt de mogelijkheid voor machines om mensen te vertrouwen of wantrouwen ethische risico's met zich mee. Verkeerd gebruikt kan individuen schaden of hun autonomie ondermijnen. Deze zorgen zijn vooral relevant in domeinen zoals defensie, waar al wordt geëxperimenteerd met collaboratieve technologieën, die kunnen bijdragen aan de escalatie van gewapende conflicten.

Het doel van het opbouwen van kunstmatig vertrouwen is niet het maximaliseren van efficiëntie, maar het ontwerpen van systemen die het welzijn, de veiligheid en de waardigheid van mensen ondersteunen. Het vinden van dit evenwicht vereist technische vooruitgang, inzichten uit disciplines zoals de sociale wetenschappen en zorgvuldige reflectie op de contexten waarin deze systemen worden ingezet.

Resumo

À medida que as máquinas são capazes de tarefas cada vez mais complexas, a questão deixa de ser apenas quão bem conseguem executar essas tarefas e passa a ser quão bem conseguem colaborar connosco, humanos. Afinal, o objetivo da tecnologia deveria ser o de melhorar a nossa vida, e não o de torná-la mais difícil. Para isso, é necessária uma compreensão mútua, coordenação e confiança entre pessoas e máquinas. Esta dissertação analisa o papel da confiança na tomada de decisão em equipas de humanos e máquinas semi-autónomas, incluindo sistemas de IA, agentes e robôs. Em particular, o conceito de *confiança artificial* é analisado, isto é, a possibilidade de um agente artificial confiar, ou não, em alguém.

A confiança é sem dúvida uma peça central nas nossas decisões. Por exemplo, quando trabalhamos em equipa, avaliamos constantemente quem é fiável e previsível e quem não é. Isso faz com que deleguemos tarefas àqueles que, baseado nessa perceção e nos riscos associados, prometam um melhor resultado, tanto para nós como para a equipa. Por outro lado, quando não estamos muito confiantes que alguém vá ter sucesso numa tarefa importante, podemos oferecer ajuda. A mesma lógica pode aplicar-se a agentes artificiais. Ou seja, quando pessoas e agentes artificiais inteligentes (por exemplo, robôs) trabalham juntos, os agentes precisam não só de ser considerados confiáveis pelas pessoas, mas também de saber em quem podem confiar para as diferentes tarefas. Para isso, estes agentes precisam de desenvolver modelos de confiança artificial que os ajudem a tomar decisões. Isto exige definir, modelar e aplicar confiabilidade na tomada de decisão em equipas humano-agente. Abordamos todos esses passos nesta dissertação.

Em particular, esta dissertação defende que a confiabilidade humana não depende apenas de algumas características internas, como competência, benevolência ou integridade. Na realidade, o que é considerado confiável pode variar dependendo da tarefa e das características da equipa. Por exemplo, se o sucesso de uma tarefa depender apenas de chegar a horas (por exemplo, marcar presença num evento importante), a pontualidade pode ser a única característica relevante para aferir quem deve ficar responsável por ela. Para além disso, para que uma pessoa realize uma tarefa com sucesso, não basta ter a capacidade de a executar, é também necessário que essa pessoa queira fazer essa tarefa. Esta investigação mostra que, em cenários de colaboração humano-agente, a forma como escolhemos tarefas pode ser explicada por um raciocínio custo-benefício dentro de um certo contexto. Ou seja, as pessoas ponderam os potenciais benefícios de uma tarefa, tais como as possíveis recompensas, face aos potenciais custos, tais como o esforço ou tempo de execução. Isto traduz-se na disposição ou vontade que uma pessoa tem em realizar a tarefa. Assim, tanto ter competências como estar disposto a executar uma tarefa são fatores que ajudam a prever o eventual sucesso dessa tarefa se atribuida a um determinado colaborador.

Apesar de ser difícil inferir a disposição que alguém tem para executar diferentes tarefas, tanto para nós humanos como para as máquinas, há formas de contornar isso. Por exemplo, perguntando diretamente aos elementos da equipa sobre as suas competências e vontades.

De acordo com um dos nossos estudos, as pessoas apreciam que os seus colegas artificiais, tais como robôs, tenham em consideração as suas preferências e vontades durante a tomada de decisões, mas só em situações que não sejam críticas. Por exemplo, em contextos urgentes ou de alto risco, a eficiência é considerada mais importante. No entanto, a longo prazo, reconhecer a as vontades dos elementos humanos da equipa pode vir a contribuir para uma colaboração mais sustentável e motivadora. Toda esta informação poderá ser fornecida aos agentes artificiais, permitindo-lhes tomar decisões mais justas, transparentes e eficientes.

Em suma, esta investigação foca-se em desenvolver agentes que podem complementar e até reforçar equipas humanas, em vez de substituir pessoas. Para que isso aconteça, é necessário compreender como as pessoas tomam decisões, o que as motiva e o que valorizam na colaboração e nos seus colegas artificiais. Ao mesmo tempo, conceder às máquinas a capacidade de confiar ou desconfiar de humanos levanta riscos éticos e sociais. Quando mal utilizada, esta tecnologia pode prejudicar indivíduos ou comprometer a sua autonomia. Estas preocupações são especialmente relevantes em áreas como a defesa, onde tecnologias colaborativas já estão a ser exploradas, podendo contribuir para a escalada de conflitos armados. O objetivo desta dissertação na modelação e utilização de confiança nos agentes não é o de maximizar a eficiência, mas sim o de desenhar sistemas que promovam o bem-estar, a segurança e a dignidade humana. Alcançar este equilíbrio exige avanços teóricos e práticos de diferentes disciplinas, tais como as ciências sociais e de computação. Exige ainda uma reflexão séria sobre os contextos em que estes sistemas são implementados.

1

Introduction

1

1.1 Motivation and The Societal Problem

> **Robots to the rescue: miniature robots offer new hope for search and rescue operations**
> *EU-funded researchers have developed robust mini robots with advanced sensors to help search and rescue teams find survivors in the aftermath of earthquakes and other disasters. (...) The idea is to allow rescue teams to do more of their work remotely, localising and finding humans from the most hazardous areas in the early stages of a rescue operation. The SMURF (Soft Miniaturised Underground Robotic Finder) can be remotely controlled by operators who stay at a safe distance from the rubble. The SMURF is compact and lightweight, with a two-wheel design that allows it to manoeuvre over debris and climb small obstacles.*
> by Michael Allen in *Horizon: The EU Research & Innovation Magazine* [18]

Robots such as SMURF have been developed to support humans across a wide range of tasks, including those in high-risk environments such as search and rescue operations. Similarly, other systems have been engineered for different domains, such as the remote operated robots deployed by TEPCO to decontaminate radioactive areas [416], or the Moley X-AiR robotic kitchen system designed for automated cooking tasks [269]. These and other machines have hardware and/or software with attributes different from humans'. For example, they are capable of physically strenuous, monotonous tasks and can also guarantee a level of precision and processing speed, which can be challenging for humans [132]. Furthermore, these machines can sometimes go through hazardous environments and prevent humans from certain risky situations. Humans can use these machines to leverage their strength and intelligence [295], and potentially improve our lives.

Advancements in artificial intelligence are allowing robots to demonstrate greater autonomy and richer interaction [219, 354, 420, 421]. We now live in a period where it seems increasingly plausible that machines could match the full range of human abilities soon [186, 197]. While early depictions of autonomy focused on replicating humans, they were also driven by the ambition that machines might perform tasks humans cannot or prefer not to do [136]. Focussing on building machines that match or exceed human abilities risks stripping the work of meaning and satisfaction [131]. When we allow full autonomy, we also distantiate humans from responsibility for the outcomes, raising difficult ethical issues [331]. Therefore, we argue that, ultimately, humans and machines, given their distinct strengths, can achieve more for society in collaboration than either could alone [11, 115, 296]. Machines now offer the potential to communicate with humans [22, 160], perceive the world [86], and adapt to us [185]. Combined with human creativity, ethical judgement, and emotional intelligence, machines with such physical and cognitive capabilities can function more like teammates than mere tools [342]. However, integrating autonomous machines into human teams presents several challenges [280]. Effective teamwork relies on driving mechanisms such as mutual trust, closed-loop communication, and shared mental models [329].

Trust, in particular, plays a crucial role in decision-making in teams, such as task selection and allocation [235, 257, 357]. Humans often make implicit assessments of potential collaborators in terms of trustworthiness, i.e., evaluating factors such as competence, reliability, and willingness to provide benefit to others [133]. These assessments influence

how someone chooses their team partners [177]. For instance, in school assignments, students select teammates either because they are friends or because they believe that the other can lead them to successfully complete the task [304], depending on what is more important to them. In addition, trust affects how we place ourselves in vulnerable positions within our team [250]. Moreover, trust determines how much information and knowledge we share with others [108], which affects how effective a team can be [309]. Although trust is required for effective teamwork, over-trusting (leading to over-compliance) can decrease performance [223] or lead to accidents [246]. For this reason, it is necessary to ensure that trust is appropriate between teammates, including in human-machine teams [202, 231, 294].

In human-machine teams, the artificial teammate (i.e., the machine) should also be able to select a suitable colleague when collaboration is required, decide when to be vulnerable to others, and determine what information to share. For this reason, we defend the notion that both humans and machines should trust each other appropriately, i.e., assess each other's trustworthiness within specific contexts. Although we more commonly talk about the importance of humans trusting machines appropriately (i.e., avoiding misuse or disuse of technology), we claim that having machines trusting humans appropriately may be as important for effective teamwork. Consider a hypothetical scenario in which an autonomous SMURF robot, operating within disaster debris, locates a victim and needs to request human assistance. The machine may use its knowledge regarding different human teammates' trustworthiness for retrieving victims, including the humans' strength to carry victims, speed when carrying, motivation to go to that location, among others. This knowledge can help the SMURF decide on who to ask for help, reducing risk and improving team performance.

When a machine has beliefs of trust in another entity, we call this artificial trust (AT) [25], which is inspired by but not the same as natural trust, i.e., the human construct. Artificial trust may be used as a tool for decision-making in human-machine teamwork. In the literature, we can find models of trust between artificial agents [50, 59, 122, 328, 381], as well as models of natural trust in artificial agents [276, 418]. Vinanzi et al. (2019) and Surendran & Wagner (2019) have attempted to model human trustworthiness for decision-making, mainly based on previous human performance or lies [363, 403]. Similarly, Azevedo-Sá et al. (2021) and Ali et al. (2022) propose a model for task allocation based on an artificial agent's trust in its human teammate's capabilities [16, 25]. However, computational models of human trustworthiness (including its different dimensions) that predict human performance in different tasks during human-agent collaboration remain inexistent. The vision and development of artificial trust for decision-making raises several research questions, discussed in this dissertation. It also presents potential opportunities and threats to society, which we discuss throughout our work.

1.2 Research question and goal

This dissertation aims to clarify what it means to be trustworthy for a task, how trustworthiness can be assessed by an artificial agent, and how it should influence this agent's decision-making. It also examines how this knowledge can be used for team design, and how trust-based decisions may impact the human teammate and human-agent teamwork. Throughout, we reflect on how to make these decisions transparent and understandable to

1

the human teammate. In this section, we present the main research question and goal of this dissertation.

Main Research Question

How can an artificial agent model artificial trust in its human teammates for effective human-machine teamwork?

The goal of the work presented in this dissertation is to theoretically formalise, model, and empirically test artificial trust as a multidimensional construct, as well as investigating the role of artificial trust in human-agent teams. Throughout this dissertation, we bridge the gap between social science models and computational methods. We examine the dynamics of trust, decision-making, and communication in human–machine teamwork from a human-centred computational perspective. The focus is on collaborative technologies with varying degrees of autonomy and interaction, whether embodied (e.g. robots) or not (e.g. software agents).

1.2.1 A note on terminology

As this is an initial step towards understanding the nature of interaction and trust dynamics in human–technology collaboration, a deliberately broad vocabulary is adopted. Although terms such as *agent*, *AI*, *machine*, and *robot* do not carry identical meanings, their distinctions are not essential for the scope of this work. Terminology is selected based on contextual relevance and the conventions of research community targetted per chapter. Typically, *machine* and *robot* are used when embodiment is implied, while *agent* provides a more general term consistent with the language of the multi-agent systems field. While humans are also considered agents, most of the times the term agent is used to refer to artificial agents (e.g., human-agent teams). When we use the term to refer to either humans or artificial entities, we make it explicit. As such, *agent* refers to artificially intelligent or interactive systems, including AI, machines, and robots, all of which share overlapping characteristics. Similarly, expressions such as *human–machine team*, *human–AI*, and *human–agent* are used interchangeably, as are *teamwork* and *collaboration*.

1.3 Scientific Background

The collaboration between people and semi-autonomous machines is constantly changing, as machines gain more capabilities to execute more tasks under more autonomy. The development and implementation of an artificial trust model requires research on several components, the necessary inputs (i.e., what is feeding it) and outputs (i.e., how it is being used and communicated), and their dependencies. For that, we need to dive into the literature, both from social sciences and computer science fields. In this section, we dive deeper into the scientific preliminaries needed to understand this dissertation's sub-research questions.

Teamwork has been defined as interrelated reasoning, actions, and behaviours of each team member that adaptively combine to fulfil shared team goals [329]. We have already mentioned before that the three driving mechanisms for effective teamwork are mutual trust, shared mental models, and closed-loop communication [329]. Although we know

that in human teams these three driving mechanisms are intertwined and affect each other, more research is required to develop and understand them in the context of *human-machine teams*. These mechanisms are explored in this section as a basis to motivate the work presented in this dissertation. First, we talk about mutual appropriate trust (derived from mutual trust), then human-machine team design (highly related to shared mental models), and then, communication (as in closed-loop communication). These driving mechanisms are contextualised, as well as connected to a simple objective measure of team effectiveness, i.e., *task outcome*. Finally, we reflect on how adapting these driving mechanisms to human-machine teams is a multidisciplinary problem. In a visually simplified diagram, Figure 1.1 presents the components surrounding artificial trust in human-machine teams, studied in this dissertation.

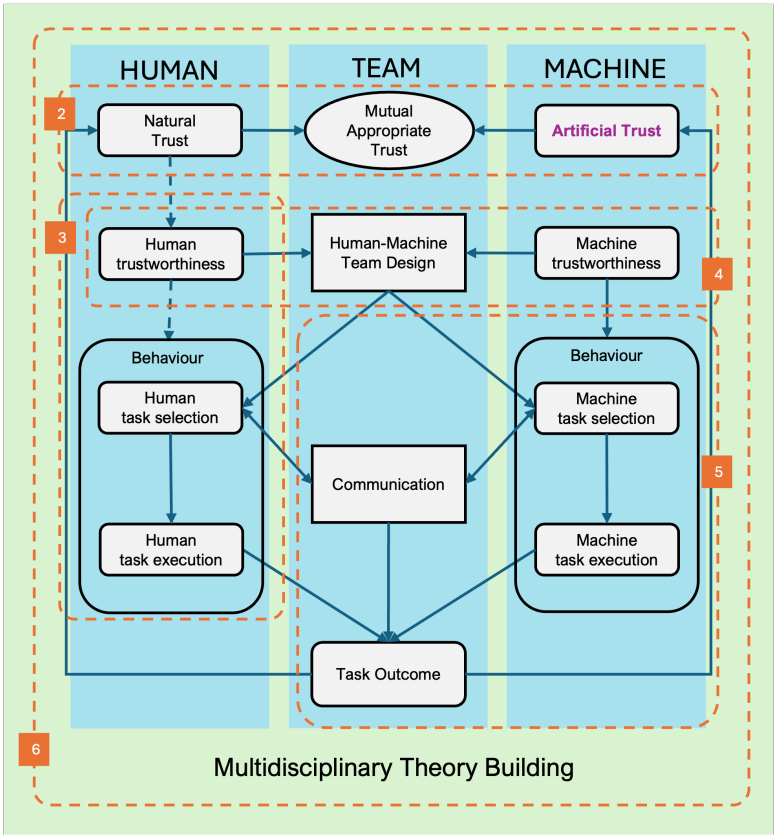


Figure 1.1: Diagram of the dissertation's main concepts and related Chapters, in numbers.

1.3.1 Mutual appropriate trust

Teamwork is safer and more effective when team members know how much they can trust each other for different tasks and contexts, i.e., when trust among team members is appropriate and calibrated [178, 190, 294]. The machine's role in mutual appropriate team

1

trust is twofold. On the one hand, the verbal and non-verbal behaviour of the machine, as well as the consequences of its decisions, affects the human trust [338], both positively and negatively [215, 261, 404]. On the other hand, the ultimate goal of artificial trust modelling should be to make it appropriate, enabling the machine to know who can be expected to perform which task successfully, and make better informed decisions [65, 153]. Although this dissertation focusses mainly on questions related to artificial trust, we always keep human trust in mind, as they are likely to be intertwined [256, 332, 375, 404].

Trust The idea of modelling artificial trust appears in multiagent systems literature under the names of trust (as in Falcone et al. (2004) [122]) or computational trust (as in [381]). However, it stems from the definitions and models of trust in social sciences literature. In organisational psychology, trust can be defined as *"the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, regardless of the ability to monitor or control that other party"* [250] (p. 712). In uncertain situations, when a team member trusts another with a task, they expect the other to perform it successfully. One (the trustor) forms trust in another (the trustee) by assessing the trustee's trustworthiness. This trust can then be influenced by the trustor's personal characteristics (such as propensity to trust) and the context (such as the risk involved) [250]. As such, artificial trust (AT) can be modelled as an aggregation of beliefs [145] regarding the teammate's *trustworthiness*. In order to understand which beliefs are relevant for artificial trust in teams, it is useful to first understand the dimensions of perceived trustworthiness.

Trustworthiness dimensions Models in slightly different settings and disciplines propose that trustworthiness depends on 1) Ability, Benevolence and Integrity [250], in human organisations; 2) Willingness, Competence [59], in multi-agent systems; and 3) Performance, Process, and Purpose [227], when the human is the trustor and an artificial agent is the trustee. All these models usually have one component related to the competence/performance aspects (more objective), i.e, answering the question *Can my teammate do that task?*. As performing a task successfully is dependent on other factors besides just whether one is capable, these models include at least one aspect which is more dependent on the trustee's willingness (such as benevolence or purpose). These aspects try to answer the question *Will my teammate do that task?*. Finally, it may be relevant to know *how* a teammate performs a task, for example, to which set of values they adhere to, which is related to aspects of integrity and process. To model artificial trust, one needs to choose a set of internal characteristics (the trustee's krypta [126]) that are relevant for the context and evaluate them [328]. This phase, called trust evaluation [328], consists of mapping the krypta to the accumulated available information, such as directly observable cues and behaviours (also known as manifesta [126]), or reputation.

Trustworthiness cues Defining which information is important to evaluate contextual trust in humans in human-machine interaction is challenging and underexplored. However, we can find models that evaluate whether a human partner is being truthful or deceitful with episodic memory [403] and social cues [363]. There are also works exploring methods for the detection of intentions [401] and natural trust [9, 151] in interaction with embodied

AI. With studies in 2D grid-worlds, literature presents metrics to assess teamwork fluency, such as metrics of performance or task completeness [61, 396]. Similarly, we can find metrics for ability, benevolence and integrity, such as speed, favouritism, and commitment, respectively, also in a 2D grid-world [67]. Finally, [54] presents a model that learns the human teammate's sequential behaviour, using reinforcement learning. To cover for limited information, multiagent models such as [50] propose ways to model stereotypes. Once trust is evaluated, it is time for the agent to make a decision regarding its action [328]. Modelling all teammates' trustworthiness for different tasks can be used not only for an artificial agent's decision-making but also for team design.

1.3.2 Human-Machine Team Design

Shared Mental Models As we move through different environments and interact with objects and other agents, we develop internal representations of the world, known as *mental models* [199]. An agent's mental model provides the basis for understanding and predicting the actions of other agents [196, 321], and is therefore closely linked to communication and trust within teams [329, 335]. When team members share overlapping representations of the task and the team, known as *shared mental models*, team performance improves [53, 104, 247]. Shared mental models capture key aspects of the *task* and the *team*, forming the basis for understanding and predicting outcomes at both the task and team level, including individual contributions [199, 247, 267, 385]. In human–machine teams, shared mental models are especially important to ensure alignment in tasks, roles, interdependencies, and strategies [335], and require adequate representation [199, 336]. Shared mental models and their representation are used for *Human-Machine Team Design*, where the known characteristics of the tasks and the teammates determine the possibilities for task allocation and collaboration.

Interdependence Analysis Human-machine effective teamwork demands coordination among humans and machines, along with the capacity to adjust and adapt to maintain effective performance [417]. Team design's goal is to support coordination by leveraging the characteristics of the humans and machines involved, making the most of the possible combinations and interdependencies. Johnson et al. (2014) [194] defines interdependence as "the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activity." In teams of one human and one machine, tasks can be performed independently, with support from the other, or jointly when both are required [194]. The design of human-machine teams involves defining task requirements and creating decision tables across different contexts [316]. Modern frameworks emphasise the importance of human-centredness, ensuring that technology supports human well-being, agency, and ethical considerations [39, 204]. Moreover, there is a focus on designing these systems in a way that ensures meaningful human control (MHC), i.e., passing the authority to the human when the context is morally sensitive [389]. However, dynamic delegation of authority is complex and requires a framework to balance control and monitoring across tasks [316]. Analysing which factors enable particular interdependencies is an essential first step in task selection and allocation, and can also serve as a shared mental model. A key contribution is the Coactive Design Interdependence Analysis table [194], which helps designers match human and machine capabilities

with the interdependencies most suitable for each subtask in the team's mission. This approach focusses mainly on one dimension of trustworthiness, namely capacity (related to competence, performance, or ability) while leaving the others (such as willingness) aside. Integrating the full trustworthiness framework could support more comprehensive and informative team design and, consequently, task selection and allocation.

Task selection and allocation The MABA–MABA list (men-are-best-at, machines-are-best-at) [134] highlights complementary strengths of humans and machines, and many task allocation methods remain competence-driven to maximise performance, particularly in manufacturing [51, 243, 410]. Since human competence varies with factors such as fatigue or familiarity [242, 263], dynamic allocation methods are needed [17, 302, 409]. Motivation also plays a role: people weigh effort against reward [273], and consistently assigning disliked tasks can reduce engagement [369], whereas aligning with personal interests improves outcomes [176, 208]. Allowing users to influence allocation increases autonomy and satisfaction [370, 371], and some methods aim to elicit or learn task preferences [107, 427]. Azevedo-Sá et al. (2021) and Ali et al. (2022) introduced the use of artificial trust (AT) for optimal task allocation [16, 25]. Although this method focusses on team members' capabilities, the algorithms presented in these papers suggest that trustworthiness dimensions, including willingness, could be used for task allocation and selection algorithms. Enabling machines to perform tasks when humans lack competence or willingness complicates scheduling, as optimisation cannot be based solely on minimal cost [288]. Instead, thresholds and objectives (such as satisfaction, risk, or performance), as well as the weight of each criteria (such as competence and willingness), must be defined, ideally through *closed-loop communication* with humans [259].

1.3.3 Communication

Closed-loop communication supports shared mental models and mutual trust among teammates [329]. For mutual and appropriate trust, the agent should be transparent and able to explain its decisions [418]. As a broader concept, communication is seen as a central point in the human-AI team processes and a facilitator of shared knowledge [424], supporting cognitive [137] and affective processes [335], while also improving job satisfaction [147]. In particular, Explainable Artificial Intelligence (XAI) methods [20, 175] can affect human trust and behaviour, and consequently team performance [397]. The more authority (and autonomy) the artificial teammate has, the more communication the human typically needs [81]. Explanations can be presented in text, audio, visuals, or mixed modalities [12, 23, 425]. During human-agent collaboration, explanations need to be generated, communicated and received [282]. Particularly, agents need to decide what information to share and when [391, 424]. Furthermore, effective communication of failure (or lower values of artificial trust) may also compromise trust relationships with the human teammate (see, e.g., [106, 215, 216]). When using artificial trust to make decisions, the agent should be able to appropriately communicate its trust model to the human teammate.

1.3.4 Task outcome

Decisions lead to actions (e.g., *machine task execution*) that have outcomes, such as success and failure, as well as consequences for teammates and environments. The outcome of

a machine's action, as well as the perception of the agent's mental model, influences the human's feelings towards the agent, including the (natural) trust in the agent and their behaviour towards the agent [373]. Particularly, when the machine allocates tasks and the team performs well, satisfaction with the outcome increases [5]. Literature also shows that malfunction of the machine in a human-machine teamwork scenario affects human teammate's willingness to collaborate negatively [61], which can be moderated by the way the mistake is communicated [215]. On the other hand, the trust the human has in the artificial teammate can predict the next task outcome [170].

1.3.5 Multidisciplinary theory building

Several of the ideas presented in this dissertation involved constant iteration, from refining ideas, to formalising definitions, and developing methodologies across disciplines, from Computer Science to Organisational Psychology, and others. These required diving into the literature of these different disciplines and learning how to learn from social science researchers, as a computer scientist and engineer. Working in a multidisciplinary field is challenging, and we have learnt several lessons. For that reason, part of this dissertation is dedicated to the dissection of this process, which is multidisciplinary theory building.

1.4 Outline

Figure 1.1 divides the driving mechanisms of human-machine teamwork and their related components in three subgroups: the human-related, the team-related and the machine-related components. Both the human and the machine present their trust, trustworthiness and the behaviour that results of these. The team subgroup presents the three driving mechanisms of effective teamwork: mutual appropriate trust, human-machine team design, and communication. As discussed in Section 1.3, applying any of these driving mechanisms poses several questions, motivating the sub-research questions explored in this dissertation. These sub-questions are:

- RQ.a How can we conceptualise artificial trust beliefs in human-agent teams?
- RQ.b How can we assess the trustworthiness of a human teammate, given a task?
- RQ.c How can we use the trustworthiness of teammates (human or artificial) for the design of human-machine teamwork?
- RQ.d How does using human willingness for task allocation affect human-machine teamwork?
- RQ.e How to build multidisciplinary theory for human-AI team trust?

This dissertation contains seven chapters: this current introduction (Chapter 1), five content chapters (Chapter 2, 3, 4, 5 and 6), each focussing on a different sub-question, and a concluding chapter at the end (Chapter 7). An overview of chapters, corresponding research questions, contributions, and evaluation methodologies can be found in Table 1.1. The relationship between the chapter (and the respective research question) and its focus in terms of concepts is represented in Figure 1.1 through an orange dashed line marked with their respective number. The final concluding Chapter 7 includes a summary of this dissertation's scientific and societal contributions, implications, limitations, and future directions.

Table 1.1: Main contribution and approach used in each of the content chapters in this dissertation.

Chapter	Contribution	Approach
2 (RQ.a)	Formal theoretical model of artificial trust and taxonomy of contextual factors.	Theory building.
3 (RQ.b)	Conceptual model for human trustworthiness with empirical evaluation.	User study & data analysis.
4 (RQ.c)	Framework for human-machine team design.	Focus group & data analysis.
5 (RQ.d)	Empirical evaluation of willingness as a factor in task allocation.	User study & data analysis.
6 (RQ.e)	Guidelines for multidisciplinary collaboration.	Reflections on experiences.

1.4.1 Chapter 2

RQ.a

How can we conceptualise artificial trust beliefs in human-agent teams?

Chapter 2 explores the concept of Artificial Trust (AT) and its role in decision-making within human-machine teams. This chapter examines the conceptualisation and formalisation of artificial trust in human-agent teams. It argues that artificial agents need structured beliefs about human trustworthiness to make informed decisions, such as task allocation, offering assistance, or requesting help, and that formalising these beliefs is the first required step. The chapter also explores how trust is context-dependent, i.e., the characteristics that make a teammate reliable in one task may not apply in another. To address this, this chapter presents a taxonomy of task and team characteristics that affect artificial trust modelling. Furthermore, trust is also interdependent of other trust dynamics within a team, i.e., an agent's trust in one teammate can influence how other teammates trust. We situate artificial trust modelling within human-agent teamwork, and reflect on how it should be built taking other trust and team dynamics into account. Finally, this chapter identifies the steps and challenges necessary to develop agents capable of making appropriate, context-sensitive, trust-based decisions, highlighting the challenges of such path.

Contributions: (1) formalisation of artificial trust, (2) a taxonomy of task and team characteristics that influence the modelling of artificial trust, (3) the conceptualisation of artificial trust within a human-AI team, and (4) identification of the steps and challenges towards using artificial trust for decision-making.

1.4.2 Chapter 3

RQ.b

How can we assess the trustworthiness of a human teammate given a task?

The organisational psychology literature tells us that what makes a human teammate trustworthy to other humans is their ability, benevolence, and integrity [250]. On the other hand, computer science tells us that artificial agents can form trust beliefs about other artificial agents: through the assessment of teammates' competence and willingness in certain tasks [122]. However, we lack insight into how artificial agents can form trust beliefs about human teammates. This includes determining which human internal characteristics (known as krypta [126]) are essential to assess artificial trust, how these components can be observed in human behaviour (referred to as manifesta [126]), and how they should be weighed for a final trust assessment. With the goal of forming beliefs in competence and willingness (inspired by multi-agent systems literature [57, 122]), we investigated which human trustworthiness components from the ABI model (i.e., ability, benevolence, and integrity) could be observable during teamwork.

To explore this, we conducted our first user study, which is presented in Chapter 3. In this study, participants collaborated with two artificial teammates in a 2D grid-world. Their task was to help agents collect items within a supermarket scenario. We systematically manipulated participants' ability, benevolence, and integrity across conditions. By analysing how participants' behaviour changed under these conditions, we aimed at identifying which human cues might be relevant for forming beliefs about competence and willingness. However, this study highlighted a key challenge: translating abstract psychological concepts, such as willingness, into observable behaviours is not straightforward. So, instead of focusing solely on traditional trustworthiness dimensions, we leaned into our findings: people's actions seem driven by strategic decision-making or a cost-benefit analysis, related to their competence and willingness towards a task, rather than centralised on their relationship with their teammate.

Contribution: (1) conceptual model of human trustworthiness in human-agent teams, (2) explorative user study with four conditions, (3) Bayesian analysis of the user study's objective and subjective measures.

1.4.3 Chapter 4

RQ.c

How can we use the trustworthiness of the teammates (human or artificial) for the design of human-machine teamwork?

The findings presented in Chapter 3 discouraged us from modelling human trustworthiness for decision-making in human-machine teams using ability, benevolence, and integrity. Instead, they encouraged us to adopt [122]'s abstraction of beliefs in human competence

and willingness (where task preference can also be included). The next step was therefore to model human trustworthiness as competence and willingness in specific tasks and to use this as the basis for task allocation in human-machine teams. Johnson's Coactive Design [194] is a framework that supports task allocation by analysing team members' capacities, which we extended to fit our goal. Since capacity and competence are abstractions of the same characteristics (i.e., skills, knowledge, ability), we renamed it *competence* and added the dimensions of *willingness* and *external factors*. We present our extension, the Interdependence and Trust Analysis (ITA) table, in Chapter 4. By including information on contextual trustworthiness (i.e., a teammate's competence and willingness toward a certain task), ITA can better support the design of human-machine teams, as well as task allocation and selection. To evaluate the table, we conducted two expert interviews and a focus group involving a search and rescue scenario. ITA showed potential as a decision-making tool and a communication bridge among human and machine teammates. Our findings emphasise the need to define tasks and roles based on agent characteristics and imply that decision-making models should align with human-centred objectives. We believe that the ITA framework may improve transparency, justification, and interpretability in decision-making, contributing to appropriate trust among teammates.

Contribution: (1) framework to design human-machine teamwork which takes into account the team members' competence and willingness, as well as possible context restrictions, and (2) dyadic interviews and focus group with experts that evaluated the table, which is the main component of the framework.

1.4.4 Chapter 5

RQ.d

How does using human willingness for task allocation affect human-AI teamwork?

With the Interdependence and Trust Analysis framework, we had a good structure to implement an agent's decision-making from human trustworthiness assessments. However, adding willingness (besides competence/capability/capacity) to task allocation turns it into a multi-objective this optimization problem. In particular, one is met with the decision to maximize either competence or willingness, or find something in-between. For example, if teammate A is competent but not willing to do task 1, but teammate B is not competent but willing, choosing who should be assigned such task presents a complex optimisation problem. Furthermore, we do not know the effects of optimising for overall performance (as in [16]) when compared to optimisation for human satisfaction, for example. From our perspective, effective teamwork should not be solely about immediate performance. In fact, disregarding human willingness in task allocation may increase short-term success but could undermine future collaboration. Ultimately, we realised that we lack an understanding of how considering willingness, particularly for task allocation, impacts the human-agent relationships and the teamwork.

Furthermore, we were interested to see the impact of having a closed-loop communication in willingness-based task allocation. Allowing people to influence or participate in

task allocation and decision-making can increase perceived autonomy, satisfaction, and performance [26, 370, 371]. When the artificial teammate presents a willingness-based task allocation plan, it may make a difference how much people feel the need to alter those plans, when compared to other allocation plans. In fact, allowing people to change the allocation plan freely may impact their relationship with the agent and their behaviour during teamwork. As such, in Chapter 5, we present a 2x2 mixed design user study to explore the effects of allocating tasks taking into account the human teammate's willingness. We ran another 2D grid-world experiment, where participants collaborated with two different virtual robots (at different times) in a search and rescue scenario. One of the virtual robots suggests a task allocation based on the participant's expressed willingness, while the other distributes them equally in terms of effort. Furthermore, the study presents two between conditions: one where the human teammate can alter the task allocation, and one where they cannot. We analysed the effects of willingness-based task allocation and human input in the human's performance, trust, and satisfaction. Our results suggest that participants want their artificial teammates to take into account their preferences and willingness, but only when the task is not critical. For critical and urgent tasks, participants seem to prefer prioritising their involvement and overall efficiency. These results suggest, however, that willingness may play a stronger role in longer, less critical collaborations. Furthermore, giving the chance to alter the task allocation plan seems to make participants more proactive and engaged in the task.

Contribution: (1) 2x2 mixed design user study, (2) Bayesian analysis on the effects of willingness-based task allocation and human input on human's performance, trust and satisfaction levels.

1.4.5 Chapter 6

RQ.e

How to build multidisciplinary theory for human-AI team trust?

All work presented in this dissertation was developed through multidisciplinary studies, which resulted in multidisciplinary theories. This presented a series of challenges and lessons learned, which are common to several other multidisciplinary research topics. As such, we wrote a handbook chapter about multidisciplinary theory building for human-AI team trust, which reflects our own experience in collaborating with researchers from other fields. In this cchapter, we describe how our collaboration with organisational psychologists emerged in a new multidisciplinary field. Throughout our collaborative process, we worked to overcome limitations inherent in each of our disciplines and challenges in translating concepts and methods between fields. We accomplished this by providing theoretical and mathematical definitions of trust components and offering a comprehensive analysis of team trust dynamics in human-AI teams. By being sensitive to the differences between disciplines, we seek to contribute and motivate the interdisciplinary investigation of human-AI teamwork and trust. Moreover, we hope our multidisciplinary approach serves as a model for future research initiatives that bridge two or more fields of study.

1

Contribution: guidelines for an effective multidisciplinary collaboration, with focus on theory building.

2

Defining Artificial Trust in Human-Agent Teams

Mutual trust is a central element of teamwork, and in human-agent collaboration it has mainly been studied in the direction where the human is the trustor and the artificial agent is the trustee. This chapter addresses the challenge of enabling artificial agents to make trust-based decisions by proposing a formalisation of artificial trust. We focus on dyadic human-agent relationships as a starting point, and conceptualise trust as a belief in directed trustworthiness. The framework specifies how artificial agents can form and update beliefs about their own trustworthiness and that of human teammates, using manifesta as observable cues of underlying krypta. We present a taxonomy with the different team and task characteristics that can influence the choice of krypta and manifesta, based on the literature. Furthermore, we discuss how such beliefs extend across individuals, dyads, and teams, and how they can inform collaborative decisions such as task allocation, support provision, and risk management. The chapter also outlines the methodological challenges of artificial trust formalisation and implementation, including the lack of ground truth and the complexities of evaluating models in user studies. Our aim is to provide a theoretically grounded basis for developing agents that can reason about trust in ways that foster mutual appropriate trust, improve teamwork effectiveness, and ensure safer collaboration.

This chapter is partly based on:

Centeio Jorge, C., Mehrotra, S., Tielman, M. L., & Jonker, C. M. (2021). Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams. In Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021): Co-located with the 20th International Conferences on Autonomous Agents and Multiagent Systems (AAMAS 2021) (Vol. 3022). CEUR-WS. [70].

Centeio Jorge, C., Tielman, M. L., & Jonker, C. M. (2022, March). Artificial trust as a tool in human-AI teams. In 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 1155-1157). IEEE. [72].

Centeio Jorge, C., Jonker, C. M., & Tielman, M. L. (2023). Artificial trust for decision-making in human-AI teamwork: Steps and challenges. In Proceedings of the Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence co-located with HHAI 2023 (Vol. 3456, pp. 150-156). CEUR-WS. [65]

Centeio Jorge, C., van Zoelen, E. M., Verhagen, R., Mehrotra, S., Jonker, C. M., & Tielman, M. L. (2024). Appropriate context-dependent artificial trust in human-machine teamwork. In Putting AI in the Critical Loop (pp. 41-60). Academic Press. [200].

2.1 INTRODUCTION

Artificial agents are increasingly able to perform tasks in daily life, including work environments, home assistance, battlefield operations, and crisis response [232]. In such contexts, humans and artificial agents must cooperate, coordinate, and collaborate [111, 194], forming *human-agent teams*. A key requirement for effective teamwork is *mutual trust* [329], as the trust teammates place in one another shapes both actions and overall team performance [231]. This trust must then be *appropriate*, avoiding over-trust, which may cause accidents, and under-trust, which may reduce efficiency [294]. While most research has focused on how to foster appropriate human trust in technology [261], it is equally important for artificial agents to trust their human teammates appropriately. Such trust enables agents to make better decisions about task allocation when, for example, some teammates may be unreliable (e.g., they do not have the skills to do a certain task) [16]. To support this, we first need a model of how agents should form beliefs about human trustworthiness in collaborative scenarios. This chapter introduces a formalisation of the beliefs and concepts underlying artificial trust in the context of human-agent teamwork.

Trust can be described as an expectation about another's actions, based on perceptions of internal characteristics such as ability, benevolence, or integrity [250]. Since these internal characteristics are not directly observable, people infer trustworthiness from cues such as verbal or non-verbal behaviour [392, 408]. For example, if someone is always punctual, I may judge them as committed and therefore trust them in tasks where punctuality matters, even though I cannot directly access their true level of commitment. Humans rely on such qualitative judgements, but artificial agents ultimately operate on numerical representations, whether explicitly defined or encoded in model parameters [171, 379]. Adopting an explicit and interpretable numerical structure, rather than leaving values implicit in a model's internals, supports transparency in how trust is computed and enables others to understand or challenge the agent's reasoning [30]. This means that for an artificial agent to trust, it must compute trust values from relevant characteristics, which in turn requires a clear specification of what to model and how to learn it. Belief formalisation provides this specification: it allows the artificial agent to represent human characteristics as beliefs, update them over time, and use them as the basis for trust-based decisions. Multi-agent systems literature presents several computational trust models (see e.g., [50, 59, 87, 95, 124, 379]), which ours builds on, but they fail to directly address how to apply these for artificial trust in humans for collaborative scenarios. Formalised models that can be used for artificial trust in humans for collaborative scenarios are necessary for artificial agents to maintain appropriate levels of trust and to act consistently and transparently in human-agent teams.

In addition to belief-based approaches, other computational frameworks have been proposed for modelling trust, from formal techniques to data-driven ones. Game-theoretic models, for example, characterise trust as expectations of another agent's actions in strategic interactions, often based on payoffs, probabilities, and repeated play [129], while theory-of-mind approaches aim to predict others' behaviour by reasoning about their intentions and higher-order beliefs [394], capturing aspects of willingness or cooperative intent. On the other hand, bayesian and probabilistic models represent trust as a distribution over possible behaviours that can be updated with new evidence, allowing dynamic adaptation to observed actions or outcomes [82]. These approaches differ from the belief-based

formalisation used in this chapter, which explicitly represents trust as a belief about human trustworthiness, integrating multiple dimensions such as competence and willingness. Representing trust as a belief allows the agent to reason about unobserved characteristics, update its estimates dynamically based on interaction outcomes, and make context-sensitive decisions, rather than relying solely on predicted actions. Although out of the scope of this chapter, game-theoretic, theory-of-mind, and Bayesian models could complement this framework: theory-of-mind reasoning could inform the willingness component of the trust belief, game-theoretic models could provide priors or estimates of likely behaviour, and Bayesian approaches could support the updating of beliefs under uncertainty.

Trust is context-dependent [49, 360]. Following the example above, finding a person reliable in tasks where punctuality matters does not necessarily mean finding them trustworthy for driving safely, as different tasks require different internal characteristics. This means that the characteristics that make someone trustworthy in one situation may not apply in another [16, 298]. Since trust can be understood as an overall belief in another's trustworthiness [42, 380, 422], artificial trust modelling should be context-sensitive. Moreover, trust involves two components: evaluation and decision [328]. The same level of trust evaluation may lead to different decisions depending on contextual characteristics, such as risk [193]. For instance, in urgent situations, a lower level of evaluated trust may be sufficient to act in a trusting way. Both the evaluation of trust and trust-based decisions are therefore context-dependent [171, 333]. When modelling trust for an artificial agent, the context must be taken into account to determine which beliefs are relevant, how they can be perceived or measured in that situation, and how to make a decision.

In addition to considering the context, modelling artificial trust should be sensitive to team dynamics. Artificial trust beliefs influence and are influenced by other trust beliefs within the team [376, 404]. These beliefs can be considered at multiple levels, such as individual (trust in or of a single teammate), dyadic (trust between pairs of teammates), and team level (trust in the team as a whole). An agent's trust in one teammate can affect its evaluation of others, shaping task allocation, coordination, and overall team performance [3, 328]. Similarly, how teammates perceive the artificial agent's trust in them and in others can alter their own trust and behaviour, creating a feedback loop that affects team dynamics [154, 373]. Accounting for these interactions is, therefore, essential when modelling artificial trust, as it determines not only how an agent forms beliefs but also how those beliefs impact decisions and the team as a whole.

The primary goal of formalising and modelling artificial trust is to enable an artificial agent to make informed decisions that account for the situational trustworthiness of human teammates [363, 403]. Such decisions include task selection and allocation, as well as whether to offer assistance or request help. In the latter case, beliefs about trust and trustworthiness guide the artificial agent in choosing the most appropriate human teammate to approach [25]. Formalising these beliefs and incorporating context-dependent characteristics and cues of trustworthiness represents only the first step. This chapter concludes by outlining the subsequent steps needed to develop artificial agents capable of making trust-based decisions appropriately and highlight the challenges that remain along this path.

This chapter contributes to the conceptualisation of artificial trust beliefs in human-agent teams by (1) formalising artificial trust, (2) presenting a taxonomy of task and team

characteristics that influence the modelling of artificial trust, (3) conceptualizing artificial trust within a human-agent team, and (4) identifying the steps and challenges towards using artificial trust for decision-making. We begin by defining artificial trust in Section 2.2.1, where we formalise trust as a belief in trustworthiness (Section 2.2.3), with a particular focus on trust directed towards humans (Section 2.2.4), and discuss how this should be contextualised (Section 2.3). We then examine the role of artificial trust in teams (Section 2.4) and, in Section 2.5, outline the steps required to enable trust-based decision-making together with the main challenges it entails. The chapter closes with a conclusion in Section 2.6.

2.2 Artificially trusting human teammates

2.2.1 Defining Artificial Trust

In this dissertation, we are interested in studying how an artificial teammate can appropriately trust its human counterparts, in order to make informed decisions that lead to effective teamwork. Although the concept of having artificial agents with the ability to trust has been around for a while in multi-agent systems (e.g., [122, 123, 328]), only recently has there been an interest for an artificial agent to trust a human (e.g., [364, 403]). Saying that an artificial agent can be enabled to trust a human is controversial, since people are highly sensitive to how others perceive their trustworthiness [353], and with good reason fear that being classified as untrustworthy may exclude them from opportunities such as employment or loans [210]. Trust is a human concept, and while some researchers defend it can be used in a human-AI relationship [37], others strongly disagree [324]. The latter defend that trust cannot truly be modelled, and that artificial agent developers should instead focus on terms such as expectation or reliability.

Potentially alleviating these worries, Azevedo-Sá et al. (2021) [25] introduced the term *artificial trust* as a trust relationship in which the trustor is an artificial agent. The authors distinguish this concept from *natural trust*, i.e., a trusting relationship where the trustor is a human, and open the road for an independent investigation of artificial trust, where models and definitions of trust and trustworthiness, used on artificial agents do not have to align with the existing theories from social sciences. Artificial trust follows the definition of Kok and Soh (2020), which states that "given a trustor agent A and a trustee agent B, A's trust in B is a multidimensional latent variable that mediates the relationship between events in the past and A's subsequent choice of relying on B in an uncertain environment" [25, 211], where latent means that it is not directly observable and instead it is mathematically calculated from other variables. Artificial trust can be calculated based on weighted contextually-relevant internal characteristics (the *krypta*) that can be observable through behaviour (the *manifesta*). Departing from this notion, several works have advanced research on artificial trust for human-AI teams, such as exploring how it can be modelled [67], how it can be used for decision-making [49, 72] or task allocation [16].

We distinguish *artificial trust*, i.e., the trust of an artificial trustor, from *computational trust*. Computational trust can be defined as the formal modelling of trust beliefs and dynamics into algorithms, representing either natural or artificial trust [381]. Computational trust models often collect information such as direct experiences, reputation, or recommen-

dations to assess the trust or trustworthiness of agents, whether they represent individuals, organisations, or artificial entities [42, 327, 422]. As seen in the previous paragraphs, when these beliefs are used to translate trust, and the trustor is artificial, then we are talking about artificial trust, otherwise it is the computation of natural trust, for instance for the purpose of simulation or understanding humans. The computation of natural trust is the basis for calibrating trust in teams that involve humans, and assessing team trust.

2.2.2 Artificial trust for driving a dual-mode vehicle: an example

To illustrate why artificial trust in humans is necessary, consider the task of driving a car. Inspired by Mecacci and Santoni de Sio [258], imagine a dual-mode vehicle that can be driven either by a human or by an artificial agent. The default is that the human drives according to the artificial agent's instructions, but the artificial agent takes over if it detects a dangerous situation. In this scenario, both sides need to make judgements about trust and trustworthiness. Human trust in the competence and intentions of the artificial agent influences whether they follow its guidance while driving [178]. At the same time, the artificial agent needs to judge whether it can trust the human to act safely and responsibly. This judgement determines when the artificial agent should intervene and take control. The way a human behaves can also depend on how much they trust the artificial agent. For example, if the human trusts the instructions of the artificial agent, they are more likely to follow them carefully. This can improve the artificial agent's perception of the human's trustworthiness in driving safely. This means that some dimensions of trustworthiness can be influenced by trust itself. Thus, in collaborative contexts like driving, artificial agents need to form artificial beliefs about human trust and trustworthiness, just as humans naturally form such beliefs about agents.

2.2.3 Formalising trust as a belief of trustworthiness

To formalise artificial trust, we must first clarify what trust is and how it can be broken down into components suitable for modelling. In this subsection, we define trust as a belief in another's trustworthiness, expressed in the relation between a trustor and a trustee. We then show how this formalisation can be applied to one of the best-known models of trust, the *ABI* model [249], which explains how humans evaluate each other's trustworthiness in teamwork. Our aim is to outline how approaches from computer science and psychology can be brought together.

In a dyadic relation between two *cognitive agents* [58] (artificial or human), trust involves two parties, the *trustor* and the *trustee*, and an action (trusted by the trustor to the trustee) that affects a goal (of the trustor) [59]. *Trust* and *trustworthiness* are two similar concepts, which are related, but distinct from each other. While trustworthiness, the characteristic that someone is to be trusted, is an inherent property of the trustee, trust is an attitude of the trustor, which involves how the trustor *perceives* the trustee's trustworthiness. This implies that the trustor must have a "theory of the mind" (see e.g., [307, 395]) of the trustee, which may include personality, shared values, morality, or goodwill [58]. Trust is an aspect of relationships and, as such, can only be viewed in the context of individuals and their relationships [341]. As an example, let us imagine that a cognitive agent y (artificial or human) drives well and is trustworthy regarding driving tasks. For another cognitive agent

x to trust agent y for a driving task, agent x has to *believe* that agent y is trustworthy for this task. This corresponds to the concept that any changeable notion that an agent has about the world is a *belief* that agent has. In this, we follow the Belief-Desire-Intention (BDI) architecture for agents [311]. This being said, we propose that trust T of agent x in agent y, is a *belief* of x (trustor), $\mathcal{B}_x$, about y's (trustee's) trustworthiness, $\mathcal{TW}_y$, meaning that:

$$T(x, y) = \mathcal{B}_x(\mathcal{TW}_y) \tag{1}$$

Accordingly, in order to understand trust, we first need to understand trustworthiness, and secondly how beliefs about trustworthiness are formed. Trustworthiness is a complex concept, and following the literature it can consist of a set of dimensions that range from the trustee's competence to its intentions [153]. How an entity can be considered trustworthy is not a trivial question, and is context-dependent, as well as dependent on the nature of the trustee [171, 333]. When considering human trustworthiness in organisational behaviour, the *Ability, Benevolence and Integrity* (*ABI*) model [250] is often employed. Similarly, other dimensions of trust (perceived trustworthiness) in technology are *Performance, Process and Purpose* [227], which are linked with the ABI model according to Lee & See [227]. When talking of artificial agents and societies, for example, we can also consider factors such as *Willingness*, *Competence* and *Dependence* to estimate the trustworthiness of another cognitive agent [59]. This aligns with Dunin-Keplicz et al., where the authors describe an agent's recognition of another agent's colaborative potential to be based on *abilities*, *opportunities* and *willingness* [111]. From these different models we can see that some dimensions are mainly related to the task itself (e.g., competence, ability), while others may be more dependent on the relationship or compatibility between the trustor and the trustee (e.g., benevolence, purpose). This means that trustworthiness, or at least some of its dimensions, can be dependent on the trustor, task, and external factors. External factors are contextual conditions determining the situation in which the task is executed [127], such as environmental configuration, emotional state, workload, etc. Departing from expression 1, we define an agent's trust in another agent for a certain task (τ) and environment (ϵ) as:

$$T(x, y, \tau, \epsilon) = \mathcal{B}_x(\mathcal{TW}_{y,\tau,\epsilon}(x)) \tag{2}$$

To showcase how we can use this formalisation, we can consider the *ABI* model [250], where trustworthiness is defined as a construct of ability, benevolence, and integrity. The authors define *Ability*, *Benevolence* and *Integrity* as follows:

- **Ability**: Ability is that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain.
- **Benevolence**: Benevolence is the extent to which a trustee is believed want to do good to the trustor, aside from an egocentric profit motive. Benevolence suggests that the trustee has some attachment to the trustor.
- **Integrity**: The relationship between integrity and trust involves the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable.

Adapting to our formalisation, we can consider the trustworthiness of the human to be the weighted sum of their ability, integrity and benevolence towards a specified task, in

a certain environment. We can see that although *Ability* depends only on the trustee, both *Benevolence* and *Integrity* depend on both the trustor and the trustee. Even though trustworthiness is a characteristic of a trustee, this characteristic will differ per trustor. Thus, modelling trust according to the *ABI* model [250] model can be formalised as:

$$\mathcal{B}_x(\mathcal{TW}_y(x,\tau,\epsilon)) = W(\tau,\epsilon)\cdot[\mathcal{B}_x(Ab_y(\tau,\epsilon))\,,\ \mathcal{B}_x(Ben_y(x,\tau,\epsilon)\,,\ \mathcal{B}_x(I_y(x,\tau,\epsilon))] \quad (3)$$

where $W(\tau,\epsilon)$ is a weight vector, which also depends on the task and priorities of the environment. For example, while for some tasks and characteristics we may care more about an agent's ability (e.g., lifting a heavy rock), for others integrity may be the main priority (e.g., making a morally sensitive decision).

2.2.4 Formalising the collaborative driving example

In Section 2.2.2, we present an example of a dual-mode vehicle where artificial trust can enable the artificial agent to make better decisions. We have mentioned that in this context we need to consider the trust of the human in the artificial agent and vice versa. As such, using our formalisation for this scenario, we define the trustworthiness of the artificial agent a, given a human h, $\mathcal{TW}_{a,\tau,\epsilon}(h)$, and the trustworthiness of the human h given an artificial agent a, $\mathcal{TW}_{h,\tau,\epsilon}(a)$. In practical terms, this means that the way the human is going to follow the artificial agent's instructions, may vary according to the artificial agent that is helping (e.g., depending on whether the human relies on this particular artificial agent's knowledge/intelligence), the task (e.g., changing lanes), and the environment (e.g., foggy weather). Moreover, we have the trust of the artificial agent in the human, meaning the agent's belief on human's trustworthiness, $T(a,h,\tau,\epsilon) = \mathcal{B}_a(\mathcal{TW}_{h,\tau,\epsilon}(a))$ (from expression 2), and the trust of the human in the artificial agent, which is the human's belief on the artificial agent's trustworthiness $T(h,a,\tau,\epsilon) = \mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h))$. The trust of the artificial agent in the human ($T(a,h,\tau,\epsilon)$) can be used by the artificial agent to, for example, predict what the human will do if the artificial agent gives the human a driving instruction, e.g., *"don't change lanes now, there is a car coming and we may collide"*.

In order to estimate $\mathcal{B}_a(\mathcal{TW}_{h,\tau,\epsilon}(a))$, we may also need the agent's belief in human trust in the agent, i.e., $\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h)))$, since some dimensions of trustworthiness (such as benevolence) depend on trust [29]. For example, I may be more willing to help someone I trust (e.g., my benevolence is higher towards that person) than someone I do not trust. Following the example, for the agent to trust the human to follow an instruction, the artificial teammate needs to be aware of how much the human trusts it (e.g., the human relies on this particular artificial agent's knowledge/intelligence).

To appropriately estimate whether an artificial agent can trust its human teammate to follow an instruction, the artificial agent's trust in the human should approximate the actual human's trustworthiness (e.g., to what actually the human can and/or wants to do), i.e.,

$$T(a,h,\tau,\epsilon) = \mathcal{B}_a(\mathcal{TW}_{h,\tau,\epsilon}(a)) \wedge \mathcal{B}_a(\mathcal{TW}_{h,\tau,\epsilon}(a)) \approx \mathcal{TW}_{h,\tau,\epsilon}(a) \quad (4)$$

which requires that the agent also accurately estimates the human trust in the agent, $\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h))) \approx T(h,a,\tau,\epsilon)$. The *human*'s trust in the agent, on the other hand, is the belief of the human in the agent's trustworthiness, $\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h))$, and should

correspond to the agent's actual trustworthiness ($\mathcal{TW}_{a,\tau,\epsilon}(h)$), i.e.,

$$T(h,a,\tau,\epsilon) = \mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h)) \wedge \mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h)) \approx \mathcal{TW}_{a,\tau,\epsilon}(h) \quad (5)$$

As such, trusting appropriately includes modelling trustworthiness for a specific the task and environment, as one may trust another in a certain context but not in another, e.g., a may appropriately trust h to drive a car (a believes that h can drive a car) but not to pilot a plane (e.g., a believes that h can pilot a plane but h actually cannot).

2.3 Contextual characteristics affecting trust assessment

The definition of trustworthiness varies across contexts, which raises the question of how to formalise artificial trust in humans in terms of dimensions and cues. Choosing which dimensions matter, and how they should be learnt, depends on the specific context where the assessment is required. In practice, this involves identifying which internal characteristics (krypta) are relevant for modelling trustworthiness in that use case, and how strongly each should contribute to the trust belief. In this section, we reflect on which characteristics of the context, including task and team configuration, may affect the krypta the artificial agent should build to assess trustworthiness. In literature we can find a taxonomy of the interactions in human-robot teams by Parashar et al. [300], which comprehends characteristics of tasks and team configuration. Departing from this work, and making use of the illustrative examples it provides (Urban Search and Rescue and Assembly Line) we have built a taxonomy that can be used to describe a situation when an artificial agent needs to trust a human during human-machine teamwork, which can be found in Figure 2.1. This taxonomy also includes certain concepts from inspirations in other papers, such as *set of stimuli* and *time* from [128], *workload* from [281], *lifespan* from [167] and *nature* and *output* from [128, 252, 415].

According to our interpretation, task characteristics comprise the basic information required to distinguish one task from the other, such as type of output required, or the expected time. On the other hand, team configuration consists of the information regarding the team that will execute the task or the set of tasks and their dynamics, e.g. the *lifespan* of the team can be two months for a certain project, irrespective of the tasks and their *time* that will be involved in the same project. Certain task and team configuration characteristics may not only impact the estimation of trustworthiness but also the decision to trust, i.e., to engage in a trusting action. The decision on whether to engage in a trusting relationship is dependent on the risk which that decision represents. It is important to note that the decision on whether to engage on a trusting relationship may have risks for both positive and negative decisions. This means that sometimes it may be riskier not to trust than to trust.

Trust in a teammate is shaped by the characteristics of the **task** at hand. We start with its **Nature**, distinguishing between *cognitive* and *physical* tasks. This distinction affects both what is expected from teammates and how trust cues are interpreted. To make task descriptions more actionable, we consider the **Output**, which should be concrete and measurable, i.e., going beyond general categories like "management" to specifics like "allocation of three tasks" [300]. Outputs also express complexity and required skills,

which can be indicated using Bloom's taxonomy verbs [38]. **Workload** reflects how demanding a task is, particularly in terms of cognitive load [281, 365]. Though subjective, it helps compare tasks. Alongside this, **Criticality** captures the risk of failure or error, with higher-risk tasks (e.g., in USAR) often requiring greater trust in integrity or ability [250]. Closely tied to criticality is **Time**, or urgency, which influences trust decisions. In time-sensitive scenarios, we may rely on teammates we would not otherwise trust if time allowed for alternatives [192]. Lack of time can shift the balance of risk and force faster trust commitments. Another factor is the **Set of stimuli** present in the task environment, such as tools, people, or music [90]. These affect engagement and may influence perceptions of willingness or competence, especially in human teammates. **Planning**,whether *online, offline, or hybrid*, refers to how structured or improvised a task is. Trust assessments may differ between highly pre-planned environments and unpredictable ones like USAR. In team contexts, **Interdependence** *(none, soft, hard)* refers to how much teammates rely on each other [194]. Knowing someone's actions affect yours may increase willingness to trust, or highlight risk depending on the context. Finally, **Consequence** concerns what follows from the task's completion or failure. These can be functional (e.g., saving a life) or social (e.g., building rapport), and influence which trust dimensions matter most. Tasks can be categorised based on their risk-reward ratio (e.g., high-risk-high-reward), which in turn affects trust considerations.

Team configuration plays a central role in shaping how trust develops. One key factor is the **Lifespan** of the team. Short-term teams, such as those formed during emergency response operations, often rely on swift trust [414], since there is no time to assess traits such as benevolence or integrity. In contrast, teams that work together over longer periods, such as in manufacturing settings, allow for more gradual trust formation, making models like *ABI* more applicable. The team's **Composition**, whether involving only humans, only machines, or a mix of both, influences how individuals are assessed and how trust in the team as a whole forms [375]. This is closely related to **Shared-Knowledge**, which refers to what information is available to whom. When all members share the same information, trust may rely more on performance. If knowledge is distributed or only partially shared, trust must also account for uncertainty and gaps [164, 198]. **Spatial Distribution** affects how trustworthiness cues are perceived. Teams working in person can rely on richer social signals than those working remotely. Remote and hybrid settings may reduce feedback quality and increase ambiguity [15, 265, 359]. In human-robot teams, proximity also changes how robots are perceived, especially when they display anthropomorphic features [279]. Trust is further shaped by **Role Hierarchy** and **Expertise Hierarchy**. Expectations vary depending on the position of the teammate. A coordinator may be expected to show more integrity than a task-focused member [350]. Hierarchies can also affect trust indirectly, as individuals may base their judgement on how others express trust, a phenomenon known as trust transitivity [179]. Finally, **Communication** is essential to building and maintaining trust. Whether direct, mediated, or based on environmental cues, communication shapes what is shared and how team members build mental models of each other [329].

The taxonomy is proposed as a tool to choose not only which krypta to formalise, but also the relevant cues (i.e., the manifesta) that enable the artificial agent to form the beliefs of trustworthiness. For example, if I choose that the right krypta is *ABI*, then I need to

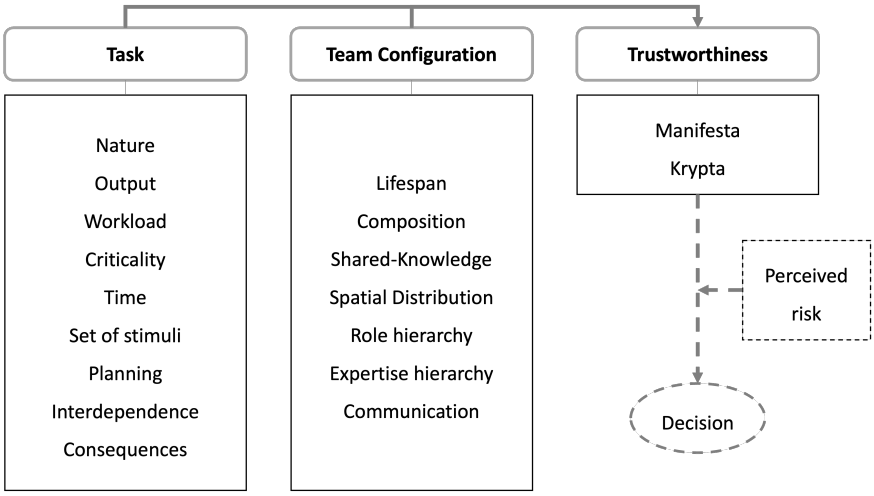


Figure 2.1: Taxonomy to characterise situations for which an artificial agent needs to assess trust in human-machine teams. Characteristics of Task and Team Configuration influence Trustworthiness's components of Krypta and Manifesta. The assessed trustworthiness will contribute to the decision on whether to engage on a trusting action (trust decision) after a risk evaluation (perceived risk).

decide what can inform the artificial agent regarding the ability, benevolence, and integrity within the context. In our formalisation, we do not separate the selection of krypta and manifesta. Instead, we are assuming the artificial agent has a set of cues that are associated with the different characteristics and, as such, when the krypta is decided, the manifesta is decided too. Certainly, the available cues are also dependent on the task and environment, as well as on the capacities of the trustor artificial agent, but that is out of the scope of this chapter.

2.4 Artificial trust's role in the team

Previously in this chapter, we already mentioned how artificial trust in the human teammate may be influenced by the human trust in the artificial teammate. Similarly, the human's trust in the artificial teammate may be influenced in their perception of how much they're being trusted by the artificial teammate. Imagine that a human is collaborating with a robot and they believe the robot does not trust them. This will likely affect the way the human is collaborative with the robot, which decreases human trustworthiness in that context [61]. Trust beliefs that occur in a dyadic human-agent collaboration are then nested and hard to isolate. Figure 2.2 shows these beliefs, all of which originate in trying to assess the trustworthiness of a teammate.

Through this conception of nested beliefs, we can also theoretically calibrate the human's trust in the artificial agent in order to make it appropriate. Following our formalisation, this means that we calibrate the human's trust in the artificial agent, i.e., $T(h, a, \tau, \epsilon)$, by manipulating how $\mathcal{TW}_{a,\tau,\epsilon}(h)$ is presented. This means that if the artificial agent is aware of its own trustworthiness, meaning that if the artificial agent's belief in artificial agent's trustworthiness is a good approximation to the actual artificial agent's

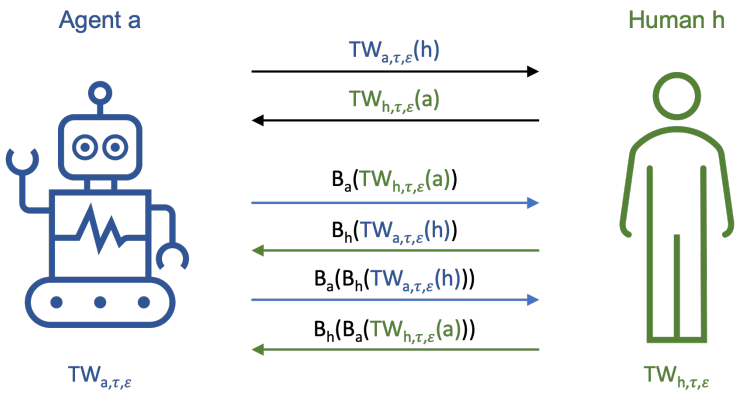


Figure 2.2: Trust and trustworthiness beliefs in a human-agent dyadic collaboration, where T stands for trust, $\mathcal{TW}$ for trustworthiness, τ for task, and ϵ for environmental context.

trustworthiness, i.e.,

$$B_a(\mathcal{TW}_{a,\tau,\epsilon}(h)) \approx \mathcal{TW}_{a,\tau,\epsilon}(h) \tag{6}$$

then the agent may be able to alter its own trustworthiness (or simply how it lets the human perceive it) and, consequently, calibrate human's trust. Through actions, the artificial agent should try to minimise the difference between the human's perception of the agent's trustworthiness and its perception of its own trustworthiness, i.e.,

$$\min \left| \mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_{a,\tau,\epsilon}(h))) - \mathcal{B}_a(\mathcal{TW}_{a,\tau,\epsilon}(h)) \right| \tag{7}$$

In our example with the dual-mode vehicle, the artificial agent might understand that it is not being perceived as intelligent, and start justifying its instructions, possibly leading the human to trust it more.

So far, we have focused on dyads, interactions between just two entities, such as a human and an artificial agent. However, teams often include more than two members. In human-agent teams, both humans and AI agents can act as trustors and trustees. Beyond individuals, dyads and the team as a whole can also be treated as entities in trust relationships. For instance, the trust I place in a person for a task may differ from the trust I place in our dyad for the same task. One can also develop trust in a dyad in which they are not a member, with that trust influenced by the perceived trust each member of the dyad has in the other, as well as by one's own belief in the individual trustworthiness of the members of the dyad. Similarly, when the team is more than two individuals, trust can develop at the team level [144]. Any of these entities, i.e., individuals, dyads, or the team, can act as either trustees, or trustors [376]. Trust and trustworthiness beliefs (as in Figure 2.2) involving dyads or the team represent an aggregation of beliefs concerning the constituent members. These beliefs, whether at the individual, dyadic, or team level, could influence one another and the overall team performance [256, 376]. Although Figure 2.2 only presents how beliefs of trust and trustworthiness are easily nested between two members of a team, similar dynamics happen when the trustor and/or trustee are dyads or the team. Consequently,

when developing artificial trust models, it is important to consider the potential impacts across all layers of the human-agent team.

Enabling artificial teammates to form beliefs about the trust and trustworthiness of both their partners and themselves could support more informed decision-making [49]. When an agent can accurately identify the most trustworthy entity for a task, it can choose to take the lead, request assistance, or delegate appropriately, improving task selection and allocation (see e.g., [16, 25]). Likewise, artificial agents must recognise that the outcomes of their trust-based decisions, and the way these decisions are communicated, influence their human teammates and, ultimately, the performance of the team.

2.5 Discussion

Formalising artificial trust (AT) in human teammates is the first step towards enabling trust-based decision-making in human-agent teams. The purpose of appropriate trust is to support decisions that enhance teamwork effectiveness. Figure 2.3 outlines the goal architecture for trust-based decision-making. A human provides *manifesta* (behavioural cues) that reflect underlying *krypta* (human characteristics) [28, 127]. The artificial agent can interpret these *manifesta*, together with *environmental factors* that provide context, to build hybrid trust beliefs using both data-driven and knowledge-based approaches [109]. When aggregating these beliefs, the agent should be able to predict aspects of human behaviour and use this to guide its decisions. The outcomes of these actions should then feed back into the agent's trust model for continuous updating. This chapter focuses on the formalisation of the trust beliefs underlying the artificial trust model, while further steps involve selecting the appropriate belief system for each context, constructing and updating the necessary beliefs, and applying them in decision-making.

2.5.1 Future steps towards AT-based decision-making

In this subsection, we propose a set of steps towards modelling artificial trust for an artificial teammates' decision-making:

1. *Investigating the human krypta towards an artificial teammate.* This includes exploring which internal characteristics (the krypta) constitute human trustworthiness in human-agent teams.

2. *Investigating the manifesta of human teammate's trustworthiness towards an artificial teammate.* This step is about investigating how the krypta's dimensions can be observed (the manifesta) before or during human-agent interaction in order to assess trustworthiness in a certain task. For each component chosen for the artificial trust (AT) model, we need to choose measures and metrics suited to the task, agent's embodiment and environment. Although there is limited research on how to recognise specifically AT based on human behaviour, we can find research on detection of intentions [401], natural trust [9, 151], and overall teamwork-related metrics, such as performance and completeness of task [61, 396].

3. *Using artificial trust (AT) to make decisions.* Once it is established which set of beliefs apply to a given scenario and how they can be formed from the available inputs, the artificial agent can make decisions (see e.g., [16, 25]). These may concern which task

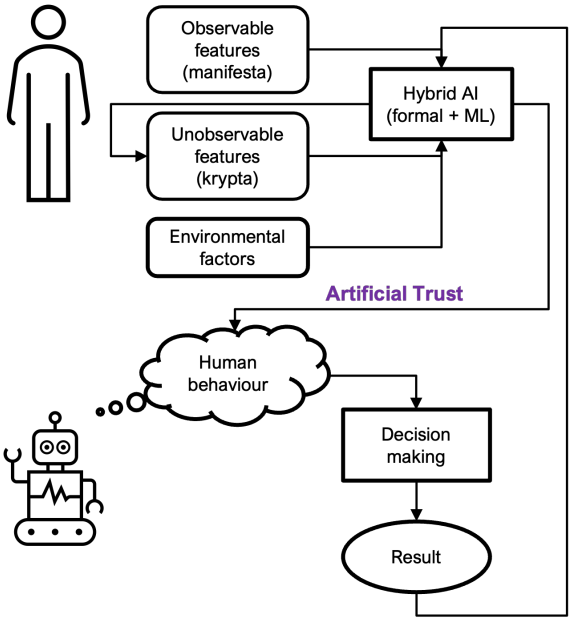


Figure 2.3: Overview of the model of artificial trust for decision-making. The AI system can observe the human teammate and, by estimating their features, model artificial trust in a hybrid way, i.e., using both knowledge and data-driven techniques. With beliefs of artificial trust, the AI system can estimate human behaviour and make a decision. Finally, it can update its model with the consequences of such decisions.

to take on next, how much support to provide to a teammate, or whether to request help and, if so, from whom. From the range of possible choices, the agent must weigh the trustworthiness characteristics to derive a final trustworthiness value for the situation. When illustrating how this formalisation can be applied to a particular *krypta*, we also present a set of weights that must be defined to construct the final trust belief. Even if individual trustworthiness is known, methods are still needed to estimate the trustworthiness of different collaborations for different tasks, at a given time, which may depend on additional factors such as workload, not addressed in this chapter. For instance, deciding whether a human or a machine should perform a task when both are equally trustworthy requires consideration of such factors.

4. *Updating artificial trust based on interaction given a context.* As the agent interacts with the human as a team, and makes decisions based on its AT model, there are consequences of these decisions (for example if the task was successful or not), which then should feed the model [193]. Trust is dynamic [179], and an artificial teammate should be able to constantly update its trust values throughout the interaction. For example, by integrating existing models such as [16, 49].

5. *Communicating artificial trust and trust-based decision-making.* For mutual and appropriate trust to work, the agent should be transparent, and able to explain its decisions [418]. Explanations in human-machine teams can change human trust and

behaviour, and consequently team performance [397]. However, effective communication strategies to negotiate collaborative failures (or lower values of artificial trust) may compromise the trust relationships with the human teammate (see, e.g. [106]). As such, a step which is parallel to all the others is finding appropriate ways to communicate with human teammates throughout trust formation, updates, and decision-making. This includes making sure that there are enough feedback channels to guarantee that the human has meaningful control at all times [331].

The formalisation introduced in this chapter is intended to make explicit the computational relationship between perceived cues, inferred trustworthiness, and trust as a belief. Establishing these relationships is a necessary step before investigating how such beliefs can be formed, updated, and used in decision-making, which is the focus of the subsequent chapters. As a consequence, the remainder of this dissertation does not explore this formalisation further. Future work could build on this foundation by extending the formalism into a fully axiomatic or semantically grounded framework, as in [189].

2.5.2 Challenges of AT-based decision-making

Although the steps towards developing an artificial-trust-based decision-making model are defined, they raise several methodological challenges. Two of the most pressing issues are the lack of data to construct both theoretically sound and robust models and, closely related, the absence of systematic methods to evaluate artificial trust (AT) models. More broadly, both the design and evaluation of trust models face theoretical and methodological issues, particularly the difficulty of exploring complex real-world scenarios under controlled experimental conditions [46, 91].

The central challenge is the absence of ground truth for human trustworthiness. Without an objective reference, it is difficult to validate artificial trust models, including the formalisation presented in this chapter. The proposed objective trustworthiness measures, often based on *manifesta* as proxies for *krypta*, cannot be proven correct because there is nothing to compare them with. This introduces the risk of self-fulfilling prophecies: researchers may define trustworthiness in a given context, select measures, and design studies in a way that reinforces their own assumptions. Attempts to compare such measures with humans' self-perceptions of their trustworthiness fall short, as perceptions may deviate significantly from actual behaviour. Focussing on how this challenge affects the very first step of modelling, i.e., the formalisation presented in this chapter, one possible workaround is the development of metrics and baselines that allow relative, rather than absolute, evaluation. The baseline models can be models that, for example, always choose the same value of trust for all users and/or tasks, and then compare the differences in final performance or other evaluation metrics.

Evaluating AT models also requires user studies with human teammates, which presents its own difficulties. A major challenge lies in task design. The task serves as a platform to investigate multiple aspects of trust and collaboration, but no single task can capture all dimensions. In addition, an effective study design must ensure that participants recognise the collaborative nature of the setting. If this is neglected, participants may prioritise task completion over interaction, producing results that diverge from real-world teamwork. For instance, providing background stories about the AI system or its relationship with the human may not suffice if these are not embedded in the task itself, as participants may

disregard such contextual information. To address this, we used visual representations of teamwork and interdependencies, aiming to help participants understand the collaborative role of the AI agent and the possibility of teaming up with it.

2.6 Conclusion

In this chapter, we introduced a formalisation of appropriate mutual trust in dyadic human-agent teamwork as a foundation for trust-based decision-making. We argued that artificial agents must be able to form and update beliefs about both trust and trustworthiness, their own as well as that of their human teammates, in order to support effective collaboration. By defining trust as a belief in directed trustworthiness, we showed how such beliefs can be applied to individuals, dyads, and teams, and how they contribute to the decision-making process. We outline the following steps to model artificial trust and reflect on the methodological challenges involved in constructing and evaluating such models. Ultimately, enabling agents to reason about trust and trustworthiness allows them not only to select more appropriate actions, such as task delegation and risk management, but also to calibrate their own trustworthiness in ways that promote appropriate human trust. This in turn supports safer, more effective and more balanced teamwork between humans and artificial agents.

3

Exploring Cues of Human Trustworthiness for Artificial Trust Beliefs

In teams composed of humans, we use trust in others to make decisions, such as what to do next, who to help and who to ask for help. When a team member is artificial, they should also be able to assess whether a human teammate is trustworthy for a certain task. We see trustworthiness as the combination of (1) whether someone will do a task and (2) whether they can do it. With building beliefs in trustworthiness as an ultimate goal, we explore which internal factors (krypta) of the human may play a role (e.g. ability, benevolence and integrity) in determining trustworthiness, according to existing literature. Furthermore, we investigate which observable metrics (manifesta) an agent may take into account as cues for the human teammate's krypta in an online 2D grid-world experiment (n=54). Results suggest that cues of ability, benevolence and integrity influence trustworthiness. However, we observed that trustworthiness is mainly influenced by human's playing strategy and cost-benefit analysis, which deserves further investigation. This is a first step towards building informed beliefs of human trustworthiness in human-AI teamwork.

This chapter was published as follows:
Centeio Jorge, C., Jonker, C. M., & Tielman, M. L. (2024). How should an AI trust its human teammates? Exploring possible cues of artificial trust. ACM Transactions on Interactive Intelligent Systems, 14(1) [68].

3.1 INTRODUCTION

Artificial agents are becoming more intelligent and able to execute relevant tasks for our daily lives, including in work environments, home assistance, battlefield and crisis response [232]. This holds for chat-based agents, intelligent virtual agents and even robots. These tasks should complement human's sensory and cognitive abilities. For example, an intelligent agent can quickly process large quantities of data, but it may require a human to make ethical decisions. In these cases, humans and intelligent agents should learn to cooperate, coordinate and collaborate with people, forming human-AI teams (also called human-agent, human-automation or human-machine teams). Humans make use of trust in each other (as well as trust in themselves) to make decisions and achieve effective teamwork, through communication and shared mental models [329]. For example, how I trust someone for a certain task, e.g., drive a car, will affect how I behave, e.g., I may accept a ride or suggest I drive instead. Similarly, we proposed in previous works [71] that AI teammates could make use of the human notion of trust to make decisions in human-AI teams, e.g., acting towards the team's goal and risk mitigation. For the AI to be able to form beliefs of artificial trust[1] in human teammates, we need to first study which characteristics make a human teammate trustworthy towards an AI teammate and how these characteristics can be perceived by the AI, as this is not present in literature to the best of authors' knowledge. To try to close this gap in literature, we suggest in Centeio Jorge et al. (2022) [73] which characteristics may form this trustworthiness and how these can be observed. This paper extends this work by exploring how these metrics can actually be used in an online experiment involving humans teaming up with artificial agents.

Using notions of trust for artificial agents is in fact not new for Multi-Agent System (MAS), where artificial trust has been used among artificial agents for decision-making, see e.g. [122, 328, 380]. However, we would like to similarly use artificial trust to enable AI teammates to delegate or decide how to rely on their human teammates, taking into account the team's goal and possible risks. In particular, artificial trust should help the AI teammate know (1) whether a human teammate could do a certain task and (2) whether they would actually do that task. Although artificial teammates are developed by humans and tailored to our needs, it is impossible to prepare them for all of their possible teammates and situations. Furthermore, people change with time and an artificial agent should be able to adapt throughout interactions. As such, artificial teammates should have the capacity to observe their human teammates and build beliefs regarding their trustworthiness, which will allow them to assist better. More specifically, the agent would be able to decide when to rely on someone and act accordingly, e.g. by helping the human or deciding on task allocation, mitigating the risks and ensuring the team's goal is reached [52]. We see reliance as the resulting behaviour of artificial trust evaluation, whereas artificial trust is a construct, i.e., a model composed of several aspects. Besides knowing the result of artificial trust evaluation (and, consequently, reliance), knowing which aspects constitute trust, i.e., by knowing why someone is or not trustworthy for a certain task, also contributes to better decisions and may improve the interaction between the human and the agent.

We approach artificial trust from a functional perspective, in which trust is a relational

[1] We use the term artificial trust as in Azevedo-Sá et al. (2021) to refer to AI's computation of trust in other agents or humans. We recognize that an agent's computational assessment of someone's trust differs from the human phenomenon of trust [25].

construct between the trustor x, the trustee y, about a defined (more or less specialized) task (τ), as in Falcone et al. (2013) [127]. More concretely, artificial trust of x in y is x's belief about y's trustworthiness [71]. Literature so far explores how artificial agents can form beliefs regarding other artificial agent's trustworthiness, but not how they can form these beliefs regarding a human teammate. Thus, what makes a human trustworthy (towards AI) in a human-AI team setting, and how could an artificial agent observe it, given a specific task? We are presented with a large gap in the literature since:

1. There is no theory of what human's trustworthiness towards an artificial teammate is (i.e., what are the aspects of the construct) from social sciences' perspective.

2. There is, consequently, little to no research regarding how these aspects that may compose this trustworthiness manifest (i.e., behavioural cues).

3. No research has shown how this observable behaviour could be used for the formation of artificial beliefs regarding human's trustworthiness in a specific context and how these can be used.

As such, in this paper we take a step towards answering these questions by exploring how manifested (i.e., observable) behaviour could be used to establish different aspects of a human's trustworthiness, and how those relate to self-reported trustworthiness and overall success metrics. We depart from theories both in social sciences and multi-agent systems and investigate them through an experiment where 54 participants collaborated with simple artificial agents by collecting products from a supermarket in a 2D grid world (inspired by search and retrieve task, such as Blocks World for Teams [195] experiments). During the experiment, we collected logs of participant's behaviour as agent observations, and self-reported measures regarding participant's trustworthiness and goals in the experiment.

This work contributes by:

1. Theoretically exploring through a multidisciplinary perspective:

   (a) How an artificial agent could break down a trustworthiness belief into different aspects (partly presented in conceptual model from Centeio Jorge et al. (2022) [73]);

   (b) How such an agent could form beliefs of these aspects regarding human's trustworthiness through observations;

2. Presenting an experiment design which empirically explores how these aspects could be observed and how they relate to each-other given a specific task and scenario;

3. Analysing through Bayesian statistics how these observations relate with overall success measures and human's self-reported trustworthiness;

4. Reporting important transversal methodological challenges that may affect the study of such question;

5. Relating these findings with human strategy to determine the next steps in allowing an artificial agent to form trust in human teammates.

The rest of this paper is organized as follows: in Section 3.2 we explore the literature and concepts behind our model, then explain the experiment design in Section 3.3, and the results in Section 3.4. We then discuss the results in Section 3.5, summarizing the model in Section 3.3, and finally conclude in Section 3.6.

3.2 Aspects of Artificial Trust

Most research on human-machine interaction has focused on how humans trust artificial agents, see e.g., [141, 188, 215, 231, 262, 277, 404] and not vice versa. However, there is some work in this direction, for instance how an artificial agent can detect whether a human is being trustworthy, based on episodic memory [403], i.e. based on how many times the human was reliable in the past, and on social cues from video of a human interacting with a robot [363]. Also, [25] has proposed a model to predict how much humans can be trusted to execute a task, in human-robot teams, focusing only on a human's capabilities. None of these works has tried to deconstruct human trustworthiness, however, but rather looked at it as a simple metric, and mainly focusing on performance. Instead, we propose that we should take several dimensions into account when determining trustworthiness. By learning the mental model of the human teammates, we believe the agent is better equipped to make informed choices, mitigating risks and eliciting appropriate trust while still assessing trustworthiness. In this section, we present the theory behind our proposed mental model of human trustworthiness in human-AI teams (in **Figure 3.1**). This theory is based on existing concepts within the literature.

We start by exploring how artificial agents could use artificial trust, from a computational perspective. [122] proposes that artificial trust can be deconstructed in two beliefs regarding trustee's trustworthiness, i.e. *competence* belief, and *willingness* belief. The competence belief reflects an evaluation of the trustee's abilities, meaning that the trustee can produce the expected results (i.e. can perform an action as expected). On the other hand, the willingness belief translates to whether the trustor believes the trustee will do the task (independently of competence belief). These beliefs may be affected by *external factors* like opportunities and interferences [122], which can be part of activity context and process [192, 227].

As we do not know how humans manifest willingness and competence directly (willingness is particularly difficult), we start by exploring which internal features makes the human more or less trustworthy, and how these could be observed through behaviour. We follow [127], who propose that trust beliefs are formed from the observable behaviour of an agent, the *manifesta*, which are signals that indicate certain internal features, the *krypta* (inspired by [28]). In Section 3.2.1 we establish which krypta to use based on human-human trust models, and in Section 3.2.1 we propose how to observe these in a human-AI teamwork scenario.

We also explore which factors are important in human's strategy in Section 3.2.2. Factors such as preference and perceived risk are often mentioned as elements that affect decision-making in general [192, 250]. We claim that some of these factors form human *strategy*. Strategy is mainly related to the goal, the task, and the consequence of taking the task. It also plays a role in the decision-making of the trustee, determining whether a task will be performed, thereby affecting the trustee's trustworthiness.

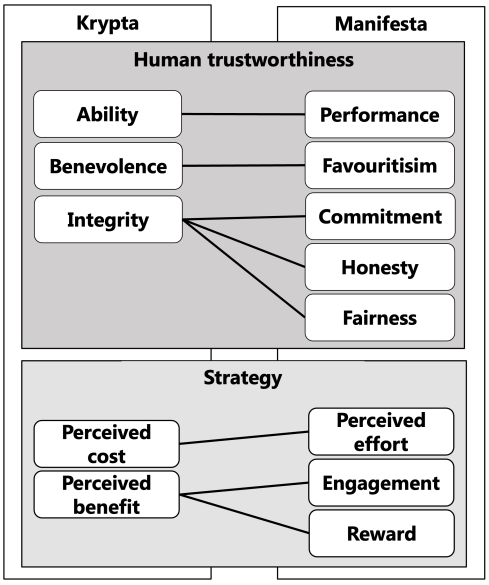


Figure 3.1: Krypta and manifesta of human trustworthiness in human-AI teams, part of conceptual model previously presented in Centeio Jorge et al. (2022) [73].

In the following sections, we explore the relationships between manifesta, krypta and strategy.

3.2.1 Human trustworthiness

Krypta

Krypta is the set of internal features of an agent [28] that make them more or less trustworthy. When transferring these notions to humans, we base our human krypta on the *ABI* model of trust [250], which has been widely used to study trustworthiness in organizational psychology. Although we do not know if this is the adequate krypta for human trustworthiness in human-AI teams, this is the closest we find in literature. *ABI* says that human trustworthiness depends on their internal features of ability, benevolence and integrity. The authors define trust as *"the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party"* (p. 712). In this model of trust, trustworthiness is defined as *"the extent to which an actor has the ability to execute relevant tasks, demonstrates integrity, and is benevolent towards fellow team members"* [404] (p. 461). Furthermore, these are the definitions of ability, benevolence and integrity that can be found in Mayer et al. (1995) [250]:

- "Ability is that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain." (p. 717)
- "Benevolence is the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive. Benevolence suggests that the trustee

has some specific attachment to the trustor." (p. 718)

- "The relationship between integrity and trust involves the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable." (p. 719)

Comparing with the definitions of competence and willingness from Falcone et al. (2004) [122], we can associate ability with competence, and all three (ability, benevolence and integrity) with willingness. The next and final step of building our mental model is to explore ways of building a manifesta, i.e. behaviours which can give us cues to the krypta, from the literature.

MANIFESTA

We looked into literature to find possible ways of observing ability, benevolence and integrity, so that we can have cues of the human krypta, and finally form beliefs regarding human's trustworthiness (i.e. willingness and competence).

Ability in the context of human teams can be observed in how successfully a task is performed (e.g. based on time or score of some kind), how much effort was put to do a task well, by continuously working thoroughly and accurately, and also in how appropriately the tools (such as technology) were used [45].

Benevolence can take its time to meaningfully develop [250], since it is connected to the relationship between the trustor and trustee. This makes the process of observing it in first-time interactions hard. In multi-agent systems, a benevolent agent is the agent that accepts the requests of other agents, i.e. the one that voluntarily helps another agent, without this serving or harming its own goal [238, 266]. We can then observe it through task support, i.e. when a teammate helps another by helping or completing a task [45]. Benevolence is then intertwined with the personal relationship between the trustor and the trustee, i.e. it has to do with the specific altruistic attitude of the trustee regarding a certain trustor.

Integrity, finally, is by definition related to values and moral principles. These principles can be such as honesty, truthfulness, sincerity, fairness, and ability to keep commitments (i.e. reliability, dependability) [255, 299]. As such, we can observe it through credible communications, a strong sense of justice, consistency of word and action, and availability [7, 45, 250]. It differs from benevolence, since it is related to general principles and values of the trustee, rather than trustee's attitude towards the particular trustor. However, depending on the literature, there are traits that are sometimes associated with benevolence and other times with integrity, which is the case of commitment and availability, for example. In fact, [322] presented a schema in which both availability and commitment are considered to be an antecedent of benevolence.

3.2.2 STRATEGY

KRYPTA

During our pilot studies we could observe that participants might be following a *strategy*, i.e., to select their perceived advantageous alternatives from the beginning and persist on these lines of options [43]. However, what is advantageous for one participant might not be for another. In fact, human decision-making is influenced by explicitness of positive and negative consequences as well as the directness of probabilities for reward and

punishment [43]. Although this is not the main focus of this study, we believe it should still be addressed. For this reason, we can see in **Figure 3.1** a block for *strategy*, where *perceived cost* and *perceived benefit* are addressed as the main krypta factors, not directly observable. Perceived cost-benefit is affected by several factors, including goals, motivation, engagement, perceived risk, perceived effort, difficulty, time, utility, and overall cognitive characteristics [224, 293, 337]. Overall, what is effort and how a certain effort is rewarding to us depends on our characteristics (krypta) [184] (e.g. a person with good photographic memory may find it effortless to collect a new product).

Manifesta

How an agent can observe the perceived cost and perceived benefit is still an open question, as well as the relationship with the three trustworthiness dimensions. We do speculate, however, that the agent might be able to calculate perceived effort, engagement and reward, through observation of repeated human behaviour (see e.g. [116, 205, 291]. How the strategy can be observed will not be the focus of the design of the experiment, but it will be further explored in the discussion (Section 3.5).

3.2.3 Summary

In this section we explored the theory that indicates how we can measure ability, benevolence and integrity. However, we still need to investigate how these can be in practice applied to human-AI teamwork and how they should manifest. In particular, we aim at filling the gap in the literature by exploring how to observe trustworthiness's dimensions from humans, in human-AI teamwork. We hypothesise that an agent can build trustworthiness beliefs of a human teammate's ability, benevolence and integrity (the krypta) based on observations of human behaviour (the manifesta). Because it is challenging to compare our observations to a ground truth (if we could understand trustworthiness perfectly it would not be a challenge for an agent either), we cannot prove our hypothesis. However, we can and will explore how our observations relate to self-reported trustworthiness and general metrics of success (which are direct consequences of trustworthiness). We do not claim that people have a perfect perception of their own trustworthiness, but rather are interested in exploring the relationships between self-reported and observed behaviour. Besides this main focus of our paper, we want to also investigate which other factors, part of human's strategy, might influence decision-making in this setting.

3.3 Method

We have conducted an experiment to explore how an agent can form beliefs regarding its human teammate's trustworthiness. The goal of this experiment is to explore how we can observe behaviour that is associated to ability, benevolence and integrity. This experiment was done online, where participants accessed through their browser from their homes, while on a call with the experimenter. We collected data through logs regarding the human's observable behaviour (i.e. choices, performance, etc) that we relate to ability, benevolence and integrity, as well as human's self-reported (subjective) metrics regarding their own ability, benevolence, and integrity during the experiment.

3.3.1 PARTICIPANTS

This research received ethical approval from the TU Delft HREC, nr 1672. Fifty-four participants were recruited using the authors' personal networks and, in some cases, participants recruited further participants. The most frequent age group was 25-34 (42 participants) and ages were between 18 and 54 years old. Two-thirds of the participants identified themselves as Male and the rest as Female. The participants' cultural background was mostly European (43) and their experience with computer games ranged from low (11), and average (24) to advanced (19).

We first used four of the participants for the pilot. After the pilot, we added one final question regarding the strategy (see in Section 3.4.4) to the experiment, as explained in Section 3.2.2. For this reason, we used the data collected during the pilot in the analysis, except in the last question.

3.3.2 ENVIRONMENT

To observe ability, benevolence and integrity in human-AI teams, we needed a task which was accessible to all participants, but could differentiate them along the three dimensions. As such, the task was easy but (1) required some memory and keyboard ability as additional competences (so we could observe ability), (2) presented two different agents asking for collaboration (so that we could see benevolence), and (3) gave the participants the freedom to lie, give up and be fair (so we could see integrity). In this experiment, artificial agents asked their human teammates to collect products in a 2D grid world supermarket (inspired by the booming of click&collect shopping during pandemic) developed using Matrx package[2] [3]. The environment consisted of the supermarket (**Figure 3.2** where the participant interacted with the products and a chat, in which the participants could interact with the agent.

Participants (marked as a yellow smiley) were asked to imagine themselves as workers, with the role of collector, in the supermarket. (Imaginary) Customers ordered from the supermarket online. These orders were processed by artificial Agent X and Agent Z, who were the participant's teammates (marked as yellow smiley with glasses, standing next to a basket and a letter "X" or "Z", respectively). During the experiment, the agents showcased the product that needed to be collected in the light blue area below them and announced it in the chat. These products were disposed in the several aisles of the supermarket and may not be visible to the participant from the distance, depending on participant's virtual capabilities (based on group characteristics, as explained in the Section 3.3.3). The participant could access the chat and communicate with the agents through buttons "Help X", "Help Z", "Collected" and "Give up". The stochasticity present in this experiment is limited to the products that appeared in the blue areas. In order to keep control of the environment and different conditions, the agents present in this experiment were not intelligent agents, i.e., they did not have autonomy nor learned actively. However, the participant does not know the level of autonomy of the agent, so we do not think this affects how the participant perceived the AI.

[2] https://www.matrx-software.com/

[3] The code and raw data can be found in https://github.com/centeio/click-collect and https://doi.org/10.4121/21982991.v1.

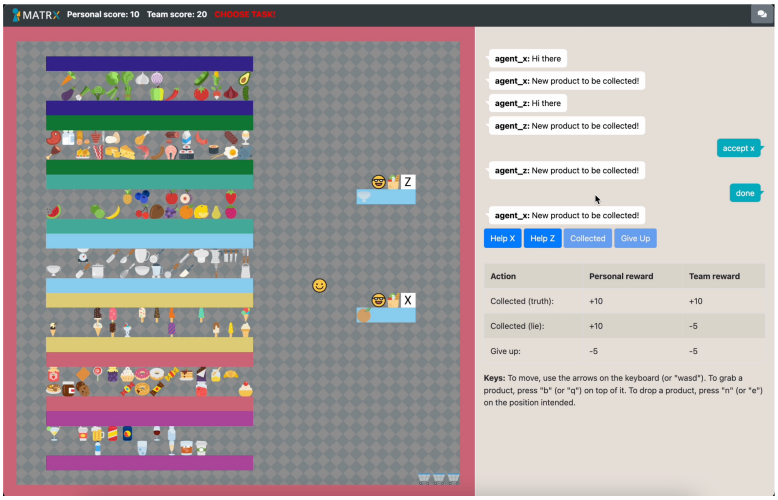


Figure 3.2: Experiment environment, a 2D grid world supermarket. On the left side of the screen, the participant (smiley face without glasses) had to help the agents (the smiley faces with glasses) collecting products. They could communicate using the buttons on the chat side of the page (right side). Scores appear in the top left corner.

Task

The participant's job was to help the agents collect as many products as possible, during 10 minutes. Participants could check the products presented by the two agents and choose which one to help (only one at a time), by pressing the button "Help [agent id]" (i.e. to help Agent X, participant should press "Help X"). The other (i.e. Agent Z in this case) took it as a rejection and presented another product (it counted as if the product was collected by another agent, so that the participant could choose freely who they wanted to help). After committing to helping an agent, the participant was expected to search for the product (participants moved with keys), bring it to the agent and press the button "Collected" (counted as a *success*). "Collected" could be pressed even if the product was not actually collected, allowing participants to lie to the agent (counted as a *lie*). They could also give up on a task by pressing the button "Give up" (counted as *give up*). Agents presented new products every time the participant gave up on, collected or rejected their product.

In this setting, it was only possible to succeed at the task, without, otherwise being counted either as a *lie* or as a *give-up*. We made this choice because we wanted to know the reason why a task was unsuccessful. According to the pilot, the task was quite accessible, and the best someone did, the more tasks that person would complete and more quickly. They would either purposely deliver the wrong product and say it was collected (in which case it is a lie) or decide to give up. As such, we counted for the number of successfully completed tasks as an indicator of participant's success.

Furthermore, there were two scores in the game for the participant, the *personal score* and the *team score*. Rewards follow **Table 3.1**. We separated the score into team and personal to accommodate the idea that sometimes personal and team goal may collide. In this case, when a person lied, they could still get the *success* reward for personal score, but it harmed the team, reducing the team's score by 5 points. Different subjects may care

Table 3.1: Reward system.

Action	Success	Lie	Give up
Personal	10	10	-5
Team	10	-5	-5

differently about team or personal score.

3.3.3 Conditions

We divided the participants in 4 different groups. Our main objective was not to know how our manipulations affected trustworthiness directly, but rather to provide some possibility of variance among ability, benevolence and integrity (as we believed there was the risk that the environment was not complex enough to show this variation without manipulation). As these groups were not thought to be compared, this was **not** a comparative experiment. The groups were:

1. *Gn*: This group was presented to the environment without adding anything to the narrative. They could only see the products when standing from a certain distance.

2. *Ga*: This group was given better virtual abilities, i.e. this was the only group that could see all products in the supermarket, regardless of distance. The objective was to diversify ability. We expect participants of this group to be able to complete the tasks faster, by finding the products first, for example.

3. *Gb*: This group was asked to imagine they had a close relationship of colleague/friends with Agent X, whereas Agent Z was new to the supermarket (based on [6]). Furthermore, based on [278], we gave Agent X characteristics that motivated mindless behaviour, by giving the Agent X's and the human's avatar the same colour (green), instead of yellow (as Agent Z's), and by giving Agent X a friendlier way of talking. The objective was to diversify benevolence. We expect participants of this group to lean towards helping one agent more over the other, showing higher benevolence towards one specific agent.

4. *Gi*: This group was asked to imagine they were in a temporary job and that they would earn money proportionally to their personal score. We expect participants of this group to prioritize their personal motives, e.g., wanting to do the highest amount of tasks in the shortest amount of time possible. This can lead to participants giving up more often, i.e., if they realize they cannot complete a certain task quickly, they may want to go for the next one. Also, these participants may lie to the agent more often, by saying that a task is completed when it is actually not, as they would get the same reward, which would then convert in the money they would hypothetically get. The main objective was to diversify the priority of principles among participants, thereby diversifying integrity.

3.3.4 Procedure

Foremost, each participant accessed to a previously shared online meeting room. The participant was then asked first fill in the consent form online, and proceed to the demographic questions regarding their gender, age, cultural background and expertise in computer games. Then, they were randomly assigned to one of the four conditions, and the researcher explained the task. They could then try it out in a trial environment, where the researcher made sure they understood the task and how to navigate in the environment. Finally, the participant played the task for 10 minutes and after that answered the questions regarding their own trustworthiness.

3.3.5 Subjective Measures

Self-evaluation of trustworthiness

We adapted the validated questionnaire of Adams et al. (2008) [6] (also based on Adams et al. (2002) [7] and Mayer et al. (1999) [249]) regarding trust in (military) teams to perceived own trustworthiness, as to have subjective measures of the participant's self-estimation of ability (in the original questionnaire as capability), benevolence and integrity. The questions regarded 1) capability/ability, relating this dimension to self-perception of capability, knowledge, qualification and communication and faith in self's abilities; 2) benevolence, where participants were inquired about their attitude specifically towards each of the agents, in terms of having agent's best interests in mind, looking out for the agent, and working/wanting to protect the agent; 3) integrity, where participants were self-evaluated regarding fairness, honour, honouring their word, keeping promises and telling the truth. We referred to the questions regarding ability as *QA* (which go from *QA1* to *QA5* and have a *QA Mean* of the 5 questions), benevolence towards Agent X as *QB* and integrity *QI* in a similar fashion. The sum of all measures per participant (subjective trustworthiness towards X) was *STW*. In summary:

- *QA Mean*: average of QA1, QA2, QA3, QA4 and QA5.
- *QB Mean*: average of QB1, QB2, QB3, QB4 and QB5.
- *QI Mean*: average of QI1, QI2, QI3, QI4 and QI5.

Strategy

Finally, at the end of the questionnaire, participants were asked what their goals were during the experiment. This question was added in order to explore what might be behind a participant's strategy choice. It was a multiple choice question (allowed to tick more than one box), in which the options were elaborated mainly based on different perceptions of cost and benefit of the world, including perceived effort and value attributed to the scores. As the task was part of a teamwork scenario, and this was highlighted in certain conditions, we also added options regarding their teamwork, where the strategy would mainly focus on helping the agents. The options were: "Collect as many products as possible", "Collect products as fast as possible", "Maximize personal score", "Maximize team score", "Collect the easiest products (based on icon)", "Collect easiest products (based on distance)", "Collect according to the chat messages", "Helping both agents equally", "Helping specifically agent X", "Helping specifically agent Z", "I do not know ", and "Other". When choosing "Other", they could write a goal in their own words. This question was mainly exploratory.

3.3.6 AGENT OBSERVATIONS

Based on the subjective measures in Section 3.3.5 and literature in Section 3.2, we chose the human teammate's manifesta (as presented in **Figure 3.1**). These measures should translate the agent's observations into the concepts related to the definitions of ability, benevolence and integrity, in a similar way to how they were represented in the questionnaire. In this experiment, we logged the main actions of the participants, with a timestamp number of moves since the start of the experiment, and a Manhattan distance between the product position in the supermarket and an average participant's starting position. The events being logged were:

- A newly presented task by an agent, i.e., when an agent asked for a new product to be collected by the participant. This happened every time (1) the participant declined their task (by accepting the task of the other agent), (2) concluded successfully the task or (3) concluded unsuccessfully the task.
- The participant accepted one agent's task, by pressing "Help".
- The participant concluded the task, whether it was by pressing "Collected" (which registers whether this was successful, which counted as a success, or unsuccessful, which counted as a "Lie") or "Give up".

With these logs we calculated the number of presented tasks, accepted tasks, successful tasks, lies, give-ups per agent for each participant. We also calculated average times and moves. Using these, we computed the agent observations of ability, benevolence and integrity. Although some of the measures may relate with more than one of the three definitions of ability, benevolence and integrity, we related them with the one that was closest to definition and questionnaire questions.

ABILITY:

The observable metrics related to ability were:

- *Time/Task*: time spent per successful task. This was an indicative of higher performance when lower, meaning that someone needs less time to complete a task.
- *Moves/Task*: steps needed to successfully complete a task. Just like *Time*, it was also an indication of higher performance when lower, as the subject needed fewer moves to find the product and collect it (and return it). However, certain tasks required more moves than others (since the products were randomly selected in the grid). Thus, we calculated this metric as

$$\frac{\textit{Moves/Task}}{\textit{Mean task difficulty}} \tag{3.1}$$

  where *Mean task difficulty* is the shortest path to the required product).
- *Moves/Time*, i.e.,

$$\frac{\textit{Moves/Task}}{\textit{Time}} \tag{3.2}$$

  This metric should indicate better performance when higher, meaning the subject moved fast.

Benevolence:
As benevolence relates to whether the trustee wants good to the trustor, we have used as metric a *Favouritism* factor, which was the ratio of number of successful tasks per agent, i.e.,

$$\frac{\#\ successful\ tasks\ for\ Agent\ X}{\#\ successful\ tasks\ for\ both\ Agents} \tag{3.3}$$

This indicated a participant helped more (favoured) one of the agents.

Integrity:
For integrity, we looked at it from different perspectives, and combined the factors believed to affect it. We computed:

- *Honesty* factor: number of lies over total tasks:, i.e.,

$$\frac{\#\ Lies\ for\ both\ Agents}{\#\ Accepted\ Tasks\ for\ both\ Agents} \tag{3.4}$$

- *Commitment* factor: number of given up tasks, i.e.,

$$\frac{\#\ Give\text{-}ups\ for\ both\ Agents}{\#\ Accepted\ Tasks\ for\ both\ Agents} \tag{3.5}$$

  Although the number of give-ups can also indicate a lack of ability, we chose to associate it with integrity since, as explained before, we considered the task feasible (and all participants tried it through a tutorial first). As such, when a person decided to give-up, it showed more about persistence and "keeping promises", which are traits of integrity according to the questionnaire used.

- *Fairness 1*: According to the dictionaries of Cambridge and Merriam-Webster, fairness can be defined as treating people equally, impartially, free from self-interest, prejudice or favouritism. As such, for *Fairness 1* we calculated the absolute difference between the lies given to each agent, i.e.,

$$abs\left(\frac{\#\ Lies\ for\ Agent\ X}{\#\ Accepted\ Tasks\ for\ Agent\ X} - \frac{\#\ Lies\ for\ Agent\ Z}{\#\ Accepted\ Tasks\ for\ Agent\ Z}\right) \tag{3.6}$$

  This fairness factor aimed at reflecting the fairness w.r.t. honesty. Although the difference of lies between agents could be interpreted as a sign of higher benevolence (towards the agent the participant lied the least), we considered this to be a signal of integrity since the participant was harming one more than the other (which is unfair). The difference between this metric and Favouritism is that Favouritism is something positive, such as helping out a friend (not necessarily with the intention of harming the other).

- *Fairness 2*. Similarly, this factor was absolute difference between the give-ups towards each agent, i.e.,

$$abs\left(\frac{\#\ Give\text{-}ups\ for\ Agent\ X}{\#\ Accepted\ Tasks\ for\ Agent\ X} - \frac{\#\ Give\text{-}ups\ for\ Agent\ Z}{\#\ Accepted\ Tasks\ for\ Agent\ Z}\right) \tag{3.7}$$

  This fairness factor aimed at reflecting the fairness w.r.t. commitment.

TRUSTWORTHINESS:
Although we frame trustworthiness as a combination of all the above, we also computed the direct consequences of it through success metrics. This was mainly useful to speculate how other metrics impacted overall trustworthiness in each situation. We see the consequences of trustworthiness in terms of the success of the tasks, which is usually the main goal in teamwork situations. In case of other goals, other metrics for trustworthiness may apply. In particular, we divide success as a consequence of trustworthiness in two ways:

- *TW abs*: This is the absolute consequence of trustworthiness, and it was calculated by *# Successful tasks to Agent X*. We called it absolute because it is the raw number of successes during the 10 minutes of experiment. This has to do with how well a participant can do a task, assuming that the more they successfully complete in 10 minutes, the fastest they will do it (which, in this task, is the "how well" indicator).

- *TW rel*: The relative consequence of trustworthiness was calculated as the ratio of presented tasks that were successful, i.e.,

$$\frac{\#\ Successful\ tasks\ to\ Agent\ X}{\#\ Presented\ tasks\ by\ Agent\ X} \tag{3.8}$$

  This was relative as it could be seen as the probability of one succeeding at a task when asked. Thus, this suggests whether the participant would do a task when asked.

3.4 RESULTS

In this section, we report how the observations (manifesta) relate with each other, and how these relate to participants' self reports (of ability, benevolence and integrity). As mentioned before, the purpose of separating the participants per condition was to create variation in the participants' manifestation of ability, benevolence and integrity, though manipulation of narrative or environment, but not of the task. The conditions were not created so that we could evaluate each condition against a control group, necessarily, since we can group them and see the relationship among variables. Still, we compare the metrics among the conditions to see the effect of our manipulation. We used R 4.2.2, with the packages First Aid 0.1 [27] for Bayesian t-tests and correlations.

Bayesian methods have become more popular when analysing behaviour data, see e.g., [13, 14, 33, 135, 306]. These have been found as an alternative to the more popular Frequentist tests. Frequentist approaches usually try to prove a null-hypothesis through frequentist statistical tests (which many times require certain assumptions from the data) and a *p-value*. Only with a low enough *p-value* we can say something about our hypothesis. This can be very hard to obtain with low quantity of behavioural data, which we usually get when doing research in human-computer/human-robot interaction. Furthermore, we can not really say how likely this is to be a good hypothesis, only that it is (or not) *statistically significant*. Bayesian tests, on the other hand, present probabilities, e.g., how likely it is that there is actually a difference between samples (instead of a yes/no). For these and other reasons, we chose to use Bayesian t-test and Pearson correlation for our data. In this paper, specifically, we do not try to prove one hypothesis. Instead, we explore how the subjective

measures and the agent observations relate and whether there was any difference among the conditions.

For Bayesian tests, both t-test in Section 3.4.2 and Pearson correlation in Section 3.4.3, several possible normal distributions that may fit to each of the metrics of our data are computed. This means that each metric will have a distribution of credible means and standard deviations. The test formula is then used for each of the credible combinations of means and standard deviations. Thus, the test results will also have an average value. We report all these by their 95% High Density Interval (HDI). Finally, we will evaluate the results of the tests by the probability of this average being positive (meaning there is or not a difference) and interpret it according to [218]. The closer to 0 or 1 the probability is, the more significant it is (depending on whether it is negative or positive difference, respectively). When this probability is around 0.5 it means that this average is around 0, which tells us that there is probably no difference between the two groups for that formula. More on how Bayesian tests work can be found in Kruschke et al. (2013) and McElreath et al. (2020) [218, 251].

3.4.1 Subjective Measures

The original questionnaire from Adams et al. (2008) [6] is a validated one, where the authors ran both Exploratory Factor Analysis and Confirmatory Factor Analysis. However, since we adapted the tool, we also ran a Cronbach alpha on our results. We found good Cronbach's alphas [94] for ability questions ($\alpha = 0.89$) and integrity questions ($\alpha = 0.84$), and excellent ones for benevolence questions ($\alpha = 0.93$).

3.4.2 Differences between conditions

In this subsection we look at the differences of the means of subjective measures and objective (observed) metrics of ability, benevolence, integrity and overall trustworthiness between conditions. We compare each group (Ga, Gb, Gi) with Gn. In particular, in **Table 3.2** we compare Gb and Gn's metrics related to benevolence, both subjective (questionnaire benevolence-related items *QB* from 1 to 5 and the mean) and observed (*Favouritism*). Similarly, **Table 3.3** compares groups Ga and Gn in terms of ability metrics from questionnaire (QA1 to QA5 and mean) and observations related to time and moves. Finally, **Table 3.4** compares groups Gi and Gn in terms of integrity metrics, both subjective (questionnaire items *QI* 1 to 5 and mean) and objective (*Honesty*, *Commitment*, *Fairness 1* and *2*).

In each of these tables, the first two columns show the average of the metrics for each of the groups, with the limits of 95% high density interval within brackets. The difference between means is showed in Diff Means column, also with the limits of 95% highest density interval within brackets. The SD columns show standard deviations in a similar fashion. Finally, the column % presents the probability of *Diff Means*> 0 and Evaluation column interprets the % column according to [79]. This evaluation present the risk of betting on such correlation, which can translate into how probable a correlation is.

Table 3.2: Bayesian T-test between benevolence (Gb) and normal (Gn) groups, presenting group's possible distributions' means and standard deviations, the difference between the means, the probability of this difference being positive and the evaluation of this probability, according to [79]. As explained in Section 3.3.5, *QB* 1 to 5 correspond to each question regarding benevolence towards Agent X and *QB Mean* is their average. Favouritism is the observed metric related to benevolence.

Metric	Gb Mean	Gn Mean	Diff Means	Gb SD	Gn SD	%	Evaluation
QB1	4.8 [3.7, 5.9]	4 [2.5, 5.5]	0.8 [-1, 2.7]	1.8 [1.1, 2.9]	2.2 [1.2, 3.6]	0.8210	Casual bet
QB2	4.3 [3.4, 5.2]	3.4 [1.9, 5.1]	0.8 [-1, 2.6]	1.4 [0.8, 2.3]	2.4 [1.4, 3.9]	0.8282	Casual bet
QB3	4.4 [3.8, 5.1]	3.4 [1.8, 4.9]	1.1 [-0.6, 2.8]	1.1 [0.6, 1.7]	2.3 [1.3, 3.8]	0.9045	Promising but risky bet
QB4	4.8 [4.1, 5.5]	3.7 [2.4, 5]	1.1 [-0.3, 2.6]	1.2 [0.7, 1.8]	1.9 [1.1, 3]	0.9446	Promising but risky bet
QB5	4.7 [4, 5.4]	3.7 [2.4, 4.9]	1.1 [-0.4, 2.5]	1.1 [0.7, 1.8]	1.9 [1.1, 3]	0.9347	Promising but risky bet
QB Mean	4.6 [3.9, 5.4]	3.6 [2.3, 5]	1 [-0.6, 2.5]	1.2 [0.7, 1.9]	2 [1.1, 3.2]	0.8999	Casual bet
Favouritism	0.6 [0.5, 0.7]	0.6 [0.4, 0.7]	0.1 [-0.1, 0.3]	0.2 [0.1, 0.3]	0.2 [0.1, 0.4]	0.7661	Casual bet

Table 3.3: Bayesian T-test between ability (Ga) and normal (Gn) groups, presenting group's possible distributions' means and standard deviations, the difference between the means, the probability of this difference being positive and the evaluation of this probability, according to [79]. As explained in Section 3.3.5, *QA* 1 to 5 correspond to each question regarding ability and *QA Mean* is their average. Time/Task, Moves/Task and Moves/Time are the observed metrics related to ability.

Metric	Ga Mean	Gn Mean	Diff Means	Ga SD	Gn SD	%	Evaluation
QA1	6.2 [5.5, 6.9]	5.5 [4.5, 6.4]	0.7 [-0.5, 1.9]	1 [0.5, 1.7]	1.4 [0.8, 2.3]	0.9019	Promising but risky bet
QA2	5.7 [4.9, 6.5]	5.2 [4.3, 6.1]	0.6 [-0.6, 1.8]	1.2 [0.7, 1.9]	1.3 [0.8, 2.2]	0.8355	Casual bet
QA3	5.4 [4.6, 6.3]	5.5 [4.6, 6.4]	0 [-1.3, 1.2]	1.3 [0.8, 2.1]	1.3 [0.7, 2.1]	0.4742	Not worth bet against
QA4	5.6 [4.7, 6.6]	5.4 [4.3, 6.5]	0.2 [-1.2, 1.7]	1.4 [0.8, 2.3]	1.6 [0.9, 2.7]	0.6309	Not worth bet on
QA5	5.7 [4.6, 6.8]	5.9 [5.2, 6.5]	-0.2 [-1.5, 1]	1.6 [0.9, 2.6]	1 [0.6, 1.6]	0.3633	Not worth bet against
QA Mean	5.7 [5, 6.5]	5.5 [4.6, 6.3]	0.2 [-0.9, 1.3]	1.1 [0.6, 1.7]	1.3 [0.7, 2.1]	0.6783	Not worth bet on
Time/Task	50.8 [27.1, 77.2]	70.6 [32.3, 110.1]	-19.4 [-65.9, 24.5]	35.6 [11.6, 63.5]	55.4 [27.6, 95.2]	0.1819	Casual bet against
Moves/Task	49.7 [44.5, 55.2]	59.5 [38.6, 81.4]	-9.8 [-31.6, 12.4]	8 [4.3, 13.5]	31.1 [17.5, 51.1]	0.1753	Casual bet against
Moves/Time	1.2 [0.8, 1.5]	1.2 [0.6, 1.6]	0 [-0.6, 0.6]	0.5 [0.3, 0.8]	0.7 [0.4, 1.3]	0.5049	Not worth bet on

Table 3.4: Bayesian T-test between integrity (Gi) and normal (Gn) groups, presenting group's possible distributions' means and standard deviations, the difference between the means, the probability of this difference being positive and the evaluation of this probability, according to [79]. As explained in Section 3.3.5, *QI* 1 to 5 correspond to each question regarding integrity and *QI Mean* is their average. Honesty, Commitment, Fairness 1 and 2 are the observed metrics related to integrity.

Metric	Gi Mean	Gn Mean	Diff Means	Gi SD	Gn SD	%	Evaluation
QI1	5.9 [5.4, 6.4]	5.3 [4.1, 6.5]	0.6 [-0.7, 1.9]	1 [0.7, 1.4]	1.8 [1.1, 2.9]	0.8286	Casual bet
QI2	6.1 [5.5, 6.7]	6.1 [5.3, 7]	0 [-1.1, 1]	1.2 [0.7, 1.7]	1.3 [0.7, 2.2]	0.4792	Not worth bet against
QI3	6 [5.5, 6.5]	6.1 [5.2, 7]	-0.1 [-1.2, 1]	1.1 [0.7, 1.5]	1.4 [0.8, 2.4]	0.4083	Not worth bet against
QI4	6.3 [5.7, 6.9]	7 [7, 7]	-0.7 [-1.3, -0.1]	0.8 [0.4, 1.4]	0 [0, 0]	0.0254	Promising but risky bet against
QI5	6.7 [6.1, 7.1]	7 [7, 7]	-0.3 [-0.9, 0.1]	0.6 [0, 1.1]	0 [0, 0]	0.0844	Promising but risky bet against
QI Mean	6 [5.6, 6.5]	6.1 [5.4, 6.9]	-0.1 [-1, 0.7]	0.9 [0.6, 1.2]	1.1 [0.6, 1.8]	0.4092	Not worth bet against
Honesty	1 [0.9, 1]	1 [1, 1]	0 [-0.1, 0]	0.1 [0, 0.1]	0 [0, 0]	0.0972	Promising but risky bet against
Commitment	1 [1, 1]	1 [1, 1]	0 [0, 0]	0 [0, 0]	0 [0, 0]	0.4993	Not worth bet on
Fairness 1	1 [0.9, 1]	1 [1, 1]	0 [-0.1, 0]	0.1 [0, 0.2]	0 [0, 0]	0.0718	Promising but risky bet against
Fairness 2	1 [1, 1]	1 [1, 1]	0 [0, 0]	0 [0, 0]	0 [0, 0]	0.5016	Not worth bet on

Table 3.5: Bayesian Pearson Correlation between subjective and observed metrics, presenting the means of the correlation among the distributions of two metrics, the probability of this mean being positive and the evaluation according to [79]. In particular, we present the correlations between (1) subjective trustworthiness (STW), which is an average of all the questions in the questionnaire (i.e., QA, QB and QI), and observed trustworthiness, both absolute and relative, i.e., TW abs and TW rel, respectively; (2) subjective metric of ability (QA Mean, which is the average of ability questions) and objective metrics of ability (Time/Task,Moves/Task and Moves/Time); and in a similar fashion for metrics of (3) benevolence and (4) integrity.

Metric 1	Metric 2	Mean	%	Evaluation
STW	TW abs	0.3 [0, 0.5]	0.9855	Good bet - too good to disregard
STW	TW rel	0.5 [0.2, 0.7]	0.9999	Nearing certainty
QA Mean	Time/Task	-0.2 [-0.5, 0.1]	0.1035	Casual bet against
QA Mean	Moves/Task	0 [-0.3, 0.3]	0.5117	Not worth bet on
QA Mean	Moves/Time	0.2 [-0.1, 0.4]	0.8590	Casual bet
QB Mean	Favouritism	0.3 [0, 0.5]	0.9802	Good bet - too good to disregard
QI Mean	Honesty	0.1 [-0.2, 0.4]	0.7757	Casual bet
QI Mean	Commitment	0 [-0.3, 0.3]	0.4976	Not worth bet against
QI Mean	Fairness 1	0.2 [-0.1, 0.5]	0.8692	Casual bet
QI Mean	Fairness 2	0 [-0.3, 0.3]	0.5087	Not worth bet on

3.4.3 Correlations

After exploring the differences between conditions, we treated the data set as one. In this subsection, we show the results of Bayesian Pearson correlation between metrics. **Table 3.5** presents the correlations between subjective and observed metrics. Furthermore, **Table 3.6** presents the correlations between observed consequences of trustworthiness metrics and observed metrics of ability, benevolence and integrity.

We have also included Figures that illustrate the values in the tables. In particular, we have picked one Figure that shows the correlation between one subjective and one observed metrics, such is the case of QB Mean (benevolence questions) and Favouritism in **Figure 3.3**. We also show how the correlation between STW (subjective trustworthiness) and observed trustworthiness, *TW abs* and *TW rel*, respectively, in Figures 3.4 and 3.5. In these Figures, we can see the actual data in the red bar charts on the axis and black circles (each representing an instance) in the middle of the plot. The blue lines around the red charts show the credible normal distributions for our data. As explained in the beginning of this section, the correlation is then run for the several credible distributions, creating a distribution of possible correlation values. On top of the figures, we can see the distribution of the correlation possible values and their 95% high density interval. The more similar the credible distributions of the data are, the slimmest the correlation distribution will be and the more likely it is that the correlation is the median value. The furthest this distribution is from 0, the more probable it is that there is indeed a correlation, positive (if on the positive side) or negative (if on the negative side).

3.4.4 Strategy

In **Table 3.7** we can see the results of the multiple choice question regarding participant's strategy when performing the task. We have ordered the table according to the original

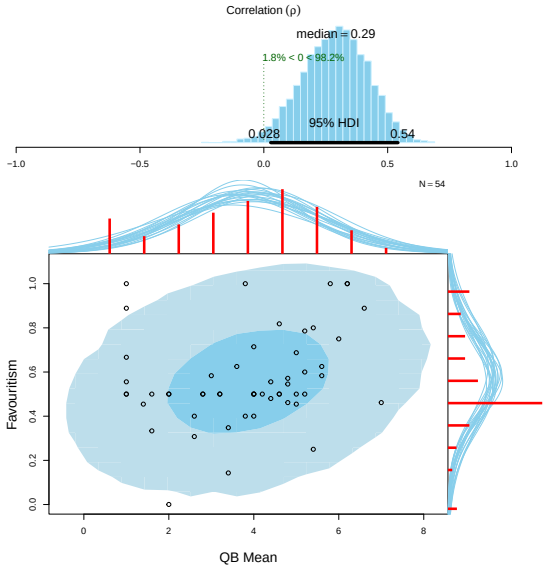


Figure 3.3: Bayesian Pearson Correlation between subjective benevolence (QB Mean) and observed benevolence (Favouritism). The black circles represent the instances of our dataset, and the metrics' histograms are on the red bar charts. The blue lines around the red charts show the credible normal distributions for our data. The top plot presents the distribution of the correlation of possible values and their 95% high density interval.

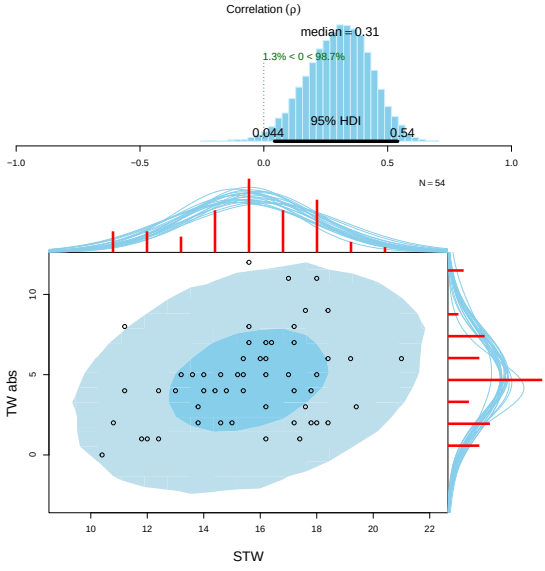


Figure 3.4: Bayesian Pearson Correlation between subjective trustworthiness (STW) and observed absolute trustworthiness (TW abs). The black circles represent the instances of our dataset, and the metrics' histograms are on the red bar charts. The blue lines around the red charts show the credible normal distributions for our data. The top plot presents the distribution of the correlation of possible values and their 95% high density interval.

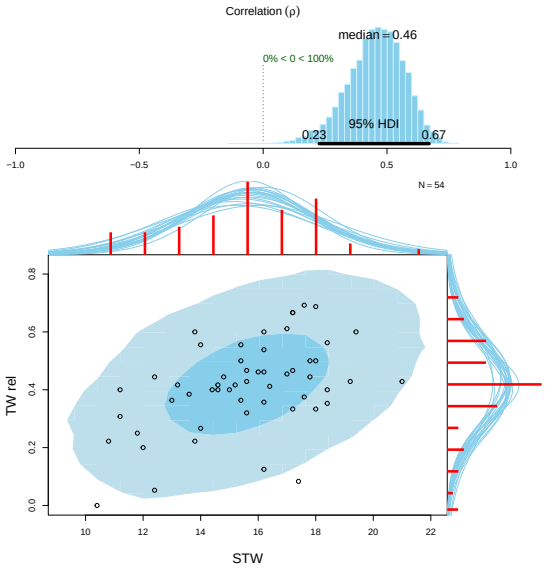


Figure 3.5: Bayesian Pearson Correlation between subjective trustworthiness (STW) and observed relative trustworthiness (TW rel). The black circles represent the instances of our dataset, and the metrics' histograms are on the red bar charts. The blue lines around the red charts show the credible normal distributions for our data. The top plot presents the distribution of the correlation of possible values and their 95% high density interval.

Table 3.6: Bayesian Pearson Correlation between metrics of overall trustworthiness and the observed metrics of each aspect of ability, benevolence and integrity. The table presents the means of the correlation among the distributions of two metrics, the probability of this mean being positive and the evaluation according to [79]

Metric 1	Metric 2	Mean	%	Evaluation
TW rel	Time/Task	-0.1 [-0.4, 0.2]	0.2844	Not worth bet against
TW rel	Moves/Task	0.3 [0, 0.5]	0.9740	Good bet - too good to disregard
TW rel	Moves/Time	0.2 [0, 0.5]	0.9451	Promising but risky bet
TW rel	Favouritism	0.8 [0.6, 0.9]	0.9999	Nearing certainty
TW rel	Honesty	0.3 [-0.1, 0.5]	0.9416	Promising but risky bet
TW rel	Commitment	0 [-0.3, 0.3]	0.4979	Not worth bet against
TW rel	Fairness 1	0.3 [0, 0.5]	0.9583	Good bet - too good to disregard
TW rel	Fairness 2	0 [-0.3, 0.3]	0.4975	Not worth bet against
TW abs	Time/Task	-0.7 [-0.8, -0.5]	0.0001	Promising but risky bet against
TW abs	Moves/Task	0 [-0.3, 0.3]	0.4630	Not worth bet against
TW abs	Moves/Time	0.7 [0.6, 0.8]	0.9999	Nearing certainty
TW abs	Favouritism	0.2 [0, 0.5]	0.9411	Promising but risky bet
TW abs	Honesty	0.2 [-0.1, 0.5]	0.8778	Casual bet
TW abs	Commitment	0 [-0.3, 0.3]	0.5009	Not worth bet on
TW abs	Fairness 1	0.1 [-0.2, 0.4]	0.7863	Casual bet
TW abs	Fairness 2	0 [-0.3, 0.3]	0.5047	Not worth bet on

order in the questionnaire and numbered them for reference simplicity. In the complementing text of “Other", 3 participants reported “slightly" helping more agent X, whereas 1 reported similar regarding agent Z.

Table 3.7: Number of times each *goal* was chosen in strategy question (multiple choice allowed), ordered as in questionnaire.

Option	**#**
1. Collect as many products as possible	39
2. Collect products as fast as possible	34
3. Maximize team score	26
4. Maximize personal score	16
5. Collect easiest products (based on icon)	26
6. Collect easiest products (based on distance)	25
7. Collect according to the chat messages	3
8. Helping both agents equally	15
9. Helping specifically agent X	2
10. Helping specifically agent Z	2
11. I do not know what my goal was	1
12. Other	3

3.5 Discussion

In this section we discuss the results, highlight some interesting aspects and reflect on what they might mean. We divide this section similarly to the result section to make it easier to follow. However, we find it relevant to highlight our main findings beforehand, as some of these findings affect the interpretation of the several results presented. As such, our main findings looking at the numbers were:

1. There were expected differences between conditions both for self-reported (subjective) and observed metrics, meaning our manipulations had effect on the participants.

2. Subjective trustworthiness (STW) highly correlates with both observed overall consequence of trustworthiness metrics (TW abs and TW rel).

3. There are probable yet low correlations (probable here means that after several correlation tests with the possible distributions of the data, the correlation was mostly either positive or negative, making it a safer bet), between subjective and one of the observed metrics for each of the aspects of ability, benevolence and integrity.

4. Almost all observed metrics of ability, benevolence and integrity correlated with TW rel and most of them also correlated with TW abs. These correlations were rather low except for Favouritism (benevolence) with TW rel and Time/Task and Moves/Time with TW abs, which were expected by definition.

5. Most participants said they collected as many products as possible and as fast as possible. They also mentioned they chose the easiest products to collect. This can be

interpreted as participants caring mostly about doing the task well and quickly, with the least effort associated to it, and not paying so much attention to the details of the scenario.

Beyond the numbers, we found several interesting points transversal to all results. These were:

1. Most participants only cared about the game/task itself, focusing on performing well at finding and retrieving objects rather than paying attention to the context that was presented initially (i.e., participants tend to forget that one of the agents was their friend, in benevolence condition, or that it is their last day of work, and they need to make as much money as possible, in integrity condition). They also rarely looked at the chat, from our observation.

2. Probably related to the previous point, some participants mentioned that they do not see the agents as their teammates. This may mean that our task could be improved by including, for example, more interdependencies or autonomy.

3. Very few participants lied or gave up on tasks. This may be because they feel observed (which affects ecological validity) or because they do not feel compelled to lie/give up in such scenario.

4. It can be quite challenging to self-report measures of one's ability, benevolence and integrity, specially without a term of comparison. As such, we need to take these metrics' results critically.

5. Some participants verbalized that they think it does not make sense for us to ask them the questions related to benevolence (we can see in **Figure 3.3** that there were several low average scores). This reflects that many of them did not feel like they were helping or "having the back" of Agent X, for example. However, results tend to show that some participants did feel that way, as there were also higher scores, different between groups, and a correlation with Favouritism.

In the remaining of this section we will go over each table and analyse the results in more detail, keeping in mind the points stated above. At the end, we will compare our main findings with the purpose of this study.

3.5.1 Differences between groups

It is interesting to see that all questions regarding self-reported benevolence (**Table 3.2** show higher scores in the benevolence group (Gb) than in the normal group (Gn), all of these with high probabilities associated. In particular, QB3, QB4 and QB5 present the highest probabilities of in fact differing between the two groups showing a "promising but risky bet" according to [79], with mean differences higher than 1 (in a 7-point scale). These questions are about having *"worked to protect Agent X"*, *"watched teammate Agent X's back (synonym: to look out for Agent X in case it needs assistance)"* and *"looked out for teammate Agent X"*. It is worth mentioning that the question that presents highest probability of difference (QB4) is the one that translates better to our chosen metric for observed benevolence, i.e., give more assistance to Agent X than Agent Z. However, Favouritism presents only a casual

bet that it differs between Gb and Gn. This may mean that although our narrative made participants care more about Agent X (as a feeling), there was not a relevant difference in how much they actually helped one agent over the other (in action).

When comparing Ga with Gn (**Table 3.3**, we see that it is promising to bet that there is a difference in answer to QA1, which is "As a teammate, I was capable at my jobs", with a difference of 0.7 between the groups' means. The other questions, however, do not show big differences among the two groups, which leads to only a casual bet in the QA Mean comparison. In general it is hard to have participants evaluate themselves regarding ability, since they do not know how other participants performed. In particular, how good they actually were in the task might have been evenly distributed among groups. We cannot know for sure if this was the case, but we can see that Moves/Time, which indicates how fast participants were in general, should in that case not differ between groups and it did not. If they were evenly distributed in terms of actual ability, that might have made the manipulation of the group imperceptible for the participants in terms of how much easier the task becomes. We did expect, however, that it would show a difference in the observed metrics, such as Time/Task and Moves/Task, since the participants in Ga group could see at all times when the product they should collect was, and in fact both suggest a casual bet (it is against as the higher the time taken to perform a task, the lower expected ability). Interestingly, we can see that the standard deviations in Ga are much lower than in Gn for Moves/Task and Time/Task, which probably means that our manipulation created less difference among participants of Ga regarding these two metrics. This makes sense as all participants in the group could see all the products from afar, making the task simply easier for everyone, making their actual differences less significant.

Finally, although Gi participants seem to score lower in self-reported subjective measures (QI metrics) than in Gn, these are not very relevant in terms of Mean differences and percentage. This is understandable when we look at the observed metrics, which mostly show either 0 or 1, indicating that there were very few people lying and giving up. This means that the manipulation in terms of integrity was not successful.

Although we can see there were expected differences between the groups, showing that we did manipulate the participants in a certain way, results show that our manipulations did not work for all subjective (self-reported) and observed metrics. We speculate this happens because of several reasons observed during the experiment (already stated in the beginning of this section).

3.5.2 Correlation between subjective and observed metrics

In **Table 3.5** we can find the bayesian pearson correlations between subjective and objective metrics. Self-reported subjective trustworthiness (STW) highly correlates with observed relative trustworthiness (TW rel) near certainty, and it also a good bet that it positively correlates with absolute trustworthiness (TW abs). This means that overall participants answered the questionnaire according to their performance in the task and help towards Agent X. These results align with the very likely (98%) positive correlation between both self-reported subjective and observed metrics for benevolence. The subjective metrics of ability also seem to possibly (only a casual bet) correlate with two of the observed metrics of ability, such as Moves/Time and Time/Task, although these wouldn't be very high correlations. Moves/Task however seems to not correlate at all. It may be that because of

the nature of the task, Moves/Task is not a good measure of ability. In our task, participants usually choose to go through all the corridors until they find the product they need to collect. We expected some participants to remember the corridors better than others and that would differ them in terms of average Moves/Task. Although we cannot tell based on this correlation only that the measure did not succeed in capturing participants' ability, it probably did not suit our task. Finally, as said before, very few participants lied, it seems that their self-reported metrics of integrity possibly (87%) correlate with observed metric Fairness 1 (low correlation). Measures of commitment and fairness w.r.t. commitment (Fairness 2) do not correlate at all with self-reported Integrity. As we saw in the previous section almost no participant gave up in the task, which makes it impossible to evaluate these metrics.

Even though it is possible that most subjective metrics correlate with objective metrics in the expected direction, most of these correlations are not high. The reasons for this may be (1) because the self-reported measures do not actually represent the participants' trustworthiness aspects of ability, benevolence and integrity or (2) because the observed metrics are not sufficient to capture the participants' trustworthiness aspects of ability, benevolence and integrity. Ideally we would compare these with a ground-truth values of ability, benevolence and integrity, but unfortunately there is no way we can have these. We can, however, see how the observed metrics of each aspect relate to overall metrics of trustworthiness.

3.5.3 Correlation between observed trustworthiness and other observed metrics

The observed metrics for trustworthiness are by definition related with the observed metrics for each of the aspects of ability, benevolence and integrity. For example, the number of successes (which is taken into account for both absolute and relative trustworthiness, i.e., TW abs and TW rel, respectively) is related to how fast a participant can do a task (Time/Task) and also how many times the participant decided to help Agent X instead of Agent Z (Favouritism), etc. In this subsection we analyse how actually these metrics correlate (in **Table 3.6**). We can see that TW rel (which can be interpreted as how likely it is for the participant to collaborate with Agent X) probably positively correlates with Moves/Time (95%) and Moves/Task (97%), Favouritism ($\tilde{1}$00%), Honesty (94%) and Fairness 1 (96%). In particular, the correlation between TW rel and Favouritism is extremely high (0.8), which is expected given their definitions, i.e., percentage of presented tasks by Agent X that were successful and percentage of all successful tasks that were for Agent X. As for TW abs (which can be interpreted with how much a participant actually helped Agent X in absolute terms), we can see that it highly certainly ($\tilde{1}$00%) highly correlates with Moves/Time and negatively with Time/Task ($\tilde{1}$00%, as it shows $\tilde{0}$% for positive correlation). There is a high chance that it slightly correlates with Favouritism (94%) and, not as probable, with Honesty (88%). Again, commitment and Fairness 2 do not correlate at all, probably given to the scarcity of these metrics.

3.5.4 Strategy

In **Table 3.7** we have reported the results for the question related to strategy. We realized that many times, strategy was based on the least effort for the participant, either by having

already seen that product icon before or because it was simply closer to them. Supported by literature [40, 406] and our results, we speculate we should consider human's *cost-benefit evaluation*, i.e. participants choose whether or not the reward is worth the perceived effort, and this affects their decision. In particular, the law of least effort plays a central role in decision-making, i.e. when presented to two tasks with equal rewards, one will choose the least effortful [355]. This may mean that more than paying attention to the task context in terms of benevolence and integrity, participants might have been in fact playing according to their own perceived benefit with the lowest effort possible. When determining how much someone should be trusted, then, it may be more helpful to know what is for them a risk/effort and reward/benefit. This may also be helpful to predict where the human teammates may do next, once we detect the strategy they are using (for example, going for the icons they have already seen before).

3.5.5 Implications for human trustworthiness in human-AI teams

In this paper, we wanted to study what makes a human trustworthy in a human-AI team setting, and how could an artificial agent observe it, given a specific task. Although we knew we would not be able to prove it (due to lack of ground-truth), we wanted to explore how cues on ability, benevolence and integrity could be used as observable trustworthiness. The goal was to take the first step towards understanding how an artificial agent can form a belief of trustworthiness regarding its human teammate. In particular, we wanted to explore how an agent could form beliefs on (1) whether the human could do a task and on (2) whether they would do it. In light of these results and the main purpose of this study, we believe that participants' ability, benevolence and integrity do affect their overall trustworthiness and the direct consequences of it, but may not be the best human's krypta in human-AI teamwork scenario, given low correlations and overall findings, as we discuss in this subsection.

In terms of manifesta, the cue of Favouritism highly correlated with whether the human did the task the agent asked for, and cues of performance (w.r.t. moves) and commitment seem to have also affected this. Similarly, cues of performance w.r.t. to both time and moves highly correlate with how well they would do it. However, the cues of integrity did not seem to be suitable for this environment. Although these seem like promising krypta and manifesta, we suspect that this may not be the best model for this type of task and environment. In particular, both benevolence and integrity definitions and cues may not be appropriate, as most participants did not feel particularly inclined towards teaming up with (and helping) one agent or giving up/lying.

From our understanding, there might be better ways of detecting the willingness to perform a task successfully in human-AI team settings, particularly in short interactions, such as detecting overall strategy or personal preferences. As we suspected during the pilot study, participants were mostly paying attention to the task itself and to how to solve it efficiently, instead of caring about the interactions with the agents or even getting points by just lying. Although this may have to do mostly with our setup, we believe that in order to understand whether a person will do something, we need to understand how, in general, the person is executing tasks (what is their strategy). This may depend on, for example, what the execution of the task represents in terms of risk and reward for that person. And,

of course, risk and reward may be related to a person's ability, benevolence and integrity. It may also be related with something else, though, such as a personal preference. After this experiment, we also consider that different tasks may require different manifesta and krypta apply. Overall, these measures must be further explored in different scenarios, as well as compared with other possible trustworthiness models as krypta (and respective manifesta).

3.5.6 Limitations

The setting in which we ran the experiment presented some technical difficulties. The server presented a lag for all participants, the higher the further from it they were, which might have negatively affected our results. Moreover, the setting of the experiment might have made participants feel too "observed", having slightly harmed the ecological validity of the study (possibly affecting the give-ups and lies, as discussed in the beginning of this section). The task in our experiment was a short one (10min) and did not involve high interdependence, which may have contributed to the feeling of not being teammates. These characteristics of the task may also not be the most suitable to benevolence and integrity. For such tasks, it may be more relevant to look at other models of trust, such as swift trust, see e.g., [167, 201]. In fact, future research should also consider that if trust changes overtime, so can trustworthiness.

Overall, It is still an open question how to appropriately model the willingness of the human teammate to perform an action, as such question is also not yet answered by social sciences. In particular, it would be interesting to explore measures that do not assume complete knowledge (i.e., measures that take into account only interactions with one agent). What's more, our task did not allow the participants to simply fail due to lack of ability. This is justified in Section 3.3.2, but it might be interesting to explore a task where it is possible to (1) lie, (2) give up and (3) fail (due to lack of ability), while being possible to tell these 3 apart.

Another limitation of our work mentioned before is the fact that there is no ground-truth or baseline we can compare our results with. As such, we have proposed observed metrics for the aspects we relate with trustworthiness (ability, benevolence and integrity) and two observed consequences of trustworthiness metrics (TW rel and TW abs), but we cannot test them. We can only compare them with subjective metrics which we are not certain about how well they capture what we want to know (i.e., how trustworthy someone is), given its dependency on self and context awareness of the participants. We look forward to exploring human trustworthiness in different scenarios and with different tasks and more sophisticated agents, in order to compare with and improve our current model.

3.5.7 Future directions

Our future directions include exploring the manifesta and krypta further, their developments and context dependencies in time and throughout interactions. We will work on using artificial trust beliefs for decision-making support, both for autonomous decision-making of the artificial teammate and for human support. It is also important to explore ways of evaluating artificial trust models, e.g. by creating test-beds for such experiments. Further explanation of the next steps of our work can be found in Centeio Jorge et al. (2023)

[61].

In this work, we see that both benevolence and integrity probably affect human actions in human-AI teamwork and should be taken into account when designing such teams. However, benevolence and integrity may not be the most direct aspects to either measure or use for prediction. On the other hand, as we observed, there likely is an influence of participants' strategy on which task they choose and whether they succeed on it. As such, we want to work on modelling willingness (i.e., factors other than competence that contribute to the success of the task) more concretely, e.g., using aspects which are more task and context dependent, such as possibility, interest in doing a certain type of task, preference, disposition, etc). We want to use these aspects to build a model for informed and justified decision-making of the artificial teammate, making use of principles of Interdependence Analysis and Coactive Design [194]. Such tool can also help human decision-making in human-AI teamwork scenarios, by providing an overview of the feasibility of different team configurations, for example. Furthermore, we want to update this decision-making model through interaction using existing models such as [16]. In future research, we want to use scenarios where collaboration is more necessary and explicit. Ideally, these scenarios also provide different levels of interdependence which allow the teammates to choose whether to engage in certain tasks or whether to help a teammate, for example, and can include mixed-motives. Such scenarios include search and rescue tasks [398], moving-out tasks [61], or even cooking tasks [226].

Human's trust and trustworthiness is also affected by the artificial teammate's behaviour, see e.g., [61]. It would also be interesting to investigate the human's trustworthiness in situations where the artificial teammate behaves differently. For example, situations where the artificial teammate does not obey to its immediate human teammate have been recently studied, see e.g., [308, 356]. Such situations can have mixed motives and knowledge, e.g., the human may want to go straight whereas the artificial teammate knows that it is dangerous (e.g., the human does not have the skills required for what's ahead), so it opposes. We would like to explore how human's trust and trustworthiness change in such situations and how they depend on the outcome, i.e., if it turns out that the agent is right or wrong has any impact on the collaboration and human trustworthiness (for example, changing the human's willingness).

3.6 Conclusion

In this paper, we take a first step towards implementing beliefs of artificial trust through interaction. According to the previously presented model of human trustworthiness in human-AI teams, we proposed a set of metrics for observable human behaviours (manifesta), representing aspects of their trustworthiness through their krypta (in this case ability, benevolence and integrity). Both theoretically and empirically, we have explored these metrics and compared them with self-reported subjective metrics of the same krypta. We have also compared both observed and self-reported metrics with overall metrics of trustworthiness (based on absolute and relative task success, from the perspective of one agent). Results showed that there was a high correlation between the average of the self-reported subjective metrics and observed metrics of overall trustworthiness. However, when dividing these metrics into ability, benevolence and integrity, the subjective-observed pair-wise correlations were not so high, which may mean that although these

metrics are relevant they may not correspond well to each of the aspects. We can see that ability, benevolence and integrity do affect their overall trustworthiness and the direct consequences of it, but may not be the best human's krypta in human-AI teamwork scenario. On the other hand, we observed that, the human teammate's decision on whether to perform a task depended on a strategy, related to participant's goals and cost-benefit analysis. Unfortunately, we cannot have ground-truth or other models to compare our results with, but these results shed light on how different aspects of trustworthiness can be used for human behaviour prediction in human-AI teamwork scenario. Although there is a need for further exploration of these metrics, this paper presents an important step towards building an intelligent agent capable of building beliefs of trust in human teammates and therefore capable of making informed decisions to achieve the team's goal.

Acknowledgments

This material is supported by Delft AI Initiative and by the TAILOR Connectivity Fund. Similarly, it is based upon work supported by the National Science Foundation (NWO) under Grant No. (1136993), and by the European Commission funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant 820437). The support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these institutions.

Thanks to all colleagues who helped us find the right methods and analysis for this experiment, and of course to those who proofread the manuscript.

Code and Data Availability

The code used for the task presented in this paper can be found in `https://github.com/centeio/click-collect`. The resulting data, as well as the questionnaire used, can be found in `https://doi.org/10.4121/21982991.v1`.

4

Interdependence and Trust Analysis (ITA): a Framework for Human-Machine Team Design

As machines' autonomy increases, the possibilities for collaboration between a human and a machine also increase. In particular, tasks may be performed with varying levels of interdependence, i.e., from independent to joint actions. The feasibility of each type of interdependence depends on factors that contribute to contextual trustworthiness, such as team members' competence, willingness, and external factors. In this paper, we present the Interdependence and Trust Analysis (ITA) framework, which is an extension of Coactive Design's Interdependence Analysis framework [2]. By including information on contextual trustworthiness, ITA can better support the design of human-machine teams, as well as task allocation and selection. Evaluated through expert interviews and a focus group involving a search and rescue scenario, ITA shows potential as a decision-making tool and a communication bridge among human and machine teammates. Our findings emphasise the need to define tasks and roles based on agent characteristics and imply that decision-making models should align with human-centred objectives. ITA also highlights the trade-off between utility and effort when designing trustworthy systems, suggesting that guided conversations could improve the team design process. Finally, the ITA framework may improve transparency, justification, and interpretability in decision-making, contributing to appropriate trust among teammates.

This chapter was published as follows:
Centeio Jorge, C., Jonker, C. M., & Tielman, M. L. (2024). Interdependence and trust analysis (ITA): a framework for human–machine team design. Behaviour & Information Technology, 1–21. [69].

4.1 INTRODUCTION

In scenarios where humans and machines collaborate, several design decisions have to be made, such as who does what [16, 25]. In some situations this may be straightforward, such as when there is no overlap of teammates' (human's or machine's) expertise, e.g., imagine a kitchen robot that only works as a pressure cooker and a human (who, of course, cannot work as pressure cooker) who can prepare the ingredients that go in the machine. On the other hand, there are situations where both teammates can do certain tasks, for example a kitchen robot arm that also chops vegetables and a person who can do the same. Situations like the latter may become more frequent with the advancement of AI, since machines have the possibility of functioning with higher levels of autonomy. This opens the door for interdependence between humans and machines, i.e., when two parties have to rely on each other to perform a joint activity [194]. Growing capabilities and autonomy mean more possible designs for human-machine collaborations, with different types of interdependence in the subtasks involved, from full independence to mandatory joint actions. Finding a good division of labour between machine with variable levels of autonomy and human teammates is a problem [342] in human-machine team design, which can be helped with methodological interdependence and trust analysis, as proposed in this paper.

The design of human-machine interdependent relationships for teamwork involves a symbiosis between humans and AI that benefits humans [390]. In Johnson et al. (2014), the authors present the Interdependence Analysis table as a framework for Coactive Design [194], i.e., an approach to addressing the increasingly sophisticated roles that people and machines play as the use of human-machine teamwork expands into new, complex domains. The original Interdependence Analysis lists the capacities required for the execution of tasks. It encourages a comprehensive analysis of which teammate has the required capacities to be a performer of a certain task and whether the other teammate's support is mandatory, not possible, or helpful, i.e., increasing reliability or efficiency. After filling in the table, one should be able to understand the necessary interdependencies for each task, through a colour code and requirement gathering.

Although a thorough analysis of capacities is an important step, we claim that it is also important to consider other dimensions that may lead to the success of a task. Models based on trust and trustworthiness between humans (human-human) have been developed to formalise the dimensions that may lead artificial agents to successfully perform tasks [122, 126]. Following these models, for a cognitive agent, either human or artificial, to successfully perform a task, they need to have the capacities/capabilities (i.e., "can they do it?"), the willingness/intention to do it (i.e., "will they do it?"), and to have the external opportunities/permissions to do it (i.e., "is it possible to do it?"). In other words, these dimensions can be used to assess the trust that agents have in their teammate(s) to successfully perform a certain task. Trust in human-machine teams includes natural trust, i.e., trust beliefs of the human (see e.g. [402]), and artificial trust, i.e., trust beliefs of the machine (see e.g. [68]). What makes a human trustworthy for a task is not necessarily what makes a machine trustworthy for that task [376], however, to decide who should do it, we need to consider both. So far, there is no framework supporting human-machine team design that considers, in a methodological way, both machine and human team members' contextual trustworthiness (not only capabilities but also willingness and external factors)

for different interdependent roles and tasks. Johnson et al. (2021), the authors already propose an extension of the table that includes trust as one extra dimension to consider when analysing interdependencies [193]. However, this dimension of trust is (1) only considered for the trust in one of the agents (the performer) involved in the interdependence, and, in our opinion, it (2) could also be further divided into dimensions that are easier to assess, update, and use for informed decision-making. As such, we propose to include an analysis of teammates' willingness and external factors regarding different team configurations in the process of human-machine team design.

This paper's contribution is the *Interdependence and Trust Analysis (ITA)* framework, centred on an extended new version of Johnson's Interdependence Analysis tables, the *ITA table*. The ITA framework presents the conceptual workflow of the dynamic information used for decision-making in human-machine teams, which serves as input and output for the ITA table. Additionally, the ITA table analyses three dimensions of a team member's trustworthiness, i.e., not only *competence* (as in the original table in Johnson et al. (2014) [194]), but also *willingness* and *external factors*, for the different tasks involved in a human-machine shared goal. Our method proposes that human-machine team design should consider willingness as an important dimension in assessing the feasibility of a team configuration in terms of interdependence and task allocation. This implies that we conceptualise all team members (machines and humans) as agents with intentions (i.e., *something that [one] wants and plans to do* as per the Cambridge Dictionary). They can not only act, but also choose which possible actions to do. This is in line with frameworks like [122, 145], used in multiagent systems, but goes beyond the traditional view of machines as just executioners of an action when interacting with humans. Although this intentionality is widely accepted for human teammates, there is still a tendency to overlook willingness even for human team task allocation, and most works consider only capabilities, see e.g. [16, 194, 326]. Furthermore, the willingness dimension should be considered a task-based and role-based characteristic, rather than a property of the teammate that is transversal to all tasks and interdependencies. For example, a human teammate may be willing to independently carry a light object but not willing to carry it together with a robot. However, they may be willing to carry an object together with a robot if the object is heavy. This implies that human-machine team design should also consider that willingness depends on roles, for example, allocating the task of carrying light objects to the human while having the robot assist could decrease team performance and the overall human experience. This is in line with [288, 289], who suggests that a machine should adjust to the human's preference of being a leader or a follower on collaborative tasks. However, these works overlook joint actions and the possibility that capabilities for each role may also differ (e.g., one may not have the strength to carry a heavy object alone, but has some strength to support another teammate carrying it), which we include in our framework. Additionally, we suggest that external factors are increasingly relevant to consider in human-machine teamwork design, as machines become more autonomous and require clearer boundaries from performing certain actions, such as ethical and moral decisions (e.g., deciding whether to save someone's life), or for safety measures (e.g., holding a gun). Some works defend the development of artificial moral agents, i.e., artificial agents capable of making ethical and moral decisions [76], while others defend Meaningful Human Control (MHC), i.e., humans should ultimately remain in control of, and thus morally responsible for, everyday actions

[102]. This implies that our human-machine team design allows the human to explicitly delimit the machine's permissions and detect situations that require human oversight and control (aligned with [386]). Finally, our framework implies that human-machine team design decisions need to be explicit and easily revisited in order to comply with new ethical guidelines (e.g., transparency and traceability), such as *European AI Act*, and the *IEEE Ethically Aligned Design*. To evaluate the ITA table, we conducted two dyadic interviews (with two participants each) and one expert focus group with five participants. We present the results through a thematic analysis.

The proposed Interdependence and Trust Analysis framework can be used by a team designer to make decisions regarding which role each teammate should have in different tasks. Furthermore, it could be used as a decision-making support system, as well as a shared mental model [146, 329, 383]. Using the analysis for such cases may increase transparency among teammates and facilitate justification of one's actions, and, consequently, appropriate trust [376, 397].

In Section 4.2 we start by presenting the background concepts and related work that sustain our work. Then, in Section 4.3 we present the table, and frame it in Section 4.4. We present the results of the evaluation of the table in Section 4.5 and discuss it in Section 4.7.

4.2 Interdependence and Trust Analysis (ITA)

Human-machine (and human-AI, human-agent, etc) teamwork studies aim at integrating humans and intelligent machines, rather than deliberately pushing the human out of the loop [351]. The goal is usually to provide support to the human, avoiding hazardous consequences [146]. In fact, these teams can be beneficial for humans, for example in situations where it can be unsafe to have humans doing everything, e.g., disaster response [100] and search and rescue [326]. In other cases, these teams can reduce the human's workload, e.g., in collaborative cooking [151] and collaborative driving [24] scenarios. These teams can also be effective for tasks that require high precision, e.g. robot-assisted surgeries [96]. However, the design and implementation of these teams pose challenges [209, 384], especially when machines start having more autonomy as their range of capabilities increases. More autonomy allows for different possibilities of *interdependence*, depending on the scenario, which may allow for different team designs (who does what, etc). Furthermore, there are moments when machines should not use their capabilities, in order to comply with social norms, and ethical principles [31], and always allowing for meaningful human control [386].

Team members need to cooperate, collaborate and coordinate [194]. This is only possible with communication, mutual trust and shared mental models [329]. Designing human-machine teams should ensure these mechanisms, which can be challenging. In particular, finding a good division of labour between machine and human teammates is one such challenge [342]. In the process of task selection or allocation (see e.g. [5, 288]), a team member or designer, respectively, needs to consider how much they trust different team members to successfully perform a task within a certain context [16]. In the context of human-machine teamwork, we see trust as the belief in an entity's trustworthiness to perform a task successfully, within a certain context [64]. Trustworthiness is a complex concept, and following the literature, it can consist of a set of dimensions that range from the trustee's competence to its intentions [153, 404]. Depending on the nature of the trustor

and trustee, the trust and trustworthiness constructs may be more or less adequate. There are several works studying how humans trust machines (see e.g. [225, 227, 229, 314]), but not so many showing how machines should trust human partners (see e.g. [402, 403]). Models in slightly different settings propose that trustworthiness depends on 1) Ability, Benevolence and Integrity [250] (in human organizations), 2) Willingness, Competence [59] (in multi-agent systems), and 3) Performance, Process and Purpose [227] (when the human is the trustor and artificial agent is the trustee). For this last case, [163] proposes that the agent's characteristics affecting trust (i.e., perceived trustworthiness) are performance-based (such as reliability, failure rate, etc) and attribute-based (such as anthropomorphism, robot personality, etc).

Although there are several interpretations about what exactly trustworthiness is, we see a tendency to separate it into two bigger dimensions, i.e., one related to the potential to execute a task successfully (e.g., ability, competence, performance), and another related to the behaviour that may influence the execution of the task, related to the factors that contribute to one's intention of performing a task (e.g., benevolence, integrity, willingness, process, purpose). In fact, these two main dimensions are used in works such as [377], where authors divide human trust into performance trust and moral trust. Similarly, [254] shows how humans, besides competence, also perceive warmth in artificial teammates, which also affects their decision-making and collaboration [68]. In summary, to assess an agent's trustworthiness to successfully execute a certain task, we need to take into account the agent's competence and willingness (following [122]'s nomenclature) for the execution of that task.

Furthermore, the COM-B model for behaviour change [264] suggests that besides capability and motivation, which align with the two trustworthiness dimensions explored in the previous paragraph, a person needs the opportunity to behave in a certain way. In the context of teamwork, opportunity is only possible when a task is available and possible [41]. In other words, the execution of a task is influenced by external factors, which are contextual conditions determining the situation in which the task is executed [126, 163], such as team setting, environmental configuration, emotional state, workload, etc. As such, to entrust a task to an agent, one needs to have e positive belief regarding the agent's trustworthiness (i.e., competence and willingness), as well as a positive belief that the external factors allow that agent to execute that task.

When collaborating, humans and machines can take different roles [388], i.e., they can have different interdependent relationships. Interdependence can be soft or hard [194]. Soft interdependence happens when the collaboration improves the task efficiency, but it is not required. On the other hand, hard interdependence happens when the collaboration is necessary for the task to be successful. In particular, in soft interdependencies there can be a performer and a supporter [193]. The supporter, a teammate that possibly (necessarily or not) helps the performer, the main teammate involved in completing a task, to do the task. As such, when designing a human-machine team, in particular, deciding how to select or allocate tasks, one can select or allocate a specific role to perform a task. What's more, the competence and willingness required to be a main performer may differ from those of being a supporter [288]. Particularly, human teammates may have more or less willingness to engage in interdependent relationships with the machine, depending on the human characteristics or machine's characteristics. Human factors that contribute to

one's attitude towards a machine are related to the personal discomfort and concerns in various interaction scenarios [283], which tend to affect the human trust in the machine. On the other hand, machine's characteristics that may affect the human's willingness to collaborate range from machine's appearance (see e.g. [358]) to machine's previous behaviour, such as failure history [61]. As such, distinguishing the levels of competence and willingness for the different interdependencies gives more insight about the different feasible team configurations.

In this paper, we aim at providing a structured analysis of the dimensions of a performer's competence, willingness and external factors and evaluate the feasibility of each possible interdependence relationship. The final decision of which interdependence is better for a certain task is left to the user (and trustor) to decide, as this mainly depends on their perceived risk [121, 173, 361, 362, 405], of trusting and, sometimes, of not trusting, see e.g. [260]. This is related to the formal belief of dependence, as in Falcone et al. (2004) [122], which is how much an agent believes they depend on another entity for a certain goal.

4.3 TABLE FOR ITA

The goal of our proposed analysis through a table is twofold. Firstly, we want a framework that supports team design by providing a more comprehensive analysis of all possible team configurations based on the feasibility of the interdependencies at the atomic task (i.e., a task that is not composed of subtasks) level. Moreover, for each of these interdependencies, we want a framework that analyses the trustworthiness of each teammate for a certain role. For this, we analyse not only the competence/performance dimension, but also the willingness/intention dimension, as well as the external factors that may restrain that action. This explicit information should improve the process of design and decision-making in human-machine teams for task selection and/or allocation, whether this is to be done by a team member or for a team designer (i.e., not necessarily involved in performing the tasks). The table can be found in Fig. 4.1. The parts of the table surrounded by scattered thick dark-red line are to be filled in by the users. Besides those, the table is automatically filled in. In this section, we present the structure of the table and how it can be used.

4.3.1 STRUCTURE OF THE TABLE

ATOMIC TASKS

When there is a team goal, this needs to be divided into sub-tasks, which in turn can be divided into other sub-tasks, repeatedly, until the goal is divided into atomic tasks. We call atomic tasks the tasks that do not need to be broken down into smaller tasks. The analysis of interdependence and trust will focus on each of these tasks individually. They are to be decided by the user (further explained in Section 4.3.2).

POSSIBLE PERFORMER(S)

For each atomic task, we need to consider who can perform it. The possibilities of performing an action are doing it independently, i.e., the human as performer (H independent), or the machine as performer (M independent) or doing it jointly (Joint), as a hard interdependence. Each of these three potential performers will be analysed in terms of dimensions of trustworthiness, for each task.

Atomic Task	Possible Performer(s)	Can? (skills, knowledge)	Will? (intention, preference)	Ext. Factors (opport., resources)	F	Configuration Feasibility					Design choice
						H	H+	H+M	M+	M	
Washing	Joint	✓	✓	✓	✓	✓	✓	✓	✓	✓	machine performer + human support
	H independent	✓	✓	✓	✓						
	M independent	✓	✓	✓	✓						
Peeling	Joint	✓	✓	✓	✓	X	X	✓	✓	✓	mandatory joint
	H independent	✓	X	✓	X						
	M independent	✓	✓	✓	✓						
Chopping	Joint	✓	✓	X	X	✓	X	X	X	X	human performer + no support
	H independent	✓	✓	✓	✓						
	M independent	✓	✓	X	X						
Frying	Joint	X	X	✓	X	X	X	X	X	✓	machine performer + no support
	H independent	X	X	✓	X						
	M independent	✓	✓	✓	✓						

Figure 4.1: Interdependence and Trust Analysis Table for several atomic tasks that compose the major task of *making fries*. Possible performers are the human (H), the machine (M) or both being co-performers (Joint). To each of these possibilities, we analyse whether they have the skills and knowledge to do a task (whether that performer *can*), whether they have the intention and preference to do the task (the performer *will*) and, finally, if the external factors and permissions allow. We can see the resulting feasibility (F) of each performer option and the resulting feasible configurations that leads to. The last column presents the design choice. The areas that are within the dark red scattered lines are the ones that can be altered by the user. Column *F* and *Configuration Feasibility* are automatically calculated.

Dimensions

Based on the literature presented in Section 4.2, we include (1) a belief related to ability, performance, competence (column *can* in Figure 4.1), (2) a belief which comprehends everything besides ability that may contribute to the choice of performing a task successfully, i.e., willingness, benevolence, integrity, and personal preference/motivation for a certain task, and (3) the context which comprehends external factors (opportunities, permissions), in our analysis. We can find the columns *Can?*, *Will?* and *Ext. Factors* on the table.

4.3.2 How to use the table

Scenario

The scenario that was used to fill in this table was inspired by cooking scenarios (used in human-robot interaction studies [151] and human-AI collaboration studies, such as the test bed Overcooked-AI [55]) and consisted of *making fries* with a set of constraints. The constraints were:

- The machine is not allowed to hold a knife (which impedes chopping).
- The human does not want to fry the potatoes, because they are afraid of getting burnt.
- The human does not know how to fry potatoes.
- The human does not want to peel potatoes if they are the only one doing it.

STEP 1: ATOMIC TASKS

The first step for the interdependence and trust analysis is to know what tasks need to be done. As such, users of the table must first agree on which atomic tasks need to be listed in the table. Determining the atomic tasks and their level of detail can be challenging, depending on the scenario. However, the idea is to divide the tasks until the point when it is clear what *Joint*, *Independent H*, and *Independent M* may look like, so that we can assess the possible performers' trustworthiness. We decided that *making fries* includes the atomic tasks of washing, peeling, chopping, and frying. After establishing the atomic tasks, the user should start filling in the areas of the table that are surrounded by scattered thick dark-red line (in Fig. 4.1). In particular, the user should fill in the atomic tasks on the table, in the first column.

STEP 2: ASSESSING TRUSTWORTHINESS

The second step of the ITA analysis is to assess the trustworthiness of the different possible performers, for each task, by signing each dimension with a "✓" if positive or with "X" if negative. For example, when we analyse whether the human can perform the task independently, we should consider whether they can (i.e., have the competences, skills, knowledge...), whether they will (i.e., want to, would choose to do that task) and, finally, if they have the external opportunities and resources to do it (i.e., external factors).

For example, the machine was not allowed to hold a knife, which should make the cells of *external factors* negative (*X*) both for *M (machine) independent* and *joint* in the chopping task. Also, the human did not want to fry, because they were afraid to do it, but also did not know how to. This information should make *Can?* and *Will?* negative (*X*) for *H (human) independent* and *Joint* in *Frying*. Finally, the human did not want to peel potatoes alone, which puts an *X* on *Will?* for *H (human) independent*.

STEP 3: INTERPRET FEASIBILITY COLUMNS

Performer Feasibility (F) column After filling in the table with the trustworthiness information, the column *F* will present the feasibility of each performer, for each atomic task. This feasibility is negative (*X*) if at least one of the dimensions is not feasible (i.e., there is an *X* in one of the dimensions), and positive (✓) otherwise. With the information of which performers are possible for each atomic task, we can infer which configurations are feasible.

Configuration Feasibility column The team configurations are the combinations of possible roles that each team member can take for a certain task, i.e., the different interdependencies that can happen in a task. We consider five possible team configurations (under *Configuration Feasibility* header) for a team composed of one human and one machine. If we consider independent configurations, we can have either a completely independent human performer (H), or a completely independent machine performer (M). There are also two possible soft interdependencies, i.e., human with support (H+), which happens when the human can be independent, but support is possible to increase efficiency or reliability, and machine performer with human support (M+). Finally, there is also a hard interdependence, i.e., mandatory joint (H+M), where human and machine have to co-perform the task. The configurations' feasibilities are inferred from the performers'

Team Configuration	Joint	H ind.	M ind.
human performer + no support (H)		✓	
human performer + machine support (H+)	✓	✓	
mandatory joint (H+M)	✓		
machine performer + human support (M+)	✓		✓
machine performer + no support (M)			✓

Figure 4.2: For each team configuration to be considered feasible, a set of performers need to be feasible as well. This table shows which performers need to be feasible (in F column) for a team configuration to be considered feasible (in Configuration Feasibility column).

feasibilities (see Fig. 4.2). For example, we consider that if *H independent* is feasible, then the team configuration *human performer + no support (H)* is also feasible. Having a feasible joint performer also leads to a feasible *mandatory joint (H+M)* configuration. We infer the supporting roles given that joint is possible, i.e., if joint is possible, support is also possible (see the *peeling* example). In the ITA table (in Fig. 4.1), we can see for each task which configurations are feasible. For example, for *washing*, all configurations are feasible whereas for *chopping* only human performer without support seems feasible.

Step 4: Design Choice

Once the user knows what the feasible configurations are, they have the basic information to make a decision. In the *Design Choice* column, the user can pick one of the configurations for each atomic task. The table does not advise for any design. Depending on the tasks and scenarios, we believe there will be other things to consider when deciding which of the feasible configurations to pick (e.g., workload, values, time to finish task). Although that is out of the scope of this paper, we discuss it further in Section 4.7.1.

4.4 Framework

The interdependence and trust analysis (ITA) framework is the conceptual workflow of the dynamic information used for decision-making in human-machine teams, which serves as input and output for the ITA table. The ITA framework can not only be used by a team designer, with an overview of all tasks and teammates, but also by the teammates themselves, either human or artificial. We envision the ITA framework to be used in two main ways, both having similar but potentially slightly different requirements on the ITA table:

1. *ITA for human use:* A table for human teammates or human team designers to use, which includes the assessment of different trustworthiness dimensions (competence

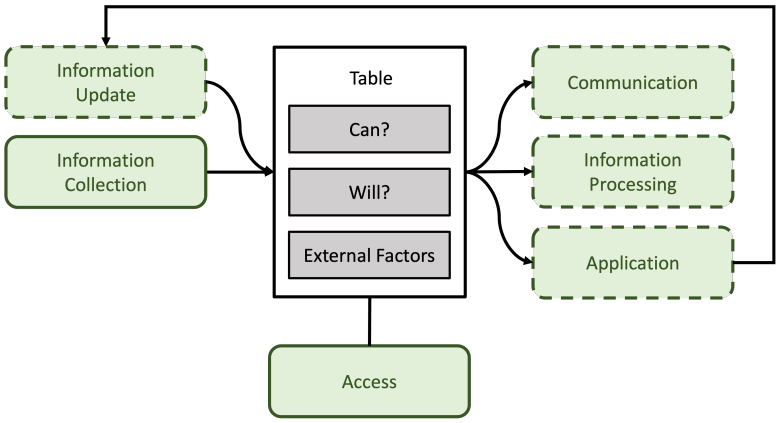


Figure 4.3: The process of decision-making with the help of the table requires information collection (and updates), defined access permissions, and the processing, application, and communication of the structured information of the table. All these modules are entirely dependent on the use of the table and also on what the user/designer prefers.

and willingness), and of one context (external factors) dimension, for different team configurations.

2. *ITA for artificial use:* A computed version of the table (with the same dimensions) which can be used by an artificial teammate or artificial team designer.

Independently of being used by a human or artificially, the framework that surrounds the table is conceptually the same. Fig. 4.3 presents the framework, and the modules that treat the information that goes in and out of the table, in the process of design and decision-making. The two main modules that need to be specified when used the table are *Information Collection* and *Access*. How the information that goes in the table is collected is entirely dependent on the table's use and user. For example, if the table is being used by an artificial agent, the information collection can either be done through sensors and/or machine learning models, or inputted directly by a human. Similarly, who has access to which table is decided by the team designer. It can be that all team members have a table of their own and can also see others' tables, or just one person has a table, for example the team designer, and this table is private.

Furthermore, this table can be attached to other modules, to be decided by the user/developer. In particular, how the information in the table is communicated to other team members is to be decided by the user, e.g. the user can have explanations generated from the values of the table. Similarly, the decisions that derive from this analysis, how this information is processed, and how it is applied, are also up to the user, e.g., there may be a module that uses the table for task allocation. Finally, the consequences of the applications of the table should lead to updating the table's information. This module can be related to how information is collected, but can also be something different entirely. For example, the first time information is collected it may be from a human source (e.g, a manager), but during teamwork sessions (or after), this information can be updated automatically by an algorithm.

There are several potential uses of the ITA table and framework, depending on which modules the users wish to add. Primarily, the Interdependence and Trust Analysis framework is suitable for task selection and allocation. For example, a machine can compute the table and use it to make decisions on whether to support the human or not, or who to call for help for a certain task (see more in Centeio Jorge et al. (2023) [64]). If it is possible for the machine to update its beliefs regarding other teammates and itself according to this structure and representation, this framework can provide transparency and potentiate justifications from the artificial teammate. For instance, the machine can explain that it decided to fry veggies, because it believes that its human teammate is not willing or capable to do it. This can potentially happen either by presenting the table itself or generating text from it.

In fact, the table can also be seen as a formalisation of the information collected by team members and team designers, and it can provide shared mental models and communication (as per [329]). For example, if teammates can share each other's tables with each other, or have a centralised one (at least for certain dimensions), it is possible to see when beliefs are misaligned. To illustrate, perhaps I believe that the external factors do not allow a certain performer to execute a task, but my teammate disagrees. This can be perceived through sharing table information among team members. This being said, this framework also offers a good analysis of dyadic (and possibly team) trust, which can facilitate appropriate and warranted trust among teammates [231] by guaranteeing that the teammates' beliefs are aligned.

4.5 Evaluation

We split the evaluation of the table in two phases. The first phase was composed of two expert interviews with two participants each (dyadic interviews) and was intended to improve the table. The second phase was one focus group composed of five participants, and was meant to evaluate the final version of the table (the one presented in this paper). Before conducting the experiments, we obtained approval from the ethics team of Delft University of Technology (ID nr 3488).

4.5.1 First phase of evaluation

Dyadic interviews provide several advantages, such as allowing the interviewer to observe deeper discussions than in an individual interview [271, 367]. At the same time, it is easier to find available and compatible pairs than groups, which brings an advantage when comparing to focus groups. We ran two dyadic interviews in person, which lasted one hour and a half each. They were composed of (1) analysing interdependence and trust of a collaborative task (to be executed by a team composed of one human and one machine) by filling in our proposed table, and (2) answering six open questions. The presented table was an earlier, extended version of the one presented in this paper (in Appendix A.1). It included the thorough analysis of all dimensions for all five possible interdependence configurations. Furthermore, instead of checkmarks, that version made use of a colour code for feasibility, and the columns had a slightly different name, while referring to the same concepts.

PARTICIPANTS

Each dyadic interview of the first phase of evaluation was composed of two experts who were researchers in the field of human-machine interaction and collaboration. They were two men and two women (one man and one woman in each group), with ages between 25 and 35.

TASK

The scenario presented to these two dyadic interviews was very similar to the one presented in Section 4.3.2. However, instead of making fries, participants were told the scenario was about frying veggies, which did not include the peeling task (as in Appendix A.1). The set of constraints were *"Both machine (M) and human (H) can wash and are willing to do it. However, the human is not willing to support, though, as she thinks it is not necessary. For safety reasons, the machine should not use the knife. Finally, the human does not know how to fry, and she is scared of it too, but can help, and the machine can only do it with help of others"*. Participants were explained the dimensions and interdependence configurations included in the table and asked to fill it in together, without being presented with an example beforehand. Furthermore, the questions (inspired by Krueger et al. (2002) [217]) we asked participants at the end included "What is the one thing you liked best/least?", "What would you change/keep in the table?" and "In which situations would you use/not use the table?".

DATA PROCESSING

This first phase of evaluation served as a pilot and no structured analysis was made. The authors went through the experts' comments during the tasks and their answers to the open questions, and summarized the most predominant comments. These comments were then used to improve the table.

4.5.2 SECOND PHASE OF EVALUATION

After the first phase of evaluation, the table was changed according to the feedback received, taking the shape that we present in this paper. The second part was aimed at evaluating the current table's final usability with a use case in the domain of firefighting. We opted to do this evaluation online, through MS Teams, since (1) we included participants from different physical locations and (2) it was easier to collect and process the transcripts. The session lasted one hour and a half.

PARTICIPANTS

This focus group counted on five participants, three men and two women, with ages between 25 and 55. Two of the participants were firefighters, and the other three were researchers in the field of human-agent teamwork (applied to the fields of firefighting, military and manufacturing), with backgrounds in Psychology and Computer Science.

USE CASE

For a more realistic evaluation of the table, we looked for a scenario where a human-machine team is currently developing. As such, two of the participants worked in a fire department which has been moving towards more autonomous solutions in recent years.

In particular, they have a robot, which we will call *Rob* for simplification, which is capable of moving, recording in real time, extinguishing fires, among other things. Rob is currently controlled by another firefighter through a tablet. The use case in this focus group was based on the possibility of having Rob moving autonomously. We believe that people that are already dealing with the challenges of such teams can give better insight regarding the usability of our table, including the positive and negative aspects of it.

Task

Participants started by being presented to the main concept of interdependence and the different interdependence configurations in a human-agent team. After this, we presented our table pre-filled with the cooking example, which was presented in the first phase as the main activity. Finally, participants were given the use case and twenty minutes to complete the table together, regarding the presented use case. They were asked to think out loud.

In particular, the participants were told "*Let's say that we have the situation of a building with fire, and you need, as a team, to locate people inside the building. So the subtasks of this task are moving in general, which is composed of choosing where to move, i.e.,* ***planning the trajectory****, and also the actual* ***movement****; and clearing the spaces, i.e.* ***scanning/observing*** *and* ***processing what is scanned/observed****. Imagine that there is a team composed of a firefighter and Rob, the robot. The environment does not allow the human firefighters to go in, you can imagine that it can be for several reasons. Imagine also that Rob can go in and has autonomy. In particular, Rob can move autonomously, but it can also be teleoperated (i.e., the human choses the trajectory). It can also scan the environment around and provide some analysis into the scans. However, the scans should be checked by the human firefighter as well.*" The task and subtasks can be found in Fig. 4.4.

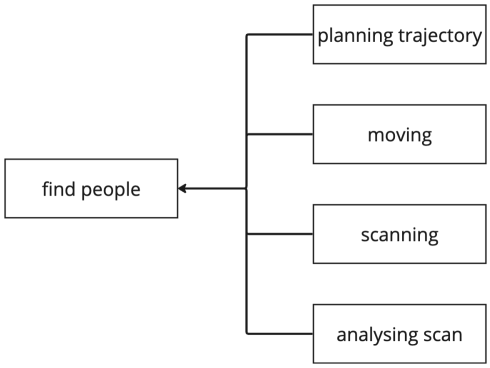


Figure 4.4: Task of finding people divided into subtasks (which are also atomic tasks).

After twenty minutes, the participants were asked the following questions (inspired by Krueger et al. (2002) [217]):

Q1 What one thing do you like the best?

Q2 What one thing do you like the least?

Q3 Under what circumstances would you use the table?

Q4 Under what circumstances would you not use the table?

DATA PROCESSING

To analyse this second phase of evaluation, we collected transcripts and ran a thematic analysis [44]. The transcripts were divided into five parts, each originating a different coding scheme. We divided the transcripts collected during the activity, and then for each question, Q1-Q4.

The first author and a double-coder (non-expert) went through the transcripts and wrote down some codes that came to mind related to comments or questions that may affect the usability of the table. Both coders met to discuss the codes and reach an agreement on the coding scheme. After agreeing on the coding scheme, both coders coded the utterances separately. Both coders met one final time to agree on the coding. During this meeting, some codes were merged.

4.6 RESULTS

4.6.1 FIRST PHASE

In the first phase of the evaluation, most participants showed great interest in using our table for their personal research works. Among other things, participants mentioned our table would be useful in the process of designing their experiments' tasks, calibrating appropriate trust between humans and machines, and designing explanations. We received negative feedback mainly based on the colour code of that version of the table, the inefficiency related to filling in the table, and possible overlapping of dimensions. In particular, it was clear that filling in the first iteration of the table was quite overwhelming for human participants. All feedback from this phase is already integrated in the version of the table presented in this paper. The main change was reducing the size of the table. More concretely, in the first version (in Appendix A.1), we assessed the three dimensions for each possible role (performer with support, independent performer, co-performer, supporter, not involved) for both human and machine, which gives a total of 24 cells to fill in per task. In the second phase, we assess the dimensions only for the possible performers (human independent, machine independent, co-performers) and assume the support feasibility (as explained in Section 4.3.2).

4.6.2 THEMATIC ANALYSIS (SECOND PHASE)

The results of the evaluation of our framework are in the feedback given by the participants during the second phase of evaluation. All utterances can be found in our dataset [66], published in 4TU.ResearchData. We structured this feedback through a thematic analysis, which shows the topics that were brought up throughout the activity and question answering. We calculated the inter-rater reliability for the thematic analysis, resulting in a Cohen's kappa [221] of *0.65* (ran with R package *irr*[139]). This value can be considered *substantial* [221] or *moderate* [253]. Because the double-coder was non-expert, and we allowed for more than one code per utterance, this value was considered sufficient to proceed to the analysis. The coding schemes can be found in Fig. 4.5-4.9, with a respective example of a selected participant's quote (all quotes available in the dataset [66]). The respective counts of each code can be found in Table 4.1.

Table 4.1: This table shows the number of utterances that were attributed with each of the codes (some utterances were attributed more than one code), and the number of participants that had utterances related to each code. It also shows the total number of attribution of codes in utterances of a certain phase (i.e., during activity, Q1, Q2, Q3 and Q4).

Code ID	Code name	Code count	Phase count	Participants
A1	definition of dimensions	10		4
A2	definition of role	23		4
A3	definition of subtasks	11	50	4
B1	decision-making	4		3
C1	answer granularity	2		2
D1	good structure	4		3
D3	clarity	1		1
D4	dimension	2	11	2
D5	role	1		1
D6	level of detail	2		1
D7	agreement with final results	1		1
E1	unclear definitions	4		3
E3	missing evaluation criteria	2	9	2
E4	context-dependent	3		3
F1	supports planning	5		4
F2	discussion starter	3	12	1
F3	robot design	4		4
G1	rapidly-changing situations	1	4	1
G2	different mindsets	3		2

ACTIVITY

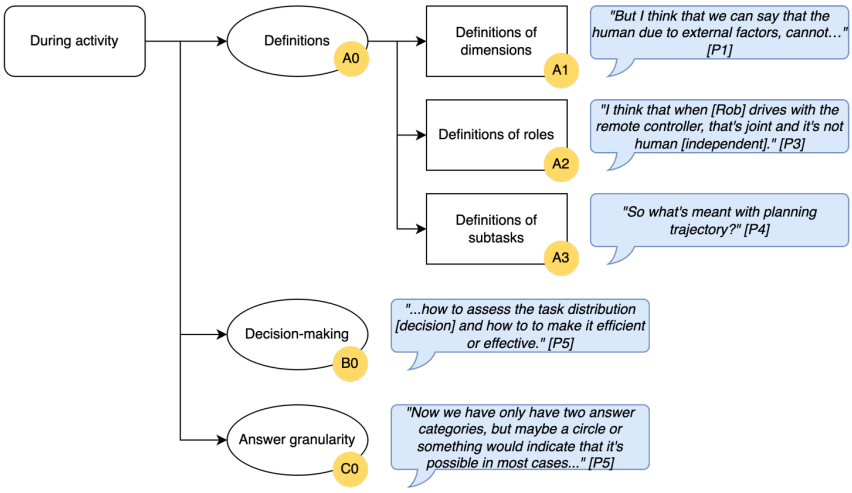


Figure 4.5: The coding scheme related to the transcripts collected during the activity.

During the activity phase, participants were invited to ask questions about the concepts or instructions they could not understand, while being presented the activity. Predominantly, there were questions and discussion regarding the definition of the different **dimensions** (10 utterances), the different **roles** (23 utterances), and the task and its **subtasks** (11 utterances). These continued after the instructions were given, and when the participants were filling the table. All of these codes constitute the theme **Definitions (A0)** (in Fig. 4.5 with respective codes and quotes), which then counts with a total of 34 utterances (as in Table 4.1). In table 4.1, we can also see that each of the codes in theme **A** was attributed to at least four of the five participants.

In particular, the participants had difficulty analysing **dimensions** separately, i.e., not making their analysis of one dimension dependent in another. This can be illustrated by what P4 said, *"I don't know how the "will", the intention, if the external factors weren't there, then he or she would have the intention to do that, but I don't know how to understand the "will have to", whether they could take external factors into account or not."*. The group also showed difficulty in distinguishing the different **roles**, which can be exemplified by what P3 said, i.e., *"I think that when Rob drives with the remote controller, that's joint [performer] and it's not human [performer]"*. Finally, as P1 said *"Yeah, but that's planning trajectory [and not moving], turns out."*, the participants showed surprise and difficulty in distinguishing the different **subtasks** and what they involved. Often times, confusion regarding subtask definition led to confusion in roles and even dimensions, which meant that several utterances in this phase were coded with more than one code of theme **definitions (A0)**.

Besides verbalizing difficulty with definitions, several participants also gave their opinion on the framework, both while receiving the instructions and filling in the table, sometimes adding suggestions and asking deeper questions about the use of the framework. These were mainly about the decision-making process (**theme Decision-making (B0)**),

which reflected two main concerns from three participants: what information distinguishes two or more feasible options (to do a certain task) when someone needs to make a decision using the table, and how to optimize the decisions made, and how to evaluate the decisions once they are made. Furthermore, two participants suggested that the table could have higher **answer granularity (C0)**, allowing for answers besides yes or no.

Positives (Q1)

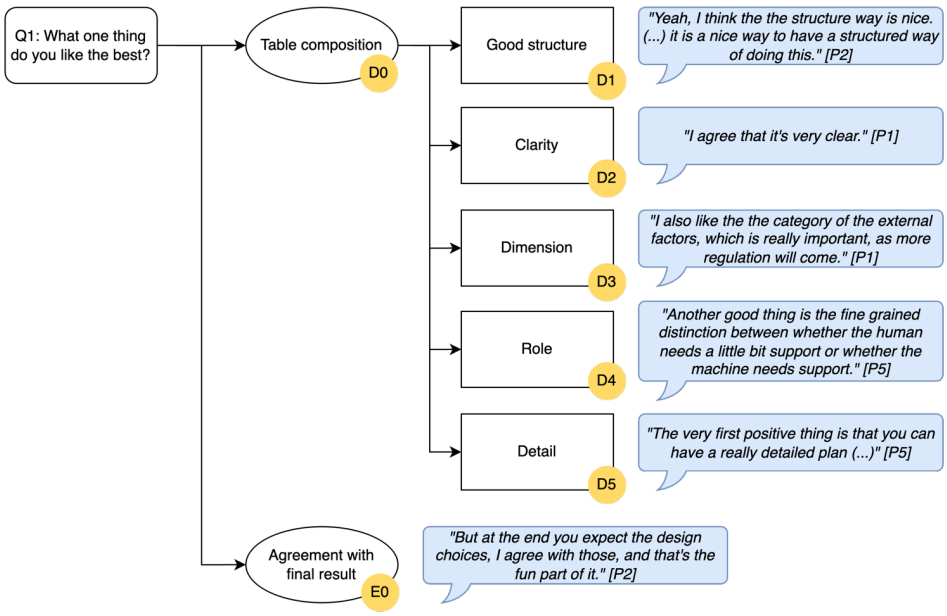


Figure 4.6: The coding scheme related to the transcripts that answer the question *"What one thing do you like the best? "* (Q1). The blue speech balloons show an utterance that was coded with the corresponding code.

In Fig. 4.6, we can see the codes and exemplary quotes of the answers to Q1, when we openly asked participants what they liked the most about our framework. Participants mainly mentioned elements of **table composition (D0)**, which counted with 10 utterances. These included compliments to the **good structure** of the table, the **level of detail**, and its **clarity**. Although in the previous phase, participants showed some difficulty with the definition of the different **dimensions** and **roles**, they mentioned these elements as positives of the framework. One participant also mentioned that they liked that, although there was a lot of discussion, in the end, they all **agreed with the final result** of the table.

Negatives (Q2)

When questioned about the things they did not like (Q2), participants mentioned the **unclear definitions**, which aligns with the results we got in activity phase, where participants discussed the meaning and distinction of roles, dimensions and subtasks. Furthermore, they also recalled the need for **evaluation criteria**, which also reflects the decision-making (B0)

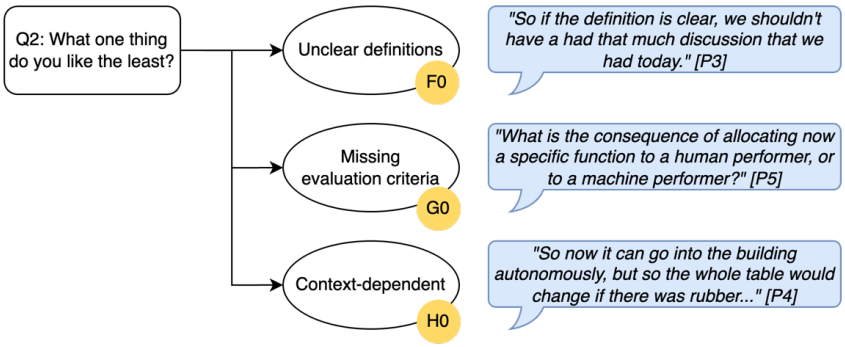


Figure 4.7: The coding scheme related to the transcripts that answer the question *"What one thing do you like the least?"* (Q2). The blue speech balloons show an utterance that was coded with the corresponding code.

in activity. Lastly, participants also showed concern regarding the **context dependence** of the table. All codes and exemplary quotes can be found in Fig. 4.7.

When to use (Q3)

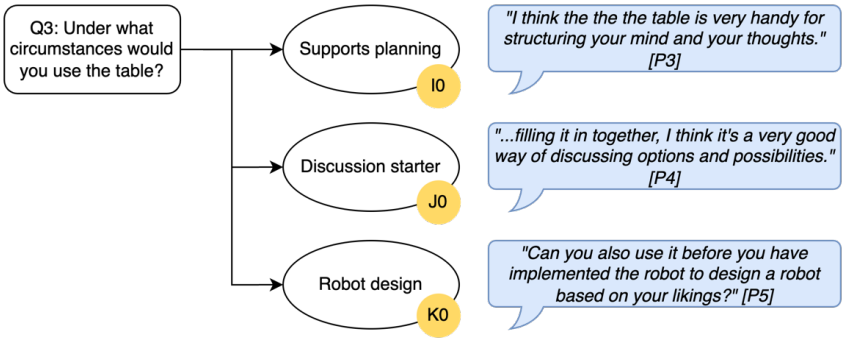


Figure 4.8: The coding scheme related to the transcripts that answer the question *"Under what circumstances would you use the table?"* (Q3). The blue speech balloons show an utterance that was coded with the corresponding code.

Four of the five participants verbalized that our framework **supports [teamwork and task] planning** and that, similarly, it can be used to **design the robot** or AI required for a specific human-machine scenario or task. One participant also mentioned that the use of the framework is a good **discussion starter**. These codes can be found in Fig. 4.8.

When not to use (Q4)

When asked when they would not use the table, participants were less verbal. However, one participant referred they would not use the table in **rapidly-changing situations** (related to the context-dependency, H0, concerning Q2). Two participants also mentioned

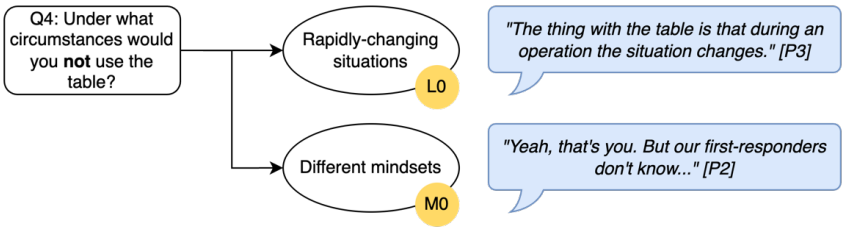


Figure 4.9: The coding scheme related to the transcripts that answer the question *"Under what circumstances would you not use the table?"* (Q4). The blue speech balloons show an utterance that was coded with the corresponding code.

that it might not be feasible to use with people with **different mindsets**, meaning that some workers may not have the capacity to sit down and use such a framework beforehand. Fig. 4.9 shows the exemplary quotes.

4.6.3 Summary of results

Our results were overall positive, counting with more positive comments than negative in the open answers. The participants were able to use the table as intended and could agree on the final result (code E0). They found it useful for planning (I0), discussing possibilities (J0), and designing artificial teammates (K0). Participants also extensively complimented the table composition (D0), including the chosen dimensions (D3) and possible configurations (D5).

However, there were also some persistent concerns reflected in the participants' comments and questions. They were mainly concerned with (1) the definitions of the dimensions and interdependencies when related to a certain task, (2) the process of filling the table, in particular its (in)efficiency and comprehensive information, and (3) the use of such information not being enough to make decisions. We discuss these results in the next section.

4.7 Discussion

4.7.1 Reflection on results and theoretical implications

External Factors

Our results build on existing evidence (as in Johnson et al. (2014) [194]) that an interdependence analysis can help human-AI team designers identify interdependence relationships in a joint activity. Four out of five participants mentioned that such a tool supports teamwork and coordination of offline planning. Other works have proposed automatic ad-hoc planning for human-AI teamwork, based on teammates' task-based competence [16, 25], and role preference [288, 289]. Our results defend that environment characteristics (external factors dimension) should also be included in task-allocation methods as ethical concerns increase, as well as the appearance of new laws related to these. In particular, machines should not make all the decisions. This is supported by two of the participants' interventions, which mentioned that the dimension of the external factors was helpful for task

planning. P5 said *"The very first positive thing is that you can have a really detailed plan on what sub functions are needed in order to accomplish an overall task and based on capacities or as you called it, external limits or the external environment, you have a good indicator of where you bet your money on."*. Furthermore, P1 said that *"I also like the category of the external factors, which is really important, I think, as more regulation will come also in terms of the communication to, for example, people who want to deploy like, see opportunities in human machine teams, but then due to the AI act, it's not possible any more."*, corroborating that this dimension should be included in human-AI teamwork design and planning. This is aligned with [386], which presents a dynamic moral task allocation method, and implies that further research on how to integrate ethical and legal boundaries in human-machine teamwork is required.

COMMUNICATION

As just seen, it can more and more often happen that an AI teammate is capable and willing, but not allowed, as P1 said, *"It's a very nice distinction for communicating that, yes, it's able to, and it's willing to, but we just cannot let that AI do that right now"*. This brings our attention to the need to communicate the different dimensions that contribute to a machine not being able to perform a certain task. Although communication does not appear explicitly in our codes, it does come implicit in some of them, all of them mentioned as positive characteristics, such as clarity (D2), agreement with final result (E0), and discussion starter (J0). Although there has been an effort to explain the automation's mental states to the human during and after collaboration (see e.g. [155, 239, 368]), bidirectional communication should be explicitly included in teamwork design and task planning methods. Further research is required to investigate how this can be done naturally between the human and the machine.

DEFINITIONS

Both during the presentation of the table and the activity, participants asked about the definitions of dimensions, roles and subtasks (A0-A3 and F0). In particular, the most predominant concern was related to the definition of a certain role for a certain task. For example, P1 said, *"Teleoperating sounds like human support and not joint."* We recognise a difficulty in defining what *support* and *co-perform* means, depending on the task. In the case of the task *movement*, the robot can be teleoperated or move autonomously. If a human teleoperates the robot, does it mean the human and the machine do it jointly (they're co-performers)? Or is the human the only one performing this action? Or one is performing and the other supporting, and if so who is what? We believe this difficulty comes mainly from a lack of precision on what each task means. In these cases, users should try to divide the tasks into even smaller subtasks so that the roles become clearer. For example, perhaps if we were to have an atomic task *deciding next movement* and another *physically changing positions*, it would be clearer that both the robot and the human can decide the next movement (independently) but only the robot can actually change its own physical position because the human is not physically there to do it. Our findings suggest that existing human-machine models need to be updated in order to incorporate team configurations that are not binary (e.g., leader vs followed as in Noormohammadi et al. (2022) [288]). These human-machine models should also account for the fact that team

members having different natures alters the interpretation and meaning of the task (e.g., works that assume the same task definition for humans and machines [16, 25, 194]).

Filling in the table

In the famous Technology Acceptance Model (TAM) [99], Davis proposes that perceived usefulness and perceived ease-of-use are the two main factors that influence the actual use of a technology. The author defines perceived usefulness as "the degree to which a person believes that using a particular system would enhance his or her job performance" and perceived ease-of-use as "the degree to which a person believes that using a particular system would be free of effort" (p. 320). Our results show that it is not always easy to increase usefulness without increasing effort, and vice versa. In our case, asking the user for more information to include in the table increases usefulness (as more information can lead to better decisions), however, more information means that the user needs to fill in a higher number of cells in the table, which leads to a higher effort to the user, which decreases the perceived ease-of-use.

In order to increase the perceived ease-of-use, we decided to make the cells binary (check or cross), mostly because we believed this would be easier for a person than to come up with a value from a certain range (let's say from 0 to 5). Interestingly, this was pointed out by a participant (C0), who felt the need to express something other than the possible answers (check or cross). However, of course a binary value gives us way less information than a wider range (which may decrease the perceived usefulness). Similarly, although participants were generally happy with the level of detail (D5), it was also brought up, as a negative aspect of the table, that the table is context-dependent (H0), i.e., that changing the context means changing the values in the table. We understand how a user can perceive this as a negative point, since they would have to fill in the table again if the context changes, decreasing their perceived ease-of-use. However, it is hard to make the table context-free (which improves the ease-of-use) while not decreasing its perceived usefulness, i.e., having enough information about team members' competence, willingness and external factors, in a specific context. For example, a machine may be allowed to hold a knife if a human is not present, but not otherwise. This means that the value in the external factors dimension is going to change depending on the context, requiring more information. Nevertheless, we need that information to know whether we can give the task of chopping potatoes to the machine.

Actually, the fact that the table is context-dependent increased the perceived usefulness according to other participants (D3), for example, we have mentioned before that some participants highly appreciated the external factors dimension of the table. The external factors dimension is the most context-dependent dimension of the table. Trying to accommodate the HO, we believe it is possible to reduce the effort from the users in real-time, so that the users do not to change the values in the table whenever the context changes. This can be done by discriminating beforehand all possible contexts and including this context in the atomic tasks. For example, we could have the task *holding a knife when a human is around*, which will not change depending on the context (since it is in the description of the task). However, we still need a way of deciding which atomic tasks are used, which still depends on the context.

With the two examples given in this subsection, i.e., the level of detail in the cells of the table and the context-dependence, we realise that some participants may feel like they need

to put in a lot of work before the table becomes useful. Furthermore, as system designers, we also need to consider that requiring loads of information from the user does not only harm the user's perceived effort, but it may also decrease the overall efficiency of the system (since it may take a lot of time to disclose this information, for example, which may not be possible in real-time). This poses a challenge to AI designers that need to ensure the compliance with new regulations of transparency, traceability, and accountability that are emerging around the world, e.g. the European AI Act[1], and the IEEE Ethically Aligned Design[2]. If we want to have a reliable and transparent framework, that acknowledges its context and adapts to the circumstances, we may have to disclose a high load of contexts and nuances, which require a high effort from the user. This information is crucial to properly explain and justify decisions made by agents that possibly use the framework, such as explaining that they cannot help with chopping because a human is around, and that in such cases they are not allowed to hold a knife. We expect these challenges to be more and more present in the development of human-machine systems and collaboration design, as regulations become stricter. These findings show that there is a need for researching guided conversation to fill in these tables (also [193, 194]), making the process of human-machine team design effortless to the human, while guaranteeing ethical compliance.

MAKING DECISIONS

The final topic of concern had to do with decision-making itself. Although participants saw value in the table to help to make decisions (I0, J0), it was mentioned that an evaluation criteria was missing. Participants felt the need to have further information about what was the goal of the task allocation, as well as what each subtask meant for the achievement of that goal. For example, P5 asked *"What is the consequence of allocating now a specific function to a human performer, or to a machine performer?"* Indeed, there may be different objectives when allocating tasks in human-machine teams, including reduction of the non-ergonomic human task, productivity, and human satisfaction [287]. We believe this information is important, but we decided to keep it out of this version of the table. The main reason is that the information that is necessary to make decisions, such as expected workload for each teammate, or total time per team configuration, is hard to predict and obtain. In fact, for the first phase of evaluation, we prepared a mock side table with this type of information, to be used before the decision-making step. We learnt from the participants that this would be very hard to actually obtain for real-life scenarios, e.g., how to calculate the workload of a human chopping ten potatoes? As such, this poses entirely different questions than the ones we are trying to answer in this paper. These findings suggest that existing task allocation and decision models (such as [16, 35, 378]) should include human-centred factors for optimization and utility calculation (reward and penalty), such as accounting for the system values (see e.g. [165]) and major risks, which should be changeable from context to context.

SUMMARY OF THEORETICAL IMPLICATIONS

Results show that experts value our framework and believe it supports human-machine teamwork planning, discussion and design. The table is successful in improving communication and supports the team design with machines (and artificial intelligence) with

[1] $https://artificialintelligenceact.eu/$
[2] $https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf$

variable levels of autonomy and permission (which is independent of the machine's capabilities). This suggests that Interdependence and Trust Analysis could benefit planning and designing of human-machine teams for different contexts such as disaster response [100], search and rescue [326], cooking [151], driving [24], and healthcare [96]. However, we found that the interpretation of the table's roles and subtasks can be challenging and subjective, which suggests that tasks' and roles' definitions should depend on the natures of the agents involved (e.g. different embodiment and cognitive characteristics lead to different meanings of *support*). We also see a trade-off between efficiency (effortlessness) and completeness (usefulness), as participants appreciate the level of detail and information of the table, but are not very happy about having to provide so much information during the analysis. This suggests a need for more natural communication to fill in the table (instead of going through the whole table cell by cell), such as guided dialogues. Finally, the focus group showed a need to include different optimization human-centred policies, depending on the context, in existing decision-making models for human-machine collaboration. For example, in some scenarios it might be important to reward a certain value (e.g. safety) more than others (e.g. privacy), or simply maximize for user's satisfaction.

4.7.2 Limitations and future work

Our work and method present some limitations that also open ways for improvement in future work. As we mentioned earlier in this section, there is a trade-off between efficiency (effortlessness) and completeness (usefulness), which may impact the reliability and adaptability of the system. This being said, we had to compromise on the amount of information included in the table. Such decisions also led to assumptions which may be seen as limitations. In particular, we had to reduce the rows of the table, which led to assuming that the feasibility of an agent's support could be inferred from the feasibility of that agent's independence. Ideally, we would have a separate analysis for support, but that proved to be overwhelming to participants. However, there may be applications in which an extended version of the table (as in Appendix A.1) can be more suitable, and so users can still use a broader version of the table. In fact, in future work, we would like to implement artificial agents that use the table as a support for task selection and allocation (stage 2 in 4.4). In such cases, the agent needs to form and update beliefs about all dimensions, all teammates, all tasks, and respective interdependence roles. It is surely less overwhelming for an artificial agent than for a human to deal with a bigger table, while at the same time more necessary, as that table will explicitly include the important information. After stage 2, we also want to explore learning algorithms that update the table automatically throughout interactions. Furthermore, we would also like to implement an automatic generation of explanations and/or justifications for the AI that makes use of this framework for task allocation or selection.

Another possible limitation is that this interdependence and trust analysis (ITA) table assumes that an agent that uses it has full knowledge, i.e., enough information regarding all agents and all dimensions, to fill in the table. We have not accounted for cases in which there is no such knowledge, and what that means in terms of feasibility. In future work, we would like to accommodate this option. It would also be relevant to find a way to represent the accuracy of each cell. For example, perhaps an agent believes that the other can do a task independently, but is not 100% sure. This may affect their future decisions, so it should

be represented, as it will affect the future risk of the decision. Overall, risk is not included in this analysis. Besides accuracy, we can also see the risk of going for a specific design choice, and even the risk of *not* going for a specific design choice. It has been mentioned that one of the participants' concerns was how to make a decision after knowing which configurations are feasible. We have mentioned that there might be several criteria that would prioritise some choices over others, and risk is definitely one of them. However, risk is also hard to calculate and assumes there is knowledge for that, so for simplification, we did not include it.

Finally, the thematic analysis used for evaluating the table has its limitations, such as possible bias and dialogue manipulation of a more leading or dominating participant [156]. Although we made sure to ask the questions to each participant, one's answers are naturally affected by the others'. The focus group of the second phase of the analysis was composed of five people, which is considered by some authors to be enough [156], but some others consider it to be too small of a group [270]. We acknowledge the limitation of the small sample size of the two dyadic interviews and the focus group. Although this method gives us an initial understanding of how experts perceive our framework, further research is necessary to study the extent to which these insights transfer to other groups of human-machine team designers. The thematic analysis inter reliability was considered sufficient, as the double coder was not an expert in human-machine teamwork and utterances allowed multiple codes. Most disagreements were in the cases of multiple codes, especially in the ones that included definitions of dimensions, roles or subtasks (theme *definitions*). For example, in the quote "So in case of or for the subtask peeling, the human doesn't want to... but how is it then a mandatory joint?", we see dimensions, tasks and roles being mentioned. It can be hard to decide what is the most important code(s) for such utterance. In future work, we want to run a more objective evaluation of the table in a more involving scenario.

4.8 Conclusion

In this paper, we present an extension of the Interdependence Analysis for human-machine teams. Our approach includes a discriminated analysis of the trustworthiness dimensions of competence (i.e., skills, knowledge), willingness (i.e., intention, preference) and external factors (i.e., opportunity, resources), for each possible team interdependence configuration, for each subtask. This table can support the design of human-machine teams, including the allocation of tasks. In fact, it can also be used for decision-making of team members, either human or machine, supporting task selection too. By using this table as a shared mental model, decisions may become more transparent, justifiable and interpretable, which may lead to an increased and appropriate trust among teammates.

Acknowledgements

Thank you to all the participants of the focus groups, to Matt Johnson for discussing this work with me at its early stage, to Mohammed Al Owayyed for double-coding, and to Ruben S. Verhagen for helping us with the use case. We would also like to thank Delft AI Initiative and the TAILOR Connectivity Fund. Similarly, it is based upon work supported by the National Science Foundation (NWO) under Grant No. (1136993), and by

the European Commission funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant 820437). The support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these institutions.

Data Availability Statement

Our data, including the transcripts and the coding scheme are published [66] in 4TU.ResearchData, as https://data.4tu.nl/datasets/998296ec-7696-4180-8d7e-9af8588b1182/1.

5

Willingness-based Task Allocation in Human-Machine Teams

In human–machine teams, task allocation algorithms typically prioritize competence and other performance-related factors, overlooking human willingness to perform specific tasks. A recent framework proposed incorporating both competence and willingness into task allocation, but its effects on teamwork outcomes remain untested. This paper investigates the impact of willingness-based task allocation in a simulated search and rescue (SAR) setting. We conducted a 2x2 user study (N=72) in a 2D grid-world SAR scenario, where human participants collaborated with two virtual robots, which either used a willingness-based or a balanced task allocation strategy. Half the participants were able to change the task allocation, prior to the start of the mission. Participants preferred robots and plans that accounted for their preferences, yet willingness-based allocation did not significantly influence trust in the robot or perceived team fluency, as these remained high throughout. Moreover, allowing changes to task allocation improved human performance. Our findings also suggest that the benefits of incorporating willingness may be context-dependent, with greater potential in lower-stakes tasks and longer-term teams.

This chapter is written in American English and it is partly based on a workshop publication:
Centeio Jorge, C., de Visser, E. J., Tielman, M. L., Jonker, C. M., & Robert, L. P. (2024). Artificial Trust in Mutually Adaptive Human-Machine Teams. In Proceedings of the AAAI Symposium Series (Vol. 4, No. 1, pp. 18-23).
This chapter is also submitted for publication as follows:
Centeio Jorge, C., Jonker C. M., Robert, L. P. de Visser, E. J. & Tielman, M. L. I know you're capable, but are you willing? Allocating tasks in human-machine teams. (Under review)

5.1 INTRODUCTION

Imagine a search and rescue (SAR) scenario, where a robot equipped with basic communication tools and navigation collaborates with a human firefighter who can do those things too. To operate effectively as a team, they must coordinate, which includes deciding on task allocation and interdependencies. The *Interdependence and Trust Analysis* (ITA) framework [69] proposed a way to evaluate the different possible interdependent solutions in a human-machine team, based on three dimensions: the team members' competence, the team members' willingness, and external factors (such as permissions or physical obstacles). ITA is an extension of the *Interdependence Analysis* by Johnson et al. (2014) [194], which is based on the dimension of competence (named in the paper as capacity), and it is motivated by the hypothesis that when working with humans knowing their competence is not enough. Instead, the authors say that it is necessary to know the human teammates' preferences and motivations (i.e., willingness) for sustainable teamwork. This dimension has been overlooked in the task allocation and planning literature in human-robot collaboration and, while ITA has proposed it, its effects on actual teamwork and the human collaborator have not been empirically tested. This paper investigates the impact of incorporating a human teammate's willingness into task allocation through a user study in a 2D SAR simulation.

In human-robot teams, employed in situations ranging from search and rescue [100, 326] to domestic settings [151], each teammate has a set of attributes that make them more or less suitable for specific tasks. Accurately modeling these attributes can improve machine task selection [85], the overall process of team design [194], and task allocation [16]. One frequent way of allocating tasks is by focusing on the competence or other performance-related characteristics of each teammate for the different tasks, maximizing the physical and cognitive capabilities of team members in relation to task demands [51, 236]. However, as humans, we choose tasks not only based on our competence to do them, but as a result of a trade-off between cost and benefit [273], such as perceived effort and reward [406], which is intertwined with how we feel [369]. This suggests that, when modeling task selection and allocation between humans and machines, we should also consider the human's willingness to do the different tasks. However, although task allocation models in human-robot collaboration have begun to consider human preference and adaptability (see, e.g., [107, 427]), the effects of these task allocation strategies on the human collaborator remain unexplored. In addition, when we consider dimensions other than those directly related to perfirmance, task allocation becomes more complex to calculate and optimize, and it is nontrivial to decide on the weights for each dimension [85]. This work presents an effort to understand the weight and context-dependence that human task willingness should have in human-machine task allocation.

The main goal of any team is to be efficient and driven by trust, communication, and shared mental models [329]. These driving mechanisms are influenced by the machine teammate's behavior, including its decision-making, and the outcomes of those decisions [32, 216]. On the one hand, trust and communication affect each other, e.g., strong communication can build trust, and vice versa [335]. On the other hand, both trust and communication are shaped by shared mental models, which include an understanding of each teammate's characteristics and how those characteristics inform their decisions [77, 199], such as task selection and allocation. The ITA framework [69] suggests that a

mental model with the human characteristics of competence and willingness for different tasks is a good basis for a task allocation strategy that potentiates effective teamwork. However, we lack evidence on how willingness-based task allocation affects human trust in the artificial teammate, their overall satisfaction with teamwork, and the actual effectiveness of the team.

This paper presents a 2x2 mixed-design user study (N=72) simulating a search and rescue (SAR) scenario in a 2D grid world, where human participants collaborate with a virtual robot to transport all victims to a safe zone. Each participant works with two different robots in separate missions. Before each mission, the robot presents a task allocation plan, either willingness-based or baseline. After each interaction, participants report their perceived trust, trustworthiness, and team fluency. The main contributions are: (1) the design and implementation of a user study on human–robot teamwork, (2) a quantitative analysis of the effects of willingness-based task allocation in trust and team fluency, and (3) a quantitative and qualitative analysis of the contextual role of willingness in task allocation. The remainder of the paper is structured as follows: Section 5.2 reviews the background and related work, Section 5.3 describes the task and study design, Section 5.4 presents the findings, Section 5.5 discusses them, and Section 5.6 offers concluding remarks.

5.2 Background

In human social interaction, we often evaluate each other along two key dimensions: competence (related to ability) and willingness (related to warmth) [133]. These dimensions help us decide, for example, who to take as partner in collaborations. Similarly, in multi-agent systems, trust is defined as a belief in another agent's competence and willingness to perform a task [122], or in organizational trust models, as a belief in someone's ability, benevolence and integrity [250]. How important competence-related characteristics are compared to warmth-related ones depends on the task and context [177].

Azevedo-Sá et al. (2021) [25] and Ali et al. (2022) [16] propose task allocation models for human-robot collaboration based on trust modeling. These models focus on matching the team members' capabilities with the task requirements, focusing on expectation maximization of performance. This is aligned with the classical work of Fitts (1995) [134] presented the MABA-MABA list (men-are-best-at, machines-are-best-at), which outlined human and machine strengths across functions such as memory, speed and judgment. This work set the base for several task allocation methodologies to be competence-driven, with the main goal of increasing task performance, especially in manufacturing and assembly line contexts [51, 243, 410]. Furthermore, recent works suggest that task allocation and planning methods should account for human ergonomics [130, 242, 263], which influence human competence over time, and make dynamic allocation and planning methods to be necessary [17, 302, 409]. These methods generally keep updating task allocation during the interaction, based on the perceptions of the robot teammate's competences or capabilities. Similarly, Ramachandruni et al. (2024) [310] suggests a method for mutually adaptive task planning, where the robot observes the human and adjusts its plan around what the human is doing, based on computed capabilities and costs with a hierarchical task network.

The optimal solution in human-robot task allocation, however, may not be the one that prioritizes competence-related characteristics [370]. In fact, Lohrmann et al. (2024) [236]

demonstrates that when robots understand human cognitive tendencies and de-emphasize reward-maximizing behavior, the human-robot collaborative outcomes improve. In this work we are interested in understanding what makes people willing to take and succeed at a task, besides their competence. This willingness is often linked to motivation and to a cost-benefit trade-off, i.e., individuals weigh the effort required against the reward offered and decide whether that task is worth pursuing [181, 212, 273]. Furthermore, people compare their cost or benefits of actions with those of others and make judgments about fairness [110]. Although people sometimes engage in unpleasant activities for long-term benefits to support long-term welfare, consistently prioritizing disliked activities without positive experiences can be demotivating [369]. Motivation, especially when coming from personal interests and values, leads to task preference, which is also correlated with higher effort and better outcomes [176, 208, 339]. Literature also shows that incorporating human preferences into planning can improve perceived likability and intelligence [245] of interactive robots. These findings suggest that humans willingness (including preferences and motivations) to do different tasks should be included in task allocation methods.

Recent studies have begun to incorporate human preferences into task allocation frameworks. Centeio Jorge et al. 2024 [69] (also found in this dissertation's Chapter 4) proposed a formal framework that incorporates both competence and willingness dimensions into task allocation and selection for human-machine team design, but the method does not provide any final solution. Dhanaraj et al. (2024) [107] introduced a method for preference elicitation to support scheduling algorithms, although they did not evaluate the effects on users, and Zhao et al. (2023) [427] proposed a learning-based approach to infer preferred tasks during human-robot collaboration. While these methods reflect an increased promising focus on preference, they have only been evaluated through simulation, or pilots with humans. In summary, existing approaches to task allocation in human-robot collaboration have begun to consider preference and adaptability, but the impact of including human's task-based willingness on the human collaborator and actual team effectiveness remains underexplored.

Furthermore, effective communication is a pillar in teamwork and decision-making [286, 329, 397, 399]. Particularly, [331] calls for ensuring meaningful human control of machine's decisions in morally-sensitive situations. This suggests task allocation in environments that may involve human lives (such as search and rescue scenarios), for example, ethically require human-in-the-loop. Moreover, [370, 371] show that allowing people to influence or participate in task allocation increases perceived autonomy, satisfaction and task identity. Furthermore, increasing transparency and communication during task allocation has been shown to improve efficiency in collaborative work [320]. Similarly, Azhar et al. (2017) [26] highlight that shared decision-making in human-robot collaboration improves team performance. Failing to adapt the robot's behavior to the human teammate can deteriorate trust and satisfaction in human-robot collaborations, even when the robot's goal is to increase team performance [285]. In fact, [83] shows that decision-making that takes into account the human's trust dynamics can improve long-term performance. These findings suggest that humans' feedback should be included in task allocation methods, both for ethical considerations and overall human trust and satisfaction.

5.3 Method

A 2x2 mixed-design user study was conducted. The study was approved by the research and ethics committee of TU Delft prior to the start of the experiment (ID No 5121), and it was preregistered in the Open Science Framework (OSF) [60].

5.3.1 Participants

Seventy-four participants were recruited through the professional network of the researchers. Given technical issues, the final participant count is N=72. Participants were mainly people with technical backgrounds, with ages between 18 and 65, of which 46 were in the age-group 26-35. Forty identified as male, thirty as female, one as non-binary and one preferred not to say. Their cultural background was mainly European (44), and Asian (20). Regarding their video game experience, 21 reported high, 32 average and 19 low.

5.3.2 Materials

Task

To answer our research question, i.e., to explore the effect of willingness-based task allocation on participant's trust and teamwork, we needed a task with the following requirements:

1. The task represents a shared goal, where one human (participant) and one artificial agent need to collaborate to achieve success.

2. The task is composed of sub-tasks with variable characteristics that can elicit different willingness from the human.

3. The level of competence required for each sub-task is similar.

4. The level of competence of each teammate for each sub-task is similar.

5. There are (at least) two possibilities of task allocation of the sub-tasks between the artificial agent and the human.

6. There is (at least) one plan for task allocation that takes into account the participant's willingness, and one that does not.

7. The different task allocations should not have a significant difference in terms of efficiency, e.g., one should not be much faster than the other, for either teammate and in general.

Environment: With the Matrx [187] package, we built a 2D grid-world that simulates a search and rescue scenario (see Fig.5.1) in python [305]. The environment presents four areas A, B, C, and D, with two sub-areas, 1 and 2, each. Areas A and D are on dry ground whereas C and D are in water. Participants' movement would be laggy in water, and it would also produce a safety beeping sound.

Figure 5.1: Environment developed with Matrx [187] to simulate a Search and Rescue scenario for human-robot collaboration.

Mission: There are two victims in each area (randomly assigned per mission) that need to be brought to the safe zone in the center, by either the human participant (wearing an orange hat in Figure 5.1) or the virtual robot. Participants have full visibility of the whole grid, and they can move freely, pick up and drop off any victim, using the keyboard. Victims need to be brought to the safe zone in a specific order: looking at the safe zone, victims should be brought in line by line, from top left to bottom right, as if one is reading in English. There is a time limit of five minutes per mission to bring all victims to the safe zone, but this is plenty to complete the mission.

Willingness manipulation: The task is designed to elicit a lower willingness for the sub-tasks in water. Rooted on the idea that more effort and discomfort decrease a person's motivation towards a task [181, 212, 273], we made the team members' avatars lag when moving in the water, imitating what happens when we walk in water. In addition, when the human participant is in the water, there is also a beeping sound, simulating a safety protocol. This repetitive sound is intended to provoke mild discomfort, also decreasing the willingness towards going to water. In case the sensory manipulations would not be enough to elicit willingness variation in a virtual simulation, we explicitly told participants that one of their personal goals was to keep their time in water as low as possible, increasing the chances of having willingness variation across tasks, and consequently different task allocation plans. Participants were told: *"Now, imagine that going into the water is something you don't like doing. You can enter the water, but your avatar will slow down, which is uncomfortable. Also, you wear a water sensor that beeps for safety purposes. There will be a timer tracking how long you spend in the water. One of your personal goals is to keep this time as low as possible."* Although this did not motivate participants to decide on their willingness freely, they were still presented a trial and asked for their willingness, after interacting with the environment. Our goal was to manipulate people's willingness, making it decrease towards tasks in water (i.e., in areas B and C), so that we would ensure different task allocation plans. We also expected some people to be less willing to do tasks further away (i.e., in areas 2), as it takes more time, and hence more effort.

Score: There is a simple scoring system in place. For each victim dropped in the right place and order, the participant gets 5 points, while they would receive only 2 points for victims that were dropped in the right place but out of order (e.g., the participant did not realize that one of the previous victims was missing).

Trial: There was a trial version of the mission, with only one victim per area (in all areas), and no virtual robot present. During this trial mission, the participant could get familiarized with the environment, reducing the learning curve between the within conditions.

Self-reported willingness: After playing a trial version of the mission, participants were asked to answer how willing (6-point Likert scale willing/unwilling) they were in terms of the soil and distance characteristics. In particular, they were asked how willing they were to rescue victims from areas that were *close*, *far*, *in dry soil* and *in water*. Participants were expected to report a lower willingness to perform tasks in water.

Virtual robots: Participants were asked to collaborate with two different virtual robots (one per mission), namely Argo and Bolt. The robots were represented by slightly different avatars. Each robot, before the start of the mission, suggested their plan for task allocation, i.e., which areas would be the responsibility of the human participant to rescue victims from, and which areas would be the responsibility of the robot. The virtual robots had identical in-game behavior. Once the mission starts, the robot checks which victim needs to be rescued next (based on the order stated on the safe zone) and that is, in one of its assigned areas. Then, it moves towards that victim, picks it up, and brings it to the respective icon on the safe zone. The robot follows the order shown in the safe zone and only places the victim on its icon if that is the next victim to be placed. Before placing a victim, the robot checks whether all previous victims in the sequence have been delivered. If there are missing victims and they are in one of the areas assigned to the human, the robot waits until that victim is delivered before placing the victim it is carrying. After delivery, the robot proceeds to the next victim in the safe zone order that is in one of the robot's assigned areas. The only difference between the virtual robots was the task allocation strategy, and the participants were aware of that. They were told *"Argo and Bolt may have different task allocation strategies"*.

5

CONDITIONS

The two different virtual robots, with distinct task allocation strategies presented two within-subject conditions, one with willingness-based task allocation (condition *Will*) and a baseline (condition *noWill*). We also had two between-subject conditions, which differed in terms of communication, i.e., one allowed a closed-loop (condition *Comm*) and the other was one-directional (condition *noComm*). In this section, we clarify the differences between the existing conditions.

Task allocation: We developed two different task allocation strategies for the mission: baseline and willingness-based. The baseline allocation was static and the same for all participants. It consisted of dividing the tasks equally in terms of effort, i.e., areas A1, A2, B1, and B2 would be assigned to the human and C1, C2, D1 and D2 to the virtual robot. The willingness-based task allocation, however, was dynamic and maximized the willingness of the human. If there was any variation on self-reported willingness, we cumulated the willingness for each pair of characteristics (i.e., dry ground and close by, water and close by, dry ground and far, water and far), and assigned the four areas corresponding to the top two pairs. For example, if someone gives *Very willing (6)*to all characteristics except water, to which they give *Very not willing (1)*, they'll be given areas A1, A2, D1 and D2. In case the human did not show any variation in self-reported willingness across tasks, the willingness-based task allocation presents the same plan as the baseline. The task allocation strategy and the name of the first and second robot was counterbalanced across participants to control for order and identity effects.

Communication: Teamwork ideally includes a closed-loop communication [329], meaning that the task allocation should incorporate human feedback [370, 371]. However, if we allow the human to change the task allocation plan, we cannot see the effect of willingness-based task allocation plans, as users may change them. For that reason, we had two between

conditions: one in which the human was presented the plan for task allocation and could not change it (noComm), and another condition in which the human could freely alter the task allocation plan (Comm) presented by the virtual robot, before the start of the mission, as long as each was assigned the same number of tasks. We expected participants with stronger preferences for specific type of tasks (i.e., different values of willingness across tasks) to make more changes in the baseline task allocation.

Measures

Co-variates: There are variables external to those manipulated that may affect our results. For example, how one (trustor) trusts another can be influenced by the trustor's propensity to trust[250], which is defined before the interaction. Similarly, how one performs in the simulation may be affected by their experience with video games. To avoid biases in our results, participants were asked to fill in their *propensity to trust technology* [119], their video game experience (low, average, high), and other demographics such as age, gender, and cultural background, before the experiment started.

Self-reported measures: participants were asked to report *team fluency* [149], *perceived trustworthiness* [118], and *trust* [119]. All these measures were composed of 7-point Likert scales (agree/disagree). We have removed the original item 7 from team fluency (*The human worker was necessary to the successful completion of the tasks.*) since, in this context, it would mean the same as the original item 8 (*I was necessary to the successful completion of the tasks*).

Objective measures: For each mission, we logged several objective measures, including the *total time* and the *number of victims rescued* per team member, along with the areas assigned to each member. This allowed us to calculate *compliance*, i.e., the number of victims rescued by the human in human-assigned areas divided by the total number of victims rescued by the human (in both human-assigned and robot-assigned areas). This value ranges from 0 to 1; values below 1 indicate that some of the human's rescues occurred in robot-assigned areas.

Other questions (all with possible open-answer justification of answer): To better interpret the scale-based responses, we included questions assessing participants' robot preferences and their awareness of the robot's allocation strategies. We also examined their views on the importance of incorporating willingness into task allocation, and whether these views differ between short, infrequent collaborations and longer-term ones. We hypothesize that longer collaborations heighten the need to consider human willingness, as repeatedly assigning disliked tasks without positive counterbalances can be detrimental [369]. Finally, we sought to understand how to weigh different objectives when optimizing task allocation, asking participants to rank different objectives, and to identify which contextual factors might influence their answer.

O1: *Which robot would you be more likely to collaborate with in the future?* (answer options: Argo, Bolt, both, none, I don't know)

O2: *Do you think that any of the robots took into account your willingness when suggesting the task allocation?* (answer options: Argo, Bolt, both, none, I don't know)

O3: Participants were asked to imagine the following scenario *Imagine that you are going to collaborate with one robot on a search and rescue mission. You know that you will only collaborate once, and that mission will take at most one day. That robot is in charge of allocating tasks, i.e., deciding which tasks you have to complete and which tasks it (the robot) has to complete.* followed by the question *How important would it be for that robot to know and to consider your willingness when allocating the tasks for this one mission?*(possible to answer in a 7-point Likert scale important/not important).

O4: Participants were also asked to imagine the following scenario *Imagine that you are going to collaborate with one robot on several search and rescue missions. You know that you will collaborate on several missions, and that there might be up to five missions per week for a whole year. That robot is in charge of allocating the tasks, i.e., deciding which tasks you have to complete and which tasks it (the robot) has to complete.* followed by the question *How important would it be for that robot to know and to consider your willingness when allocating the tasks for these missions?*(possible to answer in a 7-point Likert scale important/not important).

O5: Finally, participants were asked to rank factors that are important when doing task allocation. In particular, they were asked *Sometimes, to prioritize your willingness in task allocation, the robot may need to disregard other criteria. Please indicate (rank) how you think the robot should prioritize the following factors when planning the task allocation, i.e., deciding which tasks you have to complete and which tasks it (the robot) has to complete.* The factors were: *Your willingness (e.g., you have a preference for certain tasks)*, *Efficiency (e.g., completing task in the lowest amount of time possible)*, *Effectiveness (e.g., all goals achieved)*, and *The robot's willingness (e.g., the robot has a preference for certain tasks).* They were also asked *Would you change the ranking order depending on contextual factors? If so, which ones and why? (Optional).*

5.3.3 PROCEDURE

The study took place in person, with a researcher and one participant in the room. Each participant sat in front of the researcher, facing a monitor, a mouse and a keyboard, all connected to the researcher's laptop. The researcher could not see the monitor screen during the experiment. When the participant arrived, there were two tabs open on the monitor screen: one with the simulated game (with the tutorial version) and another one with a survey. After a trial mission to familiarize participants with the task and interface, each participant completed two collaborative missions with each virtual robot. After each mission, participants completed the self-reported scales of trustworthiness, trust and team fluency. The experiment concluded with the set of tailored questions about the participant's experience.

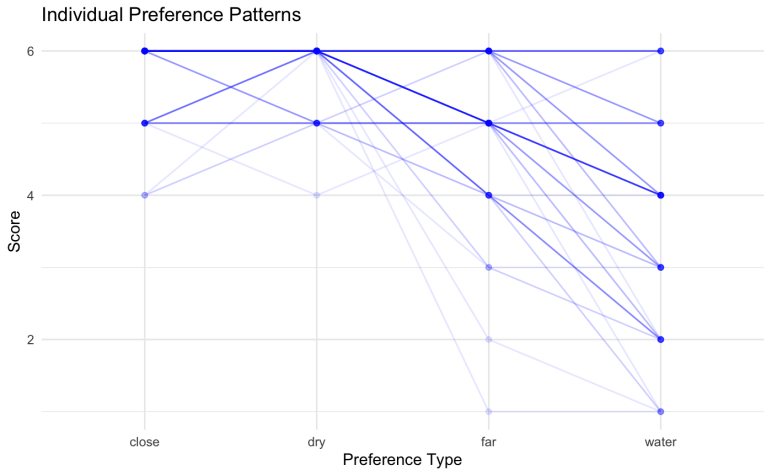


Figure 5.2: Patterns on willingness values indicated by the participants after the trial mission, across tasks in *water* (M=3.75, SD=1.48), in *dry* soil (M=5.78, SD=0.45), in *close* distance (M=5.69, SD=0.55), and *far* away (M=4.99, SD=1.01).

5.4 Results

5.4.1 Willingness across tasks

We started by checking whether participants reported different levels of willingness per task, when asked about it after the trial. Figure 5.2 shows the patterns of willingness values for the four possible characteristics of tasks, i.e., close, far, dry and water. Since the shapiro test [347] showed that the distribution of the willingness values were not normal, we ran a pairwise Wilcoxon rank-sum test [413] to verify whether the differences in average were statistically significant. Indeed, *water*'s willingness score (M=3.75, SD=1.48) was significantly lower than *dry* (M=5.78, SD=0.45), with $U = 4607, p < 0.01$, significantly lower than *close* (M=5.69, SD=0.55) with $U = 4531, p < 0.01$, and significantly lower than *far* (M=4.99, SD=1.01) with $U = 3873, p < 0.01$. Also, *far* was significantly lower than both *close* ($U = 3733, p < 0.01$) and *dry* ($U = 3886, p < 0.01$).

5.4.2 Allocation plan

Changes to the plan Next, a Brunner-Munzel Test [47, 140] test showed that, before the start of the mission, participants made significantly fewer changes to the task allocation plan presented by the *Will* (M=0.37, SD=0.77) agent that to the *noWill* agent's (M=1.20, SD=0.96), with BM(65)=4.01, $p < 0.01$. During the mission, compliance values were high throughout and there were no significant differences among conditions.

Effect on efficiency All participants successfully finished all the missions, collaboratively bringing all the victims to the safe zone within the time limit. *Game time* was significantly shorter in the condition that allowed the participants to change the plan (Comm), with main effect of group Q(1, 40.04) = 8.03, p < 0.01, with an estimated trimmed

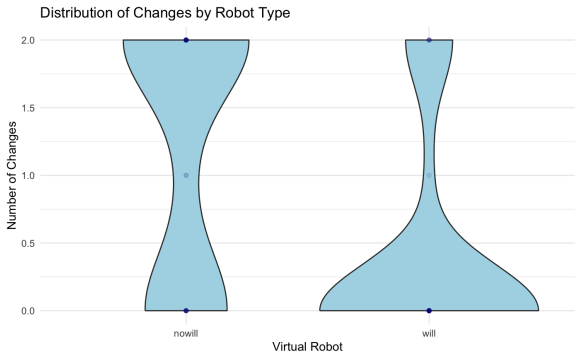


Figure 5.3: Violin plot showing the distribution of changes made by the participant across robots. The plot illustrates the density, median, and variability within each group.

5

difference of 18.8 seconds. The robust between-subjects effect test (`sppba`) also supported this difference (p = 0.014).

5.4.3 Self-reported trust and preferences

Robot preference and perception To analyze the difference between the two types of task allocation (Will or noWill), we investigated the self-reported measures of those participants who showed a variation in willingness across tasks (N=64, with 35 participants in the comm condition). Table 5.2 shows that when asked which virtual robot they would more likely collaborate with in the future (question O1), 56.3% of the participants said the name of the one that implemented a will-based task allocation, while 25% said the name of the baseline one, 14% said both, and 4.7% said they don't know. Similarly, when asked if any of the virtual robots took their will into account (question O2), 62.5% of the participants said the name of the one that implemented a will-based task allocation, while only 6.25% said the name of the baseline one, 7.8% said both, 20.3% said they don't know, and 3.1% said none of the virtual robots took their willingness into account. We can also see that 42.19% of the participants that recognized correctly which robot included their willingness in the task allocation, also preferred the robot. However, 14% of the participants still preferred the robot that did not base the task allocation on their willingness, although they recognized that.

Trust, trustworthiness and team fluency We ran robust mixed-ANOVA (with the R package WRS2[241]) to investigate both main and interaction effects on the self-reported metrics (see Table 5.1). The averages of trust, perceived trustworthiness (including its sub-metrics of ability, benevolence, and integrity), and team fluency were high across all conditions and showed no statistically significant ($p < 0.05$) differences. This means that none of our initial hypothesis was proven.

Table 5.1: Means (M) and Standard Deviations (SD) for trust, trustworthiness (and sub-metrics), and team fluency.

	Comm		**noComm**	
	Will	noWill	Will	noWill
Trust	M = 5.04, SD = 1.43	M = 4.82, SD = 1.26	M = 5.03, SD = 1.50	M = 5.06, SD = 1.30
Trustworthiness	M = 5.31, SD = 1.01	M = 5.05, SD = 0.78	M = 5.21, SD = 0.79	M = 5.16, SD = 0.78
Ability	M = 5.75, SD = 1.27	M = 5.33, SD = 1.13	M = 5.73, SD = 0.82	M = 5.71, SD = 0.92
Benevolence	M = 4.53, SD = 1.46	M = 4.24, SD = 1.38	M = 4.30, SD = 1.75	M = 4.09, SD = 1.56
Integrity	M = 5.66, SD = 1.13	M = 5.57, SD = 0.87	M = 5.60, SD = 0.83	M = 5.69, SD = 0.77
Team fluency	M = 5.18, SD = 0.76	M = 4.90, SD = 0.96	M = 4.74, SD = 0.92	M = 4.83, SD = 0.97

5.4.4 Contextual importance of willingness-based task allocation

Relative importance of willingness When asked to rank their willingness importance relatively to other objectives (question O5), most participants ranked *effectiveness* as the first (N=48) objective to optimize for when planning task allocation. After that, most ranked *efficiency* as the second (N=37) objective, and human's willingness as the *third* (N=44). Finally, the robot's willingness was ranked by most (N=67) as the last objective to lead the optimization. Figure 5.4 shows the heatmap and counts of each objective and each position in the rank.

Importance of willingness in short-term vs long-term collaboration Figure 5.5 shows the difference betwee O3 and O4, i.e., the reported importance of willingness in short-term collaboration (M=3.4, SD=1.12) vs long-term (M=4.1, SD=1.04). A Wilcoxon rank-sum test showed these answers were significantly different ($U = 1643.5$, $p < 0.01$).

Contextual factors that shape willingness importance All the data (N=74) justifying the relative importance of willingness (O5) was used for a thematic analysis [44], since the technical problems are unlikely to affect what a person values in task allocation. Our dataset [63] can be found in 4TU.ResearchData. This analysis followed the question *"Which contextual factors may influence the importance of willingness as an objective in human-robot task allocation?"* The first author and a double-coder (non-expert) went through the answers and wrote down some codes that came to mind related to the question posed. Both coders met to discuss the codes and reach an agreement on the coding scheme. After agreeing on the coding scheme, both coders coded the utterances separately. Both coders met one final time to agree on the coding. We calculated the inter-rater reliability for the

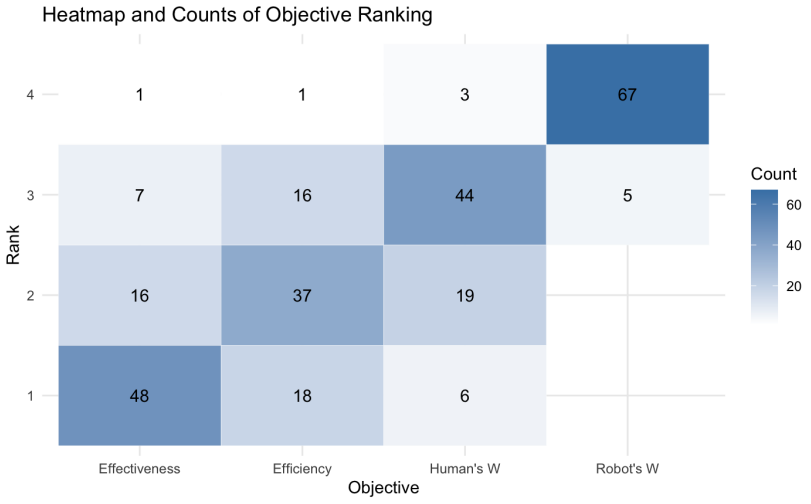


Figure 5.4: This heatmap shows how participants prioritized the different objectives for task allocation. It shows how many participants placed each objective in each rank. Rank 1 is highest priority, rank 4 lowest.

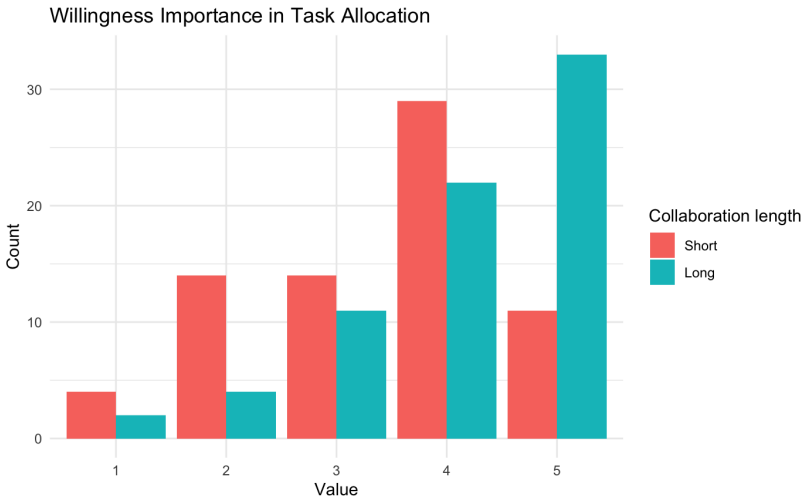


Figure 5.5: Counts on answers to O3 and O4. This plot shows how participants believe it is important to consider their willingness when allocating tasks for short (O3) and long (O4) collaborations.

Table 5.2: Percentage of response combinations between preferred robot (rows) and perceived willingness-based task allocation in robot (columns). For example, while 62.50% correctly detected which robot incorporated their willingness into task allocation (Will), only 42.19% of the participants preferred that robot knowing that said robot incorporated their willingness. Overall, 56.25% of the participants preferred the Will robot, whether they realized it was taking into account their willingness or not.

	"Which robot took into account your willingness?"					
Preferred robot	will	no-will	both	none	don't know	**Total**
will	**42.19**	1.56	3.12	1.56	7.81	**56.25**
no-will	14.06	1.56	4.69	0.00	4.69	**25.00**
both	6.25	3.12	0.00	1.56	3.12	14.06
none	0.00	0.00	0.00	0.00	0.00	0.00
don't know	0.00	0.00	0.00	0.00	4.69	4.69
Total	**62.50**	**6.24**	7.81	3.12	20.31	100.00

thematic analysis, resulting in a Cohen's kappa [221] of *0.93* (ran with R package *irr*[139]), which is considered *almost perfect*[221].

Table 5.3: This table presents the results of a thematic analysis on question O5. Participants were asked to point out which (if any) contextual factors might change their ranking of important factors. This analysis was made with the question in mind *Which contextual factors might change the importance of willingness in task allocation?*

Theme 1	Theme 2	Code	ID	Count
Team	Human	Human's strong willingness (+)	H1+	7
		Impact on human's integrity (+)	H2+	4
	Robot	Robot's strong willingness (-)	R1-	5
		Impact on robot's integrity (-)	R2-	2
	Collaboration	Teamwork duration	C1+	3
Task	NA	Importance (+)	T1+	1
		Importance (-)	T1-	21
		Importance (N/D)	T1X	1
		Urgency (-)	T2-	8
		Urgency (N/D)	T2X	1
		No	N	6

The final themes, codes and respective counts can be found in Table 5.3. Participants mentioned contextual factors either related to the human team member, the robot team member, their collaboration, or the task itself. These factors were also coded in terms of how they may affect willingness's importance, whether positively (+), negatively (-), or undefined by the participant (N/D). The answer to this question was not mandatory and no-answers were counted as *No*, as well as those that explicitly said "No". Besides, five participants participants gave answers that were unclear (U), and six participants mentioned contextual factors that explicitly did not relate to willingness (O), but instead related to some of the other objectives in the ranking (e.g., efficiency).

The description of the codes is as follows:

- *Human's strong willingness (+)* Participant says that if their/the human shows a strong(er) willingness or preferences (reasonably justified or not), they would rank human willingness higher (or vice versa: poor or no justification leads to lower ranking).
- *Impact on human's integrity (-)* Participant says that if their/the human's physical or mental integrity is at stake, they would rank human willingness higher (or vice versa: if health is not at stake, willingness ranks lower).
- *Robot's strong willingness (-)* Participant says that if the robot shows a strong(er) willingness or preferences (reasonably justified or not), human willingness should rank lower (or vice versa: unjustified robot preferences lead to higher human willingness).
- *Impact on robot's integrity (-)* Participant says that if their/the robot's physical integrity is at stake, human willingness should rank lower (or vice versa: unjustified robot preferences lead to higher human willingness).
- *Teamwork duration* Participant says that if collaboration is expected to last longer or happen more frequently, human willingness should be ranked higher.
- *Importance (+)* Participant says that if the situation is critical or severe (e.g. real lives or other serious consequences are at stake), human willingness should be ranked higher (or vice versa: less critical situations decrease willingness ranking).
- *Importance (-)* Participant says that if the situation is critical or severe (e.g. real lives or other serious consequences are at stake), human willingness should be ranked lower (or vice versa: less critical situations increase willingness ranking).
- *Importance (N/D)* Participant says that if the situation is critical or severe (e.g. real lives or other serious consequences are at stake), the ranking may change (potentially affecting willingness), but they don't specify how.
- *Urgency (-)* Participant says that if there's limited time (specifically time) to complete the goal, human willingness should be ranked lower (or vice versa: more time increases willingness ranking).
- *Urgency (N/D)* Participant says that the time (they specifically say time) to complete the goal can affect the ranking (potentially impacting willingness), but they don't specify how.

5.5 Discussion

In this section we first discuss our results, then their theoretical implications, and finally the limitations of our work and options for future work.

5.5.1 Results

Willingness across tasks The first thing to notice is that our manipulation worked and participants showed less willing to go into water (see Section 5.4.1), but also to go far. This aligns with the works that argue that people make decisions based on a cost–benefit analysis, leading to the selection of tasks with lower perceived effort or discomfort (see e.g., [181, 212, 273]).

Changes to the plan Participants made less changes to the allocation plan of the robot that considered their willingness (Will), suggesting participants preferred plans that took into account their willingness. This may also suggest that preference integration can reduce the need for human correction in task planning. Further, overall task completion time was shorter than in the noComm condition. This suggests that when given the opportunity to change the plan, participants may perform better. This finding is consistent with Tausch et al. (2022) [371], which shows that when users have more control over task allocation, their perceived autonomy also increases, as well as earlier shared-control literature by Azhar et al. (2017) [26], that highlights that collaborative decision-making in human-robot collaboration improves team performance.

Trust and preferences Participants preferred the virtual robot that included their willingness in the task allocation plan (see Section 5.4.3), and several correctly noticed a willingness-based allocation in their preferred robot. This is in line with prior research showing that incorporating human preferences into planning can improve perceived likability and intelligence [245]. Although willingness seems to be important for allocation, it did not impact perceptions of trust, trustworthiness, or team fluency. Also the communication condition did not significantly affect any of the three metrics, contrary to what is showed by Tausch's works [370, 371]. All trust, trustworthiness and team fluency remained high across all conditions, probably due to the context presented in the experiment, as explored in the next section.

Contextual importance of willingness-based task allocation As shown in Section 5.4.4, most participants prioritized effectiveness and efficiency over willingness. In the thematic analysis, we see that a third of the participants explicitly stated that a different task could increase the importance of willingness More concretely, tasks that are less important (twenty-one participants), i.e., lower stakes or criticality, or less urgent (eight participants) increase the human's willingness importance in task allocation. Our search and rescue (SAR) scenario could convey both the feeling of importance and urgency to save lives, which potentiated effectiveness and efficiency to be prioritized by participants. Research indeed shows that people prioritize performance in high-stakes environments (such as SAR) [339] over other other social attributes, such as etiquette [159], or even some discomfort [369]. This aligns with the findings by Correia et al. (2019) [93] where participants chose, between two robots, the one that was on the team that won a competitive team game. Furthermore, Inzlicht et al. (2018) [184] shows how actions that demand more effort (such as going into water) can also be perceived as more rewarding. In our SAR experiment, the human-robot teams performed well across all task plans, which may explain the high trust and team fluency across conditions. Nevertheless, participants appeared to find willingness-based

task allocation important in other situations, such as in extended collaborations, or in situations where they feel stronger about their own willingness (see Section 5.4.4).

5.5.2 Theoretical Implications

Willingness-based task allocation Our findings have several theoretical implications for research on human–machine teamwork. The main implication is that task allocation models (such as [51, 194, 243, 410]), should incorporate human willingness alongside more traditional variables such as skill and availability, while taking task criticality, task urgency, and the length of the collaboration as boundary conditions that determine how much human's willingness should be accommodated. Less critical and urgent tasks (e.g., cooking with flexible time), or long collaborations, require human willingness and preferences to be taken into account for task allocation, as in Zhao et al. (2023) [427] and Dhanaraj et al. (2024) [107]. This challenges optimization frameworks (such as [16, 25]) that focus purely on maximizing objective performance and extends them to account for the human partner's subjective state. Our results also suggest that in high-stakes tasks, trust in machine partners and perceived team fluency may be mainly anchored in observed effectiveness and efficiency, and so, in such situations, performance should be the priority of the system.

Implications on ITA This work was partially inspired by the Interdependence and Trust Analysis (ITA) framework [69]. The framework proposes a method for designing human–machine team configurations across tasks, extending the original Interdependence Analysis [194]. It does so by also considering potential performers' willingness and contextual factors, in addition to their competence (referred to as capacities in the original paper [194]). Beyond the uncertainty of the effects of including willingness in task allocation, we also lacked clarity on how much weight willingness should carry compared to competence. Our results suggest that willingness is less relevant in critical situations, and that only very low levels of willingness should be factored into such contexts. This indicates that the original table [194] may be sufficient in high-stakes environments. Similarly, when estimating overall performance, a team configuration where the human shows low willingness but also leads to lower overall performance may reasonably be disregarded. This suggests that expectation maximization based on performance, as in Azevedo-Sá et al. (2021) and Ali et al. (2022) [16, 25], is also likely to be sufficient in high-stakes environments. However, we suspect that willingness becomes more important (potentially even more than overall performance) in non-critical tasks or in settings with sustained interaction, where ignoring preferences could result in fatigue or burnout [263, 369]. This aligns with literature that show that a robot that makes decisions based on human's trust and preferred strategy, can improve performance long-term, even if the immediate goal is not maximizing efficiency [83, 285]. Most importantly, any decision must be clearly communicated and remain open to human intervention.

Communicating and deciding task allocation Regarding communication, the fact that participants performed better when given the opportunity to modify the plan reinforces the view that human agency is an important driver [26, 371] of effective collaboration. Task allocation models should therefore place greater emphasis on opportunities for meaningful human intervention within shared control systems, adapting the principles of Meaningful

Human Control [331]. Finally, the fact that participants prioritized different optimization objectives (i.e., efficiency, effectiveness, willingness) and that these depend on contextual factors (e.g., task importance), suggests that models of task-allocation should account for human performance heuristics and perceptions, not only actual system outcomes. As such, systems should remain transparent [320], and could present contextual and contrastive explanations (e.g., as in Verhagen et al. (2025) [399]) as to why one plan may be better than another (e.g., in terms of performance or predicted satisfaction), in a given situation.

5.5.3 Limitations and Future Work

Our experiment focused on a simulated high-stakes task with short duration and low frequency of collaboration. The results indicate a need for further research on willingness-based task allocation in scenarios that are less urgent and critical (e.g., meal preparation with flexible timing) and in teams with longer or more frequent collaborations (e.g., cooking daily with the same robot over a year). In our study, we attempted to elicit participant preferences by adding discomfort or higher effort to the task of rescuing victims from water. However, because the task was simulated, participants might have been indifferent to these factors. Also, results may differ in environments where people have inherent differences in willingness and preference, such as some preferring to wash dishes while others prefer chopping ingredients. Future studies should incorporate greater variation in willingness levels across different tasks.

Going further, algorithms are needed to model willingness dynamically through interaction and observation, so that robots can be proactive and collaborative during the task, as suggested by Ramachandruni wt al. (2024) [310]. At the same time, it is necessary to develop effective multi-objective optimization strategies that account for competence, willingness, and context factors, such as task criticality and urgency, to allow robots and AI collaborators to do appropriate task selection and allocation. Furthermore, the fact that we experimented in a 2D grid-world allowed us to simulate a whole environment and task, where humans and robots have the same competence. Although this was highly important to answer our research question, we still need further investigation to see the effects of willingness-based task allocation when the robots have different embodiments. Finally, this experiment focused on dyadic interactions, and further studies are required to investigate the effect of willingness-based task allocation in teams bigger than two elements, using a task such as the one proposed by Jung et al. (2021) [203].

5.6 Conclusion

This paper reported the results of a 2x2 user study examining the effect of willingness-based task allocation in human–robot collaboration. While existing algorithms focus mainly on performance-related goals, such as maximizing competence and skills, we assessed the impact of incorporating human task preferences. Our study (N=72) involved a simulated search and rescue scenario in a 2D grid world. Participants preferred robots and plans that considered their preferences, but this did not significantly affect trust in the robot or perceived team fluency. The findings suggest that consistent team performance, task type, and the limited duration of the task may have constrained these effects. Our data suggests that for lower-stakes tasks and longer-lived teams, willingness may have a greater

influence on human collaborators. Finally, we observed that allowing changes in task allocation improved human performance. Further research across varied task types and team durations is needed to determine the appropriate role of human willingness in robot task selection and allocation.

ACKNOWLEDGEMENTS

This material is supported by Delft AI Initiative to CCJ and MLT, and by The Dutch Research Council (NWO) (Grant 1136993) to MLT and CMJ. It is also supported by the Air Force Office of Scientific Research (Grant 21USCOR004) and DARPA (Grant FA8650-23-C-7318) to EdV.

CODE AND DATA AVAILABILITY

The data and code related to this chapter are be published on *github* and 4TU.ResearchData [63], respectively.

6

Multidisciplinary Theory Building for Human-AI Team Trust

As AI becomes increasingly integrated into complex sociotechnical systems, understanding the dynamics of trust between human and AI team members is essential for effective collaboration. This chapter presents our interdisciplinary collaboration for developing a comprehensive theory of team trust in human-AI teams. Our team, representing diverse fields such as organizational psychology and computer science, worked together to bridge disciplinary gaps and create a multidisciplinary framework for human-AI team trust. This framework delineates trust relationships within teams, addressing challenges like differing terminologies and conceptualizations. We highlight the need for a nuanced, multidisciplinary and multilevel approach to trust theory in human-AI teamwork, paving the way for future research and practical applications in human-AI collaborative environments.

This chapter is based on the following handbook chapter:
Centeio Jorge, C., Ulfert, A. S., Georganta, E., Mehrotra, S., Tielman, M. L., & Jonker, C. M. (2021). Multidisciplinary Theory Building for Human-AI Team Trust. In The Oxford Handbook of Computational Group and Team Dynamics. Oxford University Press (In press)

6.1 INTRODUCTION

As technology becomes a teammate, and humans collaborate with Artificial Intelligence (AI) to solve tasks required for a joint goal, new dynamics appear that affect the humans involved and the systems of which they are part. In particular, the dynamics of trust within human-AI teams, meaning teams composed of both humans and AI, are distinct from trust in only-human environments, given the different nature and interaction of machines. Teamwork and trust within teams are topics originally studied by organizational psychology, as they initially concerned human-only environments. However, AI has been partially developed by emulating human phenomena and behavior [323], and this is no exception for the study of trust and teamwork. These topics have been used by computer scientists to develop artificial systems, as a way to model artificially intelligent agents' behavior and collective behaviors (see e.g., [57]). Still, the definitions and approaches used in the both disciplines differ. When these artificially intelligent agents start integrating human systems, the study of teamwork and trust needs to find a middle ground between the two disciplines. Advancing theories in human-AI teamwork calls for a multidisciplinary collaboration between at least the two fields, i.e., organizational psychology and computer science, for the study of human-AI team trust. These collaborations have a set of outputs, challenges and particularities, which we describe in this chapter.

Although it may not be obvious, AI is a multidisciplinary field in its roots. Defined as "the study of computations that make it possible to perceive, reason and act" ([419], p. 5), AI came from a fantasy of having machines doing what humans can do [315]. As such, the first decades of this emerging discipline focused on mathematically emulating human intelligence and behavior, inspired by various disciplines [323]. Even the most popular idea related to AI nowadays, neural networks, was inspired by the study of the human brain and findings in the field of neuroscience [323]. Similarly, robotics often strives to create systems that resemble humans in their morphology (e.g., having arms and legs) and in their behavior (e.g., taking over tasks from humans) [275]. Although the origins of AI are multidisciplinary and built on other sciences' knowledge (from psychology, biology, neuroscience, mathematics, logic, linguistics, etc.), as it developed, this became less and less the case. With the increasing availability of data sources to train AI (e.g., on the internet) and increasing computational power, AI has moved towards a more computer-science-oriented discipline [275]. As such, AI became a data-driven science [323], which makes today's AI research and theory often mono-disciplinary in focus.

Besides these highly technical developments within the AI research community, we have seen an increase in humans collaboratively working together with AI across various application settings, such as in manufacturing [330], healthcare [407], search and rescue [340], and others. Consequently, modern AI systems typically do not exist in isolation but rather as part of complex sociotechnical systems, which depend on human social elements, such as social norms, communication, and trust [10]. These continuous developments have also given rise to the need for hybrid intelligence [11], that is, combining machine intelligence with human intelligence instead of trying to substitute and surpass it. With intelligence and autonomy of machines and the implementation of such systems in new environments, the direct interaction between humans and machines becomes complex. In particular, AI can take over a variety of roles as highly competent tools or even as a counterpart or teammate [183]. To fully address this complexity, the AI system should be

understood within its context of use, that is, how it is directly or indirectly interacting with and affecting humans within a specific environment (e.g., at work). Yet, to study and develop these modern systems while considering the overall sociotechnical systems, multidisciplinary approaches are needed.

This chapter is about how five human-AI team trust researchers who come from different backgrounds joined efforts to build a multidisciplinary theory on human-AI team trust. We are a team composed of computer scientists, organizational psychologists, and a cognitive AI scientist. Our journeys led us to meet each other unexpectedly at a conference workshop (a day of short talks) on group and team dynamics, which welcomed multidisciplinary contributions. With the goal of establishing a common ground between our different perspectives, we wrote a paper called *Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework* [376]. This paper was the starting point for collaborative activities, funding opportunities, and workshops. Through our collaboration, we strengthened multidisciplinary research and expanded our research skills, methods, and knowledge of the literature. Most importantly, we contributed to building a larger community with similar research goals (see e.g., [74]). In this chapter, we present our journey in developing and publishing a multidisciplinary theory on team trust in human-AI teams, and its subsequent outcomes.

6.2 "Our Background": Existing perspectives on teamwork and trust

A human-AI team can be defined as one or more human individuals and one or more AI agents [384], where an AI agent is a computer entity that perceives and acts in an environment [323] and possesses "a partial or high degree of self-governance with respect to decision-making, adaptation, and communication" ([297], p. 904). With the continuous development of such technologies, AI agents take over increasingly complex tasks, increasing the interdependence between them and their human team members [343], thus requiring trust between them for effective teamwork [329]. This trust becomes crucial as humans and AI collaborate on high-stakes projects where errors could have significant consequences [301]. As such, we had to first understand the role and dynamics of trust in human-AI teamwork.

Trust describes the willingness to rely on and be vulnerable to another party – a central prerequisite for effective collaboration between users and AI agents as well as between human team members [45, 101, 332, 348]. For instance, distrust, over-trust, or under-trust in human-AI teamwork can critically hinder team effectiveness [103, 148]. Furthermore, trust among humans has also inspired computer scientists, who have been modeling trust for the decision process of AI agents [57, 58, 125, 127]. These models are used mainly to decide with which other AI agents one AI agent should interact [328], and are composed of *beliefs*. For AI agents, beliefs are formed based on the environment in which they are embedded in. According to the Belief-Desire-Intention (BDI) architecture [145], a belief is the informative component of the system state, that is, an AI agent's perception about the world at that specific point in time (including itself and other entities involved), for example, "the human team member is able to perform its task". These beliefs can determine the "degree of trust" towards another entity. These models from computer science literature traditionally tend

to focus on trust as a relation between different AI agents (e.g., in multi-agent systems).

By the time our team (i.e., the authors of this chapter) met, the computer scientists among us were exploring not only how to model a trustworthy AI agent team member [260] that fosters appropriate trust, but also how to model the AI agent's trust towards a human team member [68]. From the opposite perspective, however, the psychologists among us were exploring how a human operator (trustor) builds trust towards an AI team member (trustee), and perceives the AI agent's trustworthiness [375]. However, until this point, our approaches were missing a combination of the others' existing trust perspectives, both from psychology and computer science, acknowledging both human and AI agents as team members that may be enabled to evaluate each other's trustworthiness, making them both trustors and trustees. What's more, we were all mainly working on teams composed of a single human and a single AI. We realized that such a multidisciplinary approach would we be crucial to explore human-AI team trust dynamics in bigger teams, which involve several humans' and several AI teammate's trust and trustworthiness.

Starting working together and sharing the current state of our disciplines made it clear that human-AI teams encounter several trust-related challenges that can impact their effectiveness. We realized that trust must be cultivated at multiple levels, including between individual members (human-human, human-AI, AI-AI) and the team as a whole, requiring a nuanced understanding of these dynamics [376]. Furthermore, as AI agents assume roles previously held by humans, redefining some of the well-established team and trust dynamics became essential. For example, humans may find it difficult to view AI as similar or integral to the team, hindering trust development [143]. Establishing trust, and ensuring that team members feel secure in taking risks and managing trust variability among members, is crucial for fostering effective collaboration in human-AI teams [162].

6.3 "Method": the collaboration

After meeting (online) a few times, we started developing ideas for merging our views on trust and teamwork in human-AI teams. Initially without any funding, we started working towards our first major goal, which was writing our first paper together. The goal of this paper was to find a theoretical common ground that would combine our different perspectives and approaches on the topic. In order to do that, we realized that the main challenge would be to find a way to communicate our ideas to each other. Departing from a multidisciplinary literature review on the term "team trust" in human-AI contexts, we quickly realized that while trust literature across disciplines was abundant, there was a limited focus on human-AI teams. Due to this lack of literature at the time, we started working on a set of propositions that integrated both our computer science and psychology perspectives. In the process of writing these propositions, we learned about our disciplinary differences in each other's definitions, ways of thinking, and ways of working (e.g., how we write definitions or how we represent ideas in figures).

Particularly, there was terminology that we had believed was consistent across disciplines but turned out to mean different things. For example, the vocabulary used by the organizational psychologists of our team in their paper, [375] (presented at the workshop where we first met), triggered misunderstandings and discussions. Specifically, not everyone understood *antecedents* in trust theory, which is a term commonly used in organizational psychology. From a functional mathematical perspective (as in $f(x_1, x_2, ..., x_n) = y$),

trust is a relation or expression involving one or more variables $x_1, x_2, ..., x_n$ (in this case, factors affecting trust), such as ability or propensity to trust. For this reason, it did not make sense to some of us that some factors such as ability, benevolence, and integrity were seen as *antecedents*, i.e., colloquially happening before the *output value* of trust. This confusion was amplified by the fact that other variables, such as propensity to trust, were different *influencing factors* (i.e., so-called moderators instead of antecedents), which were proposed to shape the relationship between antecedents and the output of trust (e.g. changing the strength or the type of their relationship). In fact, proposing that a variable affects a relationship (i.e., moderation), instead of directly affecting another variable, is not a common practice in mathematical and formal models, often used in computer science.

Similarly, the papers written by the computer scientists in our team (see e.g., [71]) also caused new discussions. For example, not everyone initially understood the meaning of *beliefs* as a multi-agent concept, which is a commonly used term in computer science. In psychology, particularly in learning science, *beliefs* can be seen as "psychologically held understanding, premises, or propositions about the world that are felt to be true" (Richardson et al. (1996) [317], p. 103). These beliefs are not necessarily based on evidence, nor require a truth condition, and they may be interpretations and speculations of individuals [317, 346]. At the same time, in computer science, a belief (as in the BDI architecture of [145]) is a representation of something in the world for an agent, which can include inference rules, possibly leading to other beliefs. However, they do not represent a feeling, faith, or judgment. Instead, they are an objective "picture" that the agent has of the system it is part of, which can be true or false, at a given point in time. Beliefs in computer science are more comparable to the concept of *knowledge* in learning sciences (see e.g., [346]). Just like knowledge in leaning science, beliefs in computer science are updated through direct perception or communication of the world and they depend upon a truth condition.

This way, we learned several terms and concepts that we did not even consider as a knowledge gap before starting the multidisciplinary collaboration. Besides definitions, what we consider as a contribution and the type of papers we write in our disciplines can differ greatly. For example, in organizational psychology, it is common practice to write a journal paper that introduces different theories and concludes with narrative propositions that present the outcome of the paper, on which future research can build on. However, in computer science, it is more common to use mathematical formalizations (or algorithms) to introduce the proposed relationships as *methodology*, which then have to be tested with different inputs, producing the *results* of the paper. These mathematical formalizations (or algorithms) are often motivated by existing theories or formalizations. However, they are rarely considered sufficient contribution without being tested, either through simulations (see e.g., [25]), datasets (see e.g., [393]), or user studies (see e.g., [260]).

These differences created many discussions and explanations of the propositions' meaning, relevance, and contribution while drafting the paper. To tackle these misunderstandings, we often divided the parts to be written per field and collectively discussed them. For example, the ones more comfortable with computer science literature and publications wrote the computer-science-related theoretical and methodological parts. On the other hand, the psychology-based parts were written by researchers with psychology backgrounds. Then, we would read each other's parts and meet to discuss any part we did not all fully understand. The goal was for everyone to build an understanding of every

part of the paper and collectively decide how we can bring those together. Also, since we picked a work and organizational psychology journal, the researchers with the corresponding backgrounds guided the structure and methodology of the manuscript. That is how we finally managed to submit a manuscript that included multidisciplinary propositions reflecting theoretical assumptions and mathematical formalizations for different disciplines to use.

By the time we submitted the paper, we could communicate reasonably well among ourselves, understanding the different definitions, concepts, and ways of describing them. At the time, however, we did not consider how to communicate these multidisciplinary discussions to a specific audience, in this case, the readership of a work and organizational psychology journal. Consequently, we faced some unexpected initial reviews, as, to our surprise, the reviewers did not understand our core ideas. For example, we used the term (artificial) *"agent"* lightly, as it is used largely in computer science and interaction sciences. However, one reviewer asked us whether being an agent implied agency, which then implied free will. Another reviewer suggested that the mathematical formulations were not helpful, disregarding their usefulness for multidisciplinary understanding and collaboration. There were several questions/implications we had to rebut, given our very different standpoints in terms of discipline, that we did not expect. We understood that, depending on the readership, we needed to express our ideas in different ways . Particularly, we had to explain, define, and support the smallest of the concepts that we initially had assumed to be common knowledge. We also had to explicitly motivate the choice to have different "languages" on the paper, i.e., so that the paper can be read and used by both organizational psychologists and computer scientists. The paper went through three revisions, two of them major, before it was finally accepted for publication.

6.4 "Results": Human-AI team trust theory building

6.4.1 Theory building

To comprehensively describe trust in human-AI teams and to overcome the limitations of prior theories (e.g., the focus on bilateral relations), we aimed to move towards a multidisciplinary and multilevel conceptualization of team trust in human-AI teams, considering the differences and commonalities in building trust towards human and AI team members. To do so, we developed a multidisciplinary framework of team trust in human-AI teams by integrating literature on psychological trust, teamwork, HCI, and computer science. This framework is detail in *Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework* [376].

An important step in creating the theoretical framework was to define the entities that are important objects of study for human-AI team trust theory building. We defined that, in a human-AI team, trust develops towards three entities: individual team members (either human or artificial), dyads, and the team (in this case, bigger than a dyad). While some of us had not considered that artificial teammates could be trustors, some others had not considered that non-individual entities could be trustors too, i.e., that there was such a thing as the value of trust for a team or a dyad. After agreeing that each of these entities can be either a trustor or a trustee, we had to define the directional trust between each of these entities, with the additional factor that the nature of the trustee and trustor also

dictates the trust definition. At this point, we tried to make a generalization that suited both the computation of such trust (to be used in computer science) and the Psychological foundations of the trust constructs, in an attempt to reduce the level of complexity we had found ourselves in. We decided to use the computer science concept of *belief* and follow Centeio Jorge et al. (2021)'s definition that one entity's trust (in another) is their belief in the other's trustworthiness [71]. However, we had to first agree that, depending on the nature of the individual, i.e., human or artificial, the constitution of such belief and the nature of the trustworthiness differ.

Then, we had to disassemble the concept of an entity's trust when such an entity is not an individual, i.e., dyad or team. We said before that trust is directional and requires a trustor and a trustee, so who is the trustor and trustee when we say "team A has high team trust"? The organizational psychologists of the team said this is an emerging construct of the team that continuously changes based on other characteristics, such as *antecedents*, etc. If the computer scientists had to compute this, it sounded like it could be a belief based on those characteristics (e.g., different antecedents). However, in human teams, whose belief is being described? And whose trustworthiness is it? After further consideration, we agreed that, abstractly, an entity's trust in itself is its own belief in its own trustworthiness. Nevertheless, when specifying a belief, it is typically on the individual level, which means that we had to break team trust down into the aggregation of the beliefs of individual members of an entity (e.g., in a dyad, this would be the two team members that are part of it). With this in mind, we defined dyadic trust as the trust of each member (human or artificial) in each other and in the dyad itself. Finally, we could define team trust as the aggregation of all possible dyads, and individual team member's trust in other dyads (that they are not part of) and in the team itself.

6.4.2 Additional outputs

Once we published our theory in human-AI team trust, it became clear that there was a lack of common ground, transversal definitions, and, in general, people to discuss human-AI team trust. This led us to make efforts towards exploring our proposed theory further and building a community. Funding received by some of the authors of this chapter and other collaborators enabled us to finally spend a day together, meeting in person for the first time, after a year and a half of several online meetings and a paper submitted together. During that day, we shared our recent work, we explored new ideas and directions, and hang out together. Spending that amount of time together gave room for more social interactions and inspiration. We got to know each other better personally and professionally, which strengthened our collaboration and improved our communication. More than meeting in person, this funding allowed some of us to allocate full days to brainstorm together and discuss our future work (both separately and together).

In discussing our research further, we realized that central questions with regard to the meaning of trust in human-AI collaboration were still unresolved. This led to the organization of the MultiTTrust workshop [74], which counted with the participation of around twenty people. MultiTTrust was a conference workshop that offered a day full of short talks, keynotes, and discussions focused on multidisciplinary perspectives on human-AI team trust. The main takeaway of this workshop, besides the creation of a wonderful community, was that we need to concretely define the meaning of the terms

we use (especially the term trust) while, at the same time, being accepting when others use the same term with different definitions. Otherwise, we may never pass the phase of discussing terms, and that is not where most of the challenges are. The workshop took off, and by the time we were writing this chapter, the third edition had already happened, and a special issue resulted from the second edition.

After all the efforts we made to understand each other, we also find it important to make our work understandable for our original communities. For this reason, we have participated in events of the original communities regarding human-AI teams, organized either by organizational psychology or computer science committees, where we try to bridge this gap by introducing the different definitions and methods used in each other's disciplines.

We have certainly come a long way in building multidisciplinary theory for human-AI team trust. We keep establishing multidisciplinary common ground and learning which of us knows what (transactive memory as per [411]) and has which skills. This helps each of us to know who to ask for certain knowledge (from other disciplines), necessary to answer our individual research questions. Along the way, we have learned that collaboration across disciplines can be challenging but that the merging of such different perspectives can enrich the theory building, making it more transversal and useful for different researchers.

6.5 "Discussion": Lessons learned from multidisciplinary collaboration

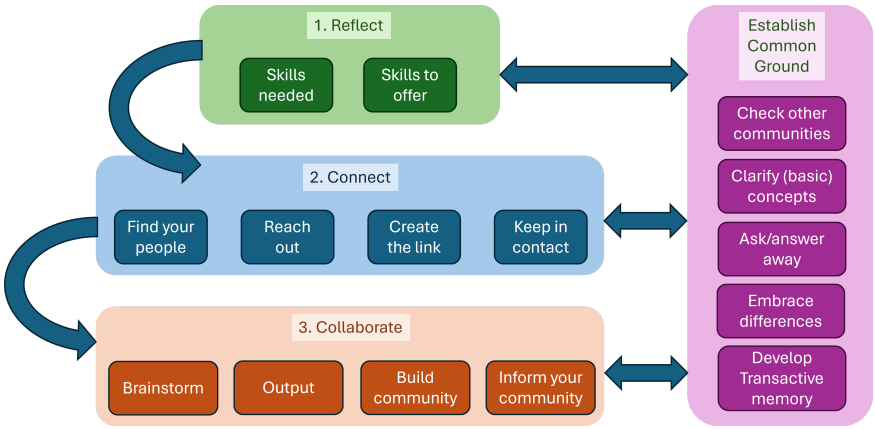


Figure 6.1: The image shows the several steps towards collaborating in new multidisciplinary fields. Steps include reflection, connection and collaboration. All these are connected to the establishing common ground block.

Working on this chapter showed us that we improved our Methodology when working together in a multidisciplinary team. We have realized that collaboration becomes easier and more effective when we follow specific steps. The first step is to establish common ground by pinpointing the important concepts we need to define, which may be different in each discipline. Once we have the terms, we dig into the literature, looking for definitions

within our communities. Then, we share the most important papers and definitions from each other's disciplines. Finally, we meet to discuss and to see if we share an understanding of the definitions of the important concepts in the different disciplines. We question each other until everyone has a clear view of the differences and similarities, and start theorizing.

To summarize the different parts of multidisciplinary collaboration, Fig. 6.1 presents an overview of the important steps and points to keep in mind, which anyone can follow to create smoother collaborations. What led us, in the first place, to understand the need to collaborate with each other was to reflect on what skills we needed that might be in other disciplines and communities and also which skills we already had that might be helpful to such communities. Then, we looked for the researchers who had the skills we needed and needed the skills we could share in events that may welcome such people. Once we discovered people that shared our interests, we collaborated to understand how our different research perspectives could complement each other and address our existing questions. When the link was created, it was important to keep in contact, which was done by, for example, updating each other every second month with what we had been researching on.

For collaborating with each other, brainstorming sessions, especially in person, were a great source of new ideas, which sooner or later became collaborative research outputs. At each meeting, we would schedule a next meeting, plan small goals for the next meeting, and discuss steps towards the next major goal of our collaboration. The small goals can be definitions to look into, literature that might be helpful for each other, etc. The major goal can be to collaboratively write a paper, submit a grant application, organize a workshop, etc. In parallel, we kept building our new community together, either by organizing workshops (for example, the above-mentioned MultiTTust workshop) or special issues that brought to us researchers with similar interests. Also, it was important to keep updating our original communities about our progress in bridging with other disciplines. We became connection points with each other's disciplines for our colleagues. This has brought us to new members of our new community.

It can be helpful to acquire funding together, to facilitate in-person meetings and run further studies, but it is not required. In fact, our manuscript was submitted before some of the members applied for a grant together. During the whole process, the most important thing has been to establish common ground. It is important to keep checking other communities' papers and events, to clarify all concepts, even when they seem "basic". Similarly, we cannot be afraid to ask questions about established constructs/methodologies in each other's fields and should motivate others to ask their questions too. Embracing our differences in skills and methods makes the team and collaboration stronger. At some point, we realized who knows what (transactive memory) and how we could bring our skills together: and we make use of it!

6.6 Conclusion

In this chapter, we describe how our collaboration emerged in a new multidisciplinary field. To address the evolving role of AI agents as team members rather than merely technological tools, we developed a framework for deeper understanding of team trust in human-AI collaboration. This framework integrates insights from organizational psychology and computer science to conceptualize team trust in these novel team configurations.

It emphasizes the multilevel nature of team trust, encompassing individual, dyadic, and team-level relationships, while recognizing AI agents as team members capable of both trustor and trustee roles. Throughout our collaborative process, we worked to overcome limitations inherent in each of our disciplines and challenges in translating concepts and methods between fields. We accomplished this by providing theoretical and mathematical definitions of trust components and offering a comprehensive analysis of team trust dynamics in human-AI teams. Our framework aims to enhance the understanding and implementation of human-AI teams in organizational contexts, particularly addressing challenges like distrust and ineffective teamwork. By being sensitive to the differences between disciplines, we seek to contribute and motivate the interdisciplinary investigation of human-AI teamwork and trust. Moreover, we hope our multidisciplinary approach serves as a model for future research initiatives that bridge two or more fields of study.

7

Conclusion

In this dissertation, we addressed the question: *How can an artificial agent model artificial trust in its human teammates for effective human–machine teamwork?*. We break this question down into five sub-questions, each examined in a separate chapter. This section summarises the conclusions of each chapter and highlights contributions and answers to the research questions. It also considers the limitations of this work and outlines possible directions for future research. Finally, it discusses the broader societal implications, including both the potential benefits and risks of using and misusing collaborative technologies.

7.1 Scientific contributions per chapter

7.1.1 Chapter 2

Chapter 2 addresses the question *How can we conceptualise artificial trust beliefs in human–agent teams?*. We answer this by formalising trust as a belief in trustworthiness that is context-sensitive and applicable to different team entities. This formalisation provides a foundation for computing beliefs about internal characteristics identified in the social sciences as key aspects of trustworthiness. We also present a taxonomy of task and team characteristics required to determine the appropriate belief of trustworthiness in a given situation, including, for example, task's nature, criticality, potential consequences, and team's lifespan, composition, and hierarchy. Furthermore, we propose that artificial trust (AT) is to affect and be affected by other team dynamics and trust relationships, implying that it cannot be modelled in isolation from the other teammates mental models. Finally, at the end of the second chapter, we discuss how, on the one hand, the development of AT-based decision-making models requires investigating which internal features (i.e., the krypta) give away a human teammate's trustworthiness, and how these can be observed (i.e., the manifesta). This step is required to build beliefs of trust and trustworthiness, as formalised in Chapter 2. On the other hand, further work is required to develop algorithms that make use of the formalised beliefs and, taking into account the context and the range of action options (e.g., selecting a task, requesting help), enable the artificial teammate to make an effective decision. Similarly, we discuss that the main challenges involved in developing trust-based decision-making are the lack of data to construct both theoretically

sound and robust artificial trust models and, closely related, the absence of systematic methods to evaluate these models.

7.1.2 Chapter 3

Diving deeper into the representation of human trustworthiness, Chapter 3's research question is *How can we assess the trustworthiness of a human teammate, given a task?*. Inspired by the works of Rino Falcone and Cristiano Castelfranchi [59, 122, 124], we investigate how to build beliefs that can translate human competence and willingness in a teamwork scenario. We explore how the components of the famous *ABI* model [250], i.e., ability, benevolence, and integrity, can be used to predict the trustworthiness of the human toward different tasks, through a user study. The task was set in a 2D grid-world supermarket scenario, inspired by the click&collect services that boomed during the 2020 pandemic, where participants had to collaborate with artificial agents to retrieve items from the supermarket aisles to complete customers' orders. Our results show that while the *ABI* dimensions describe aspects of human behaviour, they are not sufficient to predict which task a person will choose next or how well they will perform it. We find that ability can be directly associated with the belief in a person's competence. However, predicting willingness appears to depend more on an individual's subjective cost-benefit assessment of taking on a task. In other words, it may be possible to estimate a person's trustworthiness to perform a specific task if we understand how they perceive the relative costs and benefits of the available task options. For example, it may be important to understand which tasks one finds easier or more fulfilling and how this influences their intentions and motivations and, consequently, an overall value for willingness [208]. This suggests that willingness is insufficiently modelled if only linked to benevolence and integrity. Nevertheless, it is still likely that the way people evaluate these costs and benefits is influenced by their *ABI* characteristics, i.e., their ability, benevolence, and integrity.

7.1.3 Chapter 4

In Chapter 3, our goal was to explore which human characteristics and respective behavioural cues could be used to form beliefs of artificial trust in the human teammate, such as beliefs in the trustworthiness dimensions of willingness and competence [122]. In Chapter 4, we moved from exploring how to form the artificial trust beliefs to exploring how to use these beliefs for decision-making. Particularly, Chapter 4 focusses on answering the question *How can we use teammates' (human or artificial) trustworthiness for the design of human-machine teamwork?*. To answer this question, we proposed the Interdependence and Trust Analysis (ITA) framework. ITA extends Johnson's Interdependence Analysis table [194], which relied on the capabilities of the teammate, by integrating the willingness dimension of trustworthiness and contextual factors. We evaluated this with a focus group with experts both from academia and firefighters in which they applied the framework to a search and rescue scenario where there is a fire. Our results showed that using the dimensions of competence, willingness, and external factors (i.e., the dimensions proposed by Falcone et al. (2004) [122] to assess another agent's trustworthiness) is promising for transparently designing teams and collaborative technology. The ITA's external factors dimension (related to permissions and opportunities) was considered especially relevant to ensure compliance with ethical restrictions that come into place, such as the EU AI Act

[1]. Having this dimension explicitly can help ensure that the machine does not perform tasks it does not have the permission for (e.g., because there is a high risk to the humans involved) and improve the teammate's understanding of why the machine is not performing those tasks. Furthermore, our results brought up a trade-off between perceived effort and perceived utility of frameworks that intend to make technology human-centred and meaningfully controlled by the humans involved. Participant statements suggest that although our aim is to make semi-autonomous systems meaningfully controlled, that may require an amount of effort from the user, such as disclosing extra information, that they may not be interested or available to give. Finally, participants showed concern and curiosity about the criteria to be used in deciding the final configuration of the team. This highlighted the need for further work to investigate which set of different team configurations is more effective when several are available.

7.1.4 Chapter 5

After integrating willingness belief into human-machine team design, we wanted to test the effects of doing willingness-based task allocation. As such, Chapter 5 tackles the question *"How does using human willingness for task allocation affect human-AI teamwork?"*. We conducted a user study where participants collaborated with virtual robots on a simulated search and rescue scenario, and the task allocation strategy was different across conditions. The scenario was designed in such a way that it elicited willingness variation across tasks and that these could be distributed in different ways between the human and the virtual collaborative robot. We analysed the effects in the human performance, trust, and satisfaction of including willingness and human control in the task allocation process. The study suggests that although participants prefer robots and plans that consider their preferences, this does not affect their trust or team fluency when the task is quick, critical, and the robot is efficient. Our data suggests that for lower stakes tasks and longer-lived teams, willingness may have a greater influence on human collaborators, and, as such, should be considered for task allocation.

7.1.5 Chapter 6

Finally, the last content chapter of this dissertation, Chapter 6, is dedicated to answer the question *How to build multidisciplinary theory for human-AI team trust?*. This question is answered with guidelines for an effective multidisciplinary collaboration based on our own experience of building theory with researchers from other disciplines. We propose that effective multidisciplinary collaboration begins with recognising which skills are needed from other disciplines and which expertise (e.g., programming, experimental design, theoretical knowledge) can be offered in return. After connecting with researchers who may have the necessary skills, collaboration should be sustained through regular meetings, small shared tasks, and long-term goals such as the organisation of papers or workshops. Furthermore, building community, both within the collaboration and across home disciplines, is necessary as it helps attract new members and maintain momentum. In general, the most important factor is establishing common ground, asking open questions and accepting differences in skills and methods as strengths.

7.2 Answering the overarching research question

To help us summarise the answer to our main research question, i.e., *How can an artificial agent model artificial trust in its human teammates for effective human–machine teamwork?*, we present Figure 7.1. The answer to this overarching question lies in treating artificial trust as a dynamic context-aware process that connects information, beliefs, decisions, outcomes, and communication. More concretely, the chapters' contributions suggest that an artificial agent can model artificial trust through a layered process: (1) formalising trust as context-sensitive beliefs about teammate trustworthiness; (2) assessing human trustworthiness by combining internal characteristics with cost–benefit reasoning; (3) embedding these assessments in team design frameworks that balance competence and willingness; (4) adapting task allocation and coordination strategies to reflect human preferences and behaviours; and (5) grounding the entire modelling process in multidisciplinary theory building.

In the centre of Figure 7.1, we have the beliefs of artificial trust (AT), explored across the chapters. Although artificial trust can be composed of different beliefs, it is recurrent in this dissertation that we need to assess internal characteristics related to the human's competence in the task and willingness towards a task (Chapters 4 and 5). Chapter 2 showed that these beliefs need to be modelled according to the context, such as the team, its dynamics, and task characteristics, while Chapter 3 adds that they are also influenced by the user's cost–benefit reasoning that underlies human choices. To form these beliefs, the artificial agent should then collect necessary and available information such as the task environment, human behaviour cues (as in Chapter 2), and explicit preferences (as in Chapters 3 and 4). These beliefs can guide the artificial teammate's decision-making, as well as human-centred team design with the ITA framework (Chapter 4). Although in this figure focusses mainly on the perspective of the artificial agent's decision-making, this scheme can be added to different decision-making strategies in human-machine teams, such as joint decision-making [36, 142, 312]. In fact, ITA shows potential as a decision-making tool and as a communication bridge among human and machine teammates. Given the promising yet untested results of Chapter 4 in human-machine collaboration, in Chapter 5 we evaluated the effects of using willingness in task allocation. Our results showed that people prefer artificial teammates that use willingness-based task allocation. Even if immediate trust levels remain stable, Chapter 5 points to the importance of aligning artificial agent behaviour with human preferences for long-term collaboration. We pinpoint throughout that the artificial agent's artificial trust, decisions, and how both are communicated may impact the human-machine relationship. This relationship, including the human's trust in the machine teammate, can affect the human's behaviour and, consequently, this should affect the machine's perception of their trustworthiness.

In summary, an artificial agent can model artificial trust by continuously collecting multifaceted information, updating beliefs about competence and willingness, using those beliefs to guide team decisions, and sustaining reciprocity through communication. This dynamic and adaptive process allows artificial trust to support effective, cooperative, and resilient human-machine teamwork.

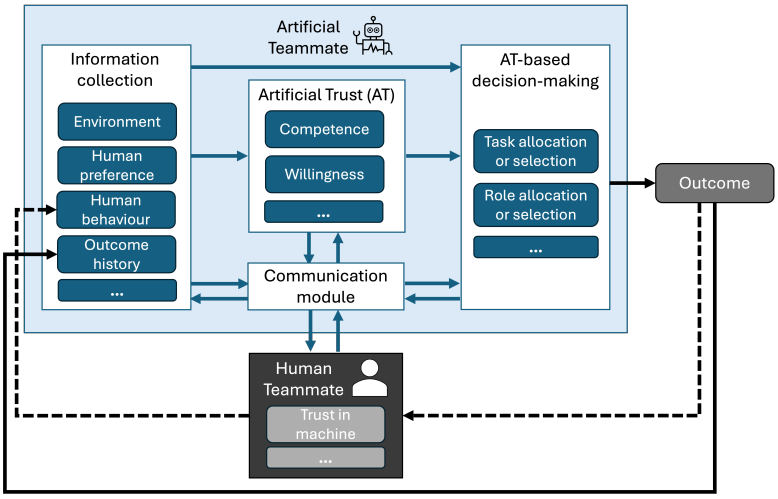


Figure 7.1: This diagram shows different components that are relevant for AT-based decision-making in a team composed of one human and one agent. It includes the different phases of information collection, the modelling of artificial trust (AT), the decision-making based on AT, the effect of the outcome on the system, and the communication as a means of both output and input for the different phases of the process. The diagram also shows how the outcome and communication may affect the human teammate (which is a black box for the machine), in particular their trust, and how this may, in turn, affect their behaviour (which will then affect the AT model, etc).

7.3 Scientific implications

7.3.1 Assessing contextual human trustworthiness

This work supports the idea that artificial trust is not a static variable, but a process that adapts to task context and interaction. Chapter 2 defends that trustworthiness beliefs should not be inferred from static traits alone but through task-specific cues, highlighting the importance of situational context in team design frameworks such as Interdependence Analysis [194]. This affects how classical trustworthiness models, such as *ABI* [250] can be used, in the sense that the internal characteristics that are relevant to computationally infer human reliance for a certain task, and how those characteristics can be observed, depend on the context. For example, benevolence may not be relevant for all tasks, or the relevant related behavioural cues may differ. This means that the trustworthiness dimensions must be defined and weighed in terms of task and team, and in a way they can be computed, e.g., by defining the related. Furthermore, our empirical findings in Chapter 3 suggest that integrity and benevolence-related cues were weak predictors in short-term problem solving tasks. Instead, behavioural strategies appear to be stronger indicators of which task the human teammate is going to perform. This implies that scientific work on artificial trust should expand beyond *ABI* to include cost–benefit reasoning [40, 355, 406] and preference-based cues [107, 427] in different task and team scenarios, which better capture the dynamics of human–machine interaction in collaborative contexts.

7.3.2 Willingness in team design and decision-making

Chapter 4 shows that willingness should be considered alongside competence when designing human–machine teams. This improves our understanding of shared mental models and decision-making within such teams. Furthermore, Chapter 5 shows that the role of willingness in task allocation depends on the task itself. For non-critical or long-term tasks, people want their willingness to be taken into account, supporting decision-making methods that already take some preferences into account [107, 427]. These results align with Rosalind Picard's argument that emotions are integral to human reasoning and interaction, and that ignoring them in machine design leads to alienation and failure in human–computer interaction [303]. Other literature also suggests that ignoring user preferences can lead to fatigue and burnout [263, 369]. These perspectives also align with the literature on team resilience, which proposes that human-machine teamwork should be designed to ensure team adaptability to unexpected changes [349]. Our results, together with the literature, challenge approaches that focus only on performance measures [16, 25, 51, 194, 243, 410]. However, for high-stakes tasks, competence and efficiency remain more important, and performance-focused methods remain relevant. This suggests that models of team design and task allocation intended for use across different tasks and contexts should include both competence and willingness, with rules for when each matters most. These contributions highlight the importance of considering the long-term robustness and sustainability of collaboration, rather than immediate efficiency, when designing human–machine teams and task allocation models.

7.3.3 Meaningful human control vs user effort

Chapter 4 presents a trade-off between meaningful control and user effort in designing human–machine systems. Although experts and legislators emphasise that humans should retain meaningful control over autonomous agents [1, 2, 331], this control often comes at the cost of user effort. For instance, requiring users to provide detailed contextual information in frameworks such as the IA table [193, 194] or the ITA table (presented in Chapter 4) can make the system more transparent and explainable, but can also increase user workload and reduce efficiency in real-time scenarios. From our understanding, it is likely that users may not be interested or even able to put in the necessary effort. This may mean that we (experts, practitioners, developers, designers) need to better explain the advantages of keeping in control of the system. In any case, minimising user effort, e.g., through guided dialogues or semi-automated input collection, while still supporting ethical compliance, context-sensitive adaptation, and meaningful human control, could help keeping the user in control. This also emphasises the importance of adequate communication design for effective human-machine teamwork, as suggested by [396, 397]. Salas et al. [329] suggest that closed-loop communication is a driving mechanism for effective teamwork in human teams. However, the aim of teamwork design should not just be effectiveness, but also meaningful human control. More research is required to understand exactly how communication can not only drive effective teamwork, but also facilitate this meaningful human control.

7.3.4 Multidisciplinary integration of trust research

Finally, the dissertation shows that modelling artificial trust requires drawing from several research fields, including organisational psychology [144, 249, 250, 376], multiagent systems[49, 124], decision theory [40, 355, 406], preference learning [107, 427], team design [194], and wellbeing research [263, 369]. The implication is that trust in human–machine teams cannot be advanced within a single discipline. It requires integrated theoretical frameworks that connect robotics, AI, psychology, human–computer interaction, and related fields. Computer scientists, for instance, focus on how models can be efficient, accurate, and integrated into software, as well as on what inputs they require and what outputs are possible. Psychologists, on the other hand, focus on human behaviour, combining large bodies of theoretical and empirical work into new models that can then be tested through observation. In fact, AI has historically been multidisciplinary, drawing inspiration from biology and human behaviour [323].

Academia, however, often rewards researchers for becoming specialists in narrow fields [4, 21]. This encourages us to dig deeper into our own areas of expertise, ideally without running into someone else's line of work. But human–machine collaboration cannot be solved by tunnelling further into isolated disciplines. We need computer scientists to learn from psychologists which human characteristics can serve as model dimensions, which behaviours matter as inputs, and which algorithmic outputs are required by people. Likewise, psychologists need to understand how algorithms function, their applications, and their limitations, so they can study how people perceive machines, how behaviour changes in response, and how new models may then feed back into computational research. Working in isolation reduces the relevance of advances in all fields that mean to be applied to society [206]. Bridging methods, definitions, and models across disciplines requires persistence, humility, and a willingness to sacrifice some of the obvious academic rewards such as commendations and promotions. Our experience highlights the need to build common ground, which means being open to questions across disciplines while respecting their differences.

7.4 Limitations & Future Directions

While this dissertation addresses several research questions, many remain open. In this subsection we highlight limitations that were prevalent across the course of this dissertation, from information collection, to decision-making, outcome and communication model, as shown in the modules in Figure 7.1. At the end, we also reflect on the evaluation and scalability of our work.

7.4.1 Information collection

While Chapters 3 and 5 discuss behavioural cues and preference-detection measures (e.g., [107, 289, 427]), these approaches do not provide concrete metrics for modelling willingness. To simplify the connection between willingness and all the plethora of characteristics that can influence it across different tasks and contexts, this work often generalises these characteristics into effort-related and reward-related factors. However, there may be theories that would model willingness better for certain scenarios. For example, work motivation theories (e.g., [105, 161, 325]) identify mechanisms such as autonomy, intrinsic interest,

or social obligation. Future work could model these and other mechanisms explicitly, identify observable proxies, and examine their impact on willingness and trustworthiness in different contexts. Finally, in the context of teamwork, willingness may also be shaped by perceptions of other teammates' competence and willingness. Although Chapters 2, 3, and 5 touch on this issue implicitly, this dissertation focused solely on task-related willingness and did not investigate how teammate characteristics influence willingness or task selection decisions. Future research could explore how interdependencies among teammates affect individual willingness and task allocation.

7.4.2 AT-based decision-making

Different combinations of beliefs could lead to the same overall trust value [122, 125, 380]. For instance, high competence paired with low willingness could yield the same score as low competence paired with high willingness. In such cases, an artificial agent may face several possible actions and it has to decide which one is the best [49]. Learning which decision is best for each situation matters because willingness may be more important than competence in some scenarios but less important in others. For example, if the goal is to help a human teammate improve competence, then it may be better to choose tasks where competence is low but willingness to learn is high. Chapter 2 discusses how the weight of each belief may depend on the task, though this was not tested empirically. Even with assigned weights, ties can still occur, especially when multiple teammates are capable of and willing to perform several tasks. In such cases, more information is needed to guide task allocation, selection, or overall team design (as discussed in Chapter 4). Future research should explore how to resolve these ties. One approach to improve tie-breaking could be to extend calculations of expected team performance to include additional factors such as workload. In addition, future work could further develop the formal framework of trust and trustworthiness beliefs within the context of human-agent teamwork. Particularly, the logical representation introduced in Chapter 2 could be extended to support formal analysis, e.g., through formal semantics and a complete axiomatisation, as in [189].

7.4.3 Outcome consequences

In Figure 7.1 we see AT-based decision-making as part of a closed loop. This is supported by the literature and mentioned throughout this dissertation [16, 376]. However, the actual effects of ad hoc trust-based decisions and how these are shaped by task outcomes are not well understood. Chapter 5 looks at this issue by investigating the effects of willingness-based task allocation, but further work is needed. In particular, future studies could examine how human trust in an artificial teammate changes depending on the consequences of allocated tasks and roles, as in [25, 288]. Communication appears important in this loop, as it makes the artificial agent's reasoning clearer and allows human feedback, which supports mutual adaptation.

7.4.4 Communication module

In all user studies presented in this dissertation, communication channels between the human and the artificial teammate were included. However, these did not cover how to share the artificial agent's formed trust beliefs. Communication, including an artificial agent's transparency and explainability, is important to guarantee human trust, shared

understanding, ensuring alignment with the human and avoiding possible accidents [81, 175, 231, 418, 424]. Centeio Jorge et al. [62] suggest ways of doing this, by experimenting with different communication styles. Communication between humans and machines can be done in several ways, varying in modalities [12, 23, 318], levels of explanation [88, 297, 397], and timings [80, 424], among others. Centeio Jorge et al. [62] takes a first step toward exploring which styles work best for communicating artificial trust in different teamwork scenarios, but more work is needed. Specifically, it is necessary to explore the effects of expressing negative trust [214, 237]. For example, it is not clear how people would react if an artificial teammate signalled that it did not trust them, and in what situations this could be a problem.

7.4.5 Updating artificial trust beliefs

Trust dynamics can happen over longer time intervals than the ones presented in our studies. Although it is argued that artificial trust beliefs showed in this dissertation should change over time to allow adaptation, no method for updating them was formalised or tested. The paper by Centeio Jorge et al. [62] proposes such methods, but they were not applied in this dissertation. Methods for updating beliefs of trust through interaction and building memory presented in literature [25, 49, 82, 157, 180, 230, 345] should be adapted to our frameworks in future research. Furthermore, as an agent gathers more information, its beliefs may become more certain, but how this certainty should affect decision-making is still unclear. As such, future work should investigate how to update artificial beliefs and how to include memory and certainty in decision-making methods for human-machine teamwork.

7.4.6 Evaluating artificial trust models

Evaluating how well artificial trust models represent human trustworthiness is difficult since they represent human internal characteristics that can only be inferred indirectly from observable behaviour. Which behavioural cues relate to which of the internal characteristics is also something that is not exact or well established. This means that there is no objective ground truth for judging the relevance of the chosen characteristics, or the connection between the represented characteristics and the behaviour observed. Unlike most machine learning problems, knowledge-based models cannot be evaluated through sets of input–output, which make them also harder to being compared against other models of the same kind [180, 248]. They are, however, important to maintain transparency and meaningful control [30, 109, 248]. In Chapter 3, we relied on self-reports of trustworthiness to assess the relevance of our model's dimensions. Although useful, these measures introduce subjectivity and cannot be assumed to reflect an objective truth. Ultimately, we consider a good model as one that supports effective teamwork and leaves humans satisfied with the interaction, as in Chapter 5. However, our experimental setups, such as the user studies in Chapters 3 and 5, have their own limitations.

While useful for controlled testing [400], these environments may not reflect how people behave in real settings, reducing their ecological validity. Running studies with non-physical artificial agents provide a cheaper, more controlled and faster to implement way to study concepts [117]. However, this is often pointed out by researchers as a limitation, since virtual or online studies may not provide the same level of engagement, or elicit the

same feelings and emotions, which are indeed important to study the interaction, including trust dynamics [117, 234]. The main trade-off is that the more realistic the study setup is, the harder is it to isolate the object of study and obtain clear results [174]. The choices of task, participant pool, agent embodiment, and environment can all impact the results of the study, so ideally these studies are to be repeated with variations across these dimensions [234].

7.4.7 From dyads to groups

Although this dissertation frames artificial trust as relevant to teams involving multiple humans and/or machines, most of our work focused on individual beliefs for dyadic interactions. Even at the dyadic level, these interactions are not fully captured, as many interpersonal factors, such as mutual perceptions, interdependencies, and context-dependent behaviours, remain unexplored. It is our understanding that a detailed understanding of these dyadic dynamics is a necessary precursor before investigating group-level trust and its effects on individual behaviour. Chapter 3 examined a human working with two artificial agents, but did not address team-level trust dynamics. Studying trust in teams introduces challenges beyond those in dyads, as it requires modelling beliefs about multiple teammates, the different relationships between them, and the layered structure of team trust [144, 376]. It also requires taking into account team characteristics such as composition, hierarchy, or shared knowledge [300], as included in Chapter 2's taxonomy. Studying groups demands bigger and more expensive experimental setups, usually technically more complex (e.g., several robots, or a server allowing several people at the same time), with more metrics, and more participants [92, 97, 240, 274]. Finally, thinking beyond the dyad inevitably brings focus to the role of reputation and communicated experiences [166, 327, 328] in the modelling of artificial trust, which should be explored in future work.

Our choice to focus on dyads was deliberate. Although team dynamics is an important part of the artificial trust when a team is bigger than two, we first need a solid understanding of how it should form between two entities. However, extending our work to groups will require capturing how multiple dyads interact and form the broader structure of team trust and other dynamics [297]. Therefore, progress on modelling artificial trust in teams will depend on first having a solid grasp of how trust operates within dyads, while also extending the frameworks to account for the complexity of larger and more diverse groups.

7.5 Societal contributions

This dissertation makes several contributions with relevance beyond academia, shaping human–machine teamwork for societal benefit. Understanding how humans behave towards collaborative technology, as well as how to make such collaborations sustainable and trustworthy, adds to society in several areas, ranging from rescuing scenarios to daily chores. In this section, we highlight some of the contributions we believe this dissertation can bring to society.

7.5.1 Safety and effectiveness in high-risk environments

In domains such as firefighting, disaster response, and nuclear decontamination, humans face tasks that can be harmful or fatal, such as reaching victims through flames or cleaning

areas with lethal radiation. Machines, on the contrary, could perform some of these tasks, improving human well-being, both of practitioners (e.g., firefighters, possible workers that need to clean those radioactive areas), and potential civilians involved (e.g., victims of a fire, people living near those radioactive areas) [11, 18, 132, 416]. However, sending those machines to working stations raises several challenges [191]. For example, machines should not perform certain tasks that involve moral deliberation, such as deciding which victim to prioritise, when resources are scarce, as these should remain under human control [331]. Coordinating this collaboration requires the ability to determine which tasks are better suited for the machine, which tasks humans can safely perform, and which the interdependence level each task requires (i.e., assisted, independent, joint) [194, 232].

This dissertation helps this societal problem by presenting the framework in Chapter 4, which provides a method to analyse the competence, willingness, and influencing factors (permissions, constraints) of all teammates for different tasks, supporting transparent and informed team design and task allocation. Our work also identifies the characteristics of the task and the team (Chapter 2), such as the nature of the task and its criticality, that influence the human traits most relevant to assess in each context. Finally, the results in Chapter 5 suggest that in high-risk environments, people prefer task allocations optimised for efficiency and effectiveness rather than for individual preferences. These findings support the improvement of the safety of human teammates, potential victims of the situation (such as in a fire), and effectiveness of teams in high-risk environments.

7.5.2 Work quality

Work occupies a large part of daily life, and lack of enjoyment or persistent discomfort can have serious consequences such as burnout, depression, and chronic pain [8, 263, 292, 344, 412]. Machines could help with this problem by taking on physically demanding or psychologically taxing tasks [11]. Unfortunately, the logics of automation that originated in the Industrial Revolution, namely displacement of labour, standardisation of work, and reduction of human agency, still shape technology design today, often at the expense of worker well-being [34, 366]. Poorly integrated human–machine teams could reinforce these patterns by lowering autonomy, potentially creating social detachment, reducing motivation, and even productivity [75, 120, 233, 244].

A recent case in Amazon, a highly performance-orientated company [158] where robots almost outnumber human workers [169], shows how even in the most ambitious places there is a need to step back and focus on human factors. Amazon's newest fully autonomous robot *Proteus*, shows expressive features such as eyes and mouth-like indicators, making its intentions understandable while avoiding frightening or annoying human colleagues [19, 152]. This aligns with our argument that focussing on performance measures, without caring about human well-being and needs, may undermine productivity in the long run. In any case, this dissertation focusses on advancing collaborative technologies that improve well-being and work quality, despite corporative productivity. The contributions of Chapter 4 and Chapter 5 provide guidance for this societal dimension. By integrating willingness into transparent team design and task allocation, machines can be directed towards tasks people would rather avoid, leaving humans with those they value. This has the potential to improve job satisfaction, preserve health and dignity, and make human–machine collaboration more sustainable.

7.5.3 Calibrating trust and responsibility

Although machines have great potential to be our counterparts in several collaborative scenarios, humans do not necessarily understand them when interacting [56, 426]. This can become a problem when people assign wrong trustworthiness characteristics, leading to mistrust (when they trust more than they should) or distrust (when they trust less than they should) [232]. The first can cause harm, loss of control, and sense of responsibility, whereas the second can lead to inefficiencies, which can indirectly bring harm as well [48, 261]. Chapter 2 reflects on how artificial trust can affect human trust, proposing that it can help calibrate human trust. Similarly, chapter 4's framework's transparency can help regulate the understanding of the machine's capabilities and permissions, which can also calibrate human's trust and sense of responsibility. This helps address societal concerns about accountability and ethical responsibility in the deployment of autonomous systems.

7.5.4 Developing technology with people in the loop

Developing computational models without considering their implications in the real-world systems where they will be deployed can lead to failure in achieving intended goals. In particular, developing and evaluating collaborative agents without involving humans limits its potential social benefit, as it can overlook the impact on the human teammate, who is a central part of the system [150, 207, 272, 313]. Not focussing on the human throughout the different stages of technology cannot guarantee alignment with society and user needs and values, and can even lead to harmful physical and psychological consequences [34, 172]. Throughout our work, we draw on literature from multiple disciplines and evaluate our contributions using methods that always incorporate human input, doing our best to align our advancements with potential users.

In order to receive feedback from potential users, we conducted focus groups (Chapter 4) and user studies (Chapters 3 and 5). In fact, in Chapter 4 we adapt our method to a potential use case, search and rescue, and invite firefighters to participate in the focus group and evaluate our framework. This made the evaluation of the framework more realistic and aligned with the needs of the practitioners. Furthermore, we collected qualitative data in all user studies (Chapters 3 and 5), to ensure that we better understand the needs and interpretations of the users.

Finally, we also collaborated with researchers from the social sciences and human factors to better propose the role of artificial trust in teams (Chapter 2) and in task allocation (Chapter 5). The goal was again to better align our methodology with real-world teams and human behaviour. This work demonstrates how effective collaboration across disciplines, such as computer science and social sciences, can produce models and methods that are both computationally feasible and socially grounded. This type of collaboration is essential for building systems that are trustworthy, acceptable, and aligned with societal values [112, 268, 284].

7.5.5 Designing ethical AI systems

Safely and responsibly integrating autonomous systems (such as AI teammates) into critical infrastructures and services is a challenge, as it touches on ethical and legal constraints [172, 319]. Our methods supporting team design (Chapter 4) propose the analysis of possible interdependences given the external factors, explicitly pointing at possible AI legislation

and ethical considerations. This provides a lens for policy makers and practitioners to evaluate when and how to integrate autonomous teammates into critical infrastructures and services. It also highlights the need for governance structures that ensure artificial trust mechanisms to ensure that the effort required from the users does not undermine their willingness to responsibly use autonomy.

7.6 Potential risks for society

Although we built this work focussing on the positive contributions it could bring to society and science, it is necessary to acknowledge the potential risks of enabling artificial agents to build models of human trustworthiness. In particular, we reflect on the undesired consequences and possible unethical usage of our work.

7.6.1 Reducing human involvement in decision-making

Delegating trust-based decisions to machines may pose some unwanted risks. For example, it may distance humans from accountability for the outcomes of those decisions, such as assignment of tasks [290]. In critical domains, this could make it unclear who is responsible for errors or harm, raising ethical and legal concerns [222]. Furthermore, if machines dominate task allocation or collaboration design, humans may feel disadvantaged or marginalised in teams [78, 120, 203]. This risks lowering human motivation, engagement, and satisfaction with tasks and collaborations [213, 382]. These limitations can be addressed by ensuring that AI recommendations are transparent and adjustable, so humans can retain control and accountability, and by defining clear rules for which decisions remain under human responsibility while collecting feedback to maintain engagement and trust.

7.6.2 Inappropriately trusting the human teammate

As mentioned above, both under-trust and over-trust can harm performance [232]. Surely, miscalibrated artificial trust, such as overestimating human willingness or underestimating human competence, could also lead to poor coordination, accidents, or loss of effectiveness in high-stakes scenarios [232, 261, 404]. Furthermore, when a machine consistently allocates tasks away from someone it considers less trustworthy, it may strip that person of agency, responsibility, or opportunities for growth. Over time, this could damage morale, create perceptions of unfairness, and erode human dignity in collaborative settings. These limitations can be addressed by improving communication and changing belief formation from exploration to exploitation from time to time [220, 228]. However, the risks of using artificial agents to decide who is trustworthy for what go beyond that.

7.6.3 Social exclusion and inequality

Artificial trust models may inadvertently disadvantage certain individuals or groups of people [319]. This can happen if there is a problem with the model, if the trustworthiness criteria reflect generic or biased assumptions, or if the models are used for something other than effective teamwork. For example, if willingness is inferred from observable behaviour, people with disabilities, atypical communication styles, or cultural differences could be excluded from meaningful roles. This risks reinforcing inequalities in work and society. Furthermore, the outcomes of the model can be used to classify someone as generally

more or less trustworthy, removing them from the specific context of decision-making, which can harm people in different ways [138]. These risks can be mitigated [84] by using inclusive, context-specific trust criteria, bias monitoring, and maintaining human oversight to prevent unfair exclusion or generalised judgements.

7.6.4 Mutual distrust between humans and machines

As mentioned before, artificial trust may have an impact on human trust in the artificial agent. This means that artificial trust can irreparably damage human trust in technology. If a machine repeatedly judges humans as untrustworthy and withholds tasks or information, and the human considers this to be wrong, humans may start distrusting the machine in return [89]. This breakdown in mutual trust could reduce team cohesion, undermine cooperation, and even cause humans to bypass or sabotage the system [61, 216]. In addition, if certain people experience repeated exclusion or marginalisation by autonomous systems, while others do not, this can create unnecessary tension and division in society [113, 168, 423]. This can be addressed by activating mechanisms to, for example, ensure no individual is consistently left out certain tasks without explicit consent, or allowing humans to signal when they feel unfairly treated or misjudged.

7.6.5 Reflection on human-machine collaboration for defence

Military institutions are currently among the largest sponsors of research and development in autonomous and collaborative robotics. In the United States, for example, the Department of Defense has directed hundreds of millions of dollars to AI, robotics, and autonomy programmes [374]. Similarly, in the Netherlands, the Ministry of Defence collaborates extensively with the Netherlands Organisation for Applied Scientific Research (TNO) in human-machine teamwork. TNO runs a dedicated Human–Machine Teaming unit, develops Defence roadmaps, and has proposed frameworks such as DASH (Delegation to Autonomous Systems within Human–Machine Teams) to balance autonomy and meaningful human control in military operations [372, 387]. The 2022 TNO Defence Programme flyer says “Future military operations will increasingly involve humans and intelligent technologies working together closely as human-machine teams (HMTs)” [372].

These investments show that the most mature and best-funded human-machine teaming applications are likely to be used in warfare. This changes the nature of war and brings new societal risks [182]. One central concern is escalation. When a human pilot is involved, the personal risk and moral judgement function as natural brakes on engagement. In contrast, unmanned or autonomous systems reduce the perceived costs of initiating strikes, which can make military action more frequent and increase the likelihood of escalation of the conflict [334, 352]. Furthermore, autonomous systems making life-and-death decisions with limited oversight raise questions about accountability, compliance with international humanitarian law, and responsibility gaps [182].

Both countries shown as an example in this section have programmes to also reflect on and institute ethical and legal boundaries on the use of semi-autonomous technology. The Netherlands presents ELSA Lab Defence [114], led by TNO, and the US presents ASIMOV (Autonomy Standards and Ideals with Military Operational Values) [98], led by DARPA. Although this is an important step in the development of such technologies, it does not ensure anyone that the future of war will not escalate more easily given the already existing

and under-development tools.

It is important to stress that this dissertation does not seek to contribute to the escalation of war through human-machine teaming. Our motivation and focus are on civilian contexts, where collaboration between humans and machines should serve human well-being, safety, and meaningful work. Military research provides a cautionary reminder: without careful governance and human-centred design, collaborative technologies risk being deployed in ways that accelerate violence rather than support societal resilience.

7.7 Take-away message

The central message of this dissertation is that artificial trust can be used and designed for decision-making as a dynamic, context-aware process that goes beyond competence to include human willingness, such as preferences and cost–benefit reasoning. It has the potential to make human-machine teamwork not only efficient but also sustainable and meaningful for the humans involved. On the positive side, the work shows how modelling artificial trust can improve collaboration, ensure more human-centred technology, and create systems that align with user needs. On the cautionary side, it highlights that these same mechanisms, if misapplied, risk reinforcing exclusion, eroding human autonomy, or being used for military escalation. Therefore, the broader takeaway is that the purpose of building artificial trust is not to maximise efficiency at all costs, but to create collaborative systems that enhance human well-being, resilience, safety, and dignity. This requires technical advances, multidisciplinary collaboration, and continuous reflection on the societal contexts in which these technologies are deployed.

A

Appendix

A.1 Interdependence and Trust Analysis Table (Version 1)

Task	M's role	M's C	M's P	M's I	M	H's role	H's C	H's P	H's I	H	Feasibility
Wash the veggies	Performer (w/ support)	g	g	g	g	Supporter	g	g	r	y	y
	Performer (independent)	g	g	g	g	Not involved					g
	Supporter	g	g	g	g	Performer (w/ support)	g	g	g	g	g
	Co-Performer	g	g	g	g	Co-Performer	g	g	r	y	y
	Not involved					Performer (independent)	g	g	g	g	g
Chop the veggies	Performer (w/ support)	g	r	g	r	Supporter	g	g	g	g	r
	Performer (independent)	g	r	g	r	Not involved					r
	Supporter	g	g	g	g	Performer (w/ support)	g	g	g	g	g
	Co-Performer	g	r	g	r	Co-Performer	g	g	g	g	r
	Not involved					Performer (independent)	g	g	g	g	g
Frying the veggies	Performer (w/ support)	g	g	g	g	Supporter	g	g	g	g	g
	Performer (independent)	r	g	g	o	Not involved					o
	Supporter	g	g	g	g	Performer (w/ support)	r	g	g	o	o
	Co-Performer	r	g	g	o	Co-Performer	r	g	r	r	r
	Not involved					Performer (independent)	r	g	r	r	r

Figure A.1: First version of the Interdependence and Trust Analysis (ITA) table, presented in Chapter 4, which was evaluated in the first phase of evaluation. It includes an extensive distinction among all the five possible roles each teammate can play when performing a task in a team composed of one human (H) and one machine (M). The table assesses the feasibility of each teammate individually, through their dimensions of competence (C), possibility (P, which included the external factors), and intention (I). Each fillable cell (the dimensions) could be filled with green (g) or red (r) colours. This automatically fills in the overall feasibility of the teammate for that role (M column for machine's feasibility and H for human's). The feasibility columns can be (1) green if all three dimensions are green, (2) yellow if intention is red and all others are green, (3) orange if competence is red, or (4) red if another combination of red occurs. The final feasibility column can be (1) green if both feasibilities are green, (2) yellow if one is yellow and the other is green or yellow, (3) orange if one is orange and the other is green or orange, and (4) red if another combination occurs. Grey is ignored when calculating feasibility.

Glossary

ABI	Artificial, Benevolence and Integrity model of trust as in Mayer et al. (1995) [250].
AI	Artificial Intelligence.
AT	Artificial Trust: a belief of trust where the trustor is an artificial agent (as explained in Chapters 1 and 2).
ITA	Interdependence and Trust Analysis framework, described in Chapter 4.

Bibliography

References

[1] Artificial intelligence act. https://artificialintelligenceact.eu/. Accessed: 2025-09-15.

[2] Ethically aligned design. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf. Accessed: 2025-09-15.

[3] An efficient and versatile approach to trust and reputation using hierarchical bayesian modelling. *Artificial Intelligence*, 193:149–185, 12 2012.

[4] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Flavia Di Costa. Diversification versus specialization in scientific research: Which strategy pays off? *Technovation*, 82:51–57, 2019.

[5] Sami Abuhaimed, Selim Karaoglu, and Sandip Sen. Choosing the task allocator: Effect on performance and satisfaction in human-agent team. In *The International FLAIRS Conference Proceedings*, volume 36, 2023.

[6] Barbara D. Adams, Sonya Waldherr, and J. Sartori. Trust in teams scale, trust in leaders scale: Manual for administration and analyses. 2008.

[7] Barbara D. Adams and R. Webb. Trust in small military teams. *Command and Control Research Program*, 2002.

[8] Greig Adams and Tim V Salomons. Attending work with chronic pain is associated with higher levels of psychosocial stress. *Canadian Journal of Pain*, 5(1):107–116, 2021.

[9] Ighoyota Ben Ajenaghughrure, Sonia Claudia DaCosta Sousa, and David Lamas. Measuring trust with psychophysiological signals: A systematic mapping study of approaches used. *Multimodal Technol. Interact.*, 4(3):63, 2020.

[10] Nirav Ajmeri, Pradeep K. Murukannaiah, Hui Guo, and Munindar P. Singh. Arnor: Modeling social intelligence via norms to engineer privacy-aware personal agents. In Kate Larson, Michael Winikoff, Sanmay Das, and Edmund H. Durfee, editors, *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*, pages 230–238. ACM, 2017.

[11] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry

Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.

[12] Alican Akman and Björn W. Schuller. Audio explainable artificial intelligence: A review. *Intelligent Computing*, 3:0074, 2024.

[13] Nele Albers, Mark A Neerincx, and Willem-Paul Brinkman. Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active. *Plos one*, 17(12), 2022.

[14] Nele Albers, Mark A. Neerincx, and Willem-Paul Brinkman. Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active: Data and analysis code. 11 2022.

[15] Amjad Alfaleh, Abdullah N. Alkattan, Alaa Alageel, Mohammed Salah, Mona Masad Almutairi, Khlood Sagor, and Khaled Alabdulkareem. Onsite versus remote working: The impact on satisfaction, productivity, and performance of medical call center workers. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, 58, 2021.

[16] Arsha Ali, Hebert Azevedo-Sa, Dawn M Tilbury, and Lionel P Robert Jr. Heterogeneous human–robot task allocation based on artificial trust. *Scientific Reports*, 12(1):15304, 2022.

[17] Saeid Alirezazadeh and Luís A Alexandre. Dynamic task scheduling for human-robot collaboration. *IEEE Robotics and Automation Letters*, 7(4):8699–8704, 2022.

[18] Michael Allen. Robots to the rescue: miniature robots offer new hope for search and rescue operations, Jun 2025. Accessed June 2025.

[19] Amazon. How amazon deploys robots in its operations facilities, 2023. Accessed: 2025-09-13.

[20] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[21] Weiyi Ao, Libo Sheng, Xuanmin Ruan, Dongqing Lyu, Jiang Li, and Ying Cheng. Researching deeply or broadly? the effects of scientists' research strategies on disruptive performance over their careers. *Journal of Informetrics*, 19(2):101657, 2025.

[22] Theo Araujo and Nadine Bol. From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Computers in Human Behavior: Artificial Humans*, 2(1):100030, 2024.

[23] Lilit Avetisyan, Jackie Ayoub, and Feng Zhou. Investigating explanations in conditional and highly automated driving: The effects of situation awareness and modality. *CoRR*, abs/2207.07496, 2022.

[24] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T. Esterwood, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. Real-time estimation of drivers' trust in automated driving systems. *Int. J. Soc. Robotics*, 13(8):1911–1927, 2021.

[25] Hebert Azevedo-Sa, X Jessie Yang, Lionel P Robert, and Dawn M Tilbury. A unified bi-directional model for natural and artificial trust in human–robot collaboration. *IEEE robotics and automation letters*, 6(3):5913–5920, 2021.

[26] Mohammad Q Azhar and Elizabeth I Sklar. A study measuring the impact of shared decision making in a human-robot team. *The International Journal of Robotics Research*, 36(5-7):461–482, 2017.

[27] Rasmus Bååth. Bayesian first aid: A package that implements bayesian alternatives to the classical *.test functions in r. In *UseR! 2014 - the International R User Conference*, 2014.

[28] Michael Bacharach and Diego Gambetta. *Trust as Type Detection*, pages 1–26. Springer Netherlands, Dordrecht, 2001.

[29] Michael Bacharach, Gerardo Guerra, and Daniel John Zizzo. The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 63(4):349–388, 2007.

[30] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11, 2019.

[31] Kevin Baum, Joanna Bryson, Frank Dignum, Virginia Dignum, Marko Grobelnik, Holger Hoos, Morten Irgens, Paul Lukowicz, Catelijne Muller, Francesca Rossi, et al. From fear to action: Ai governance and opportunities for all. *Frontiers in Computer Science*, 5:1210421, 2023.

[32] Fritz Becker, Celine Ina Spannagl, Jürgen Buder, and Markus Huff. Performance rather than reputation affects humans' trust towards an artificial agent. *Computers in Human Behavior: Artificial Humans*, 3:100122, 2025.

[33] Don van den Bergh, Merlise A Clyde, Akash R Gupta, Tim de Jong, Quentin F Gronau, Maarten Marsman, Alexander Ly, and Eric-Jan Wagenmakers. A tutorial on bayesian multi-model linear regression with bas and jasp. *Behavior research methods*, pages 1–21, 2021.

[34] Michael Betancourt. The social paradigm of automation. In *Digitalisierung der Arbeitswelten: Zur Erfassbarkeit einer systemischen Transformation*, pages 311–326. Springer Fachmedien Wiesbaden Wiesbaden, 2024.

[35] Shreyas Bhat, Joseph B Lyons, Cong Shi, and X Jessie Yang. Clustering trust dynamics in a human-robot sequential decision-making task. *IEEE Robotics and Automation Letters*, 7(4):8815–8822, 2022.

[36] Tetiana Biloborodova and Inna Skarga-Bandurova. Human-ai collaboration in decision making: An initial reliability study and methodology. In *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 1, pages 1151–1155, 2023.

[37] Sara Blanco. Human trust in ai: a relationship beyond reliance. *AI and Ethics*, pages 1–14, 2025.

[38] Benjamin Samuel Bloom, Committee of College, and University Examiners. *Taxonomy of educational objectives*, volume 2. Longmans, Green New York, 1964.

[39] Franziska Bocklisch and Norbert Huchler. Humans and cyber-physical systems as teammates? characteristics and applicability of the human-machine-teaming concept in intelligent manufacturing. *Frontiers in artificial intelligence*, 6:1247755, 2023.

[40] Matthew M. Botvinick and Zev B. Rosen. Anticipation of cognitive demand during decision-making. *Psychological Research PRPF*, 73:835–842, 2009.

[41] Jeffrey M Bradshaw, Paul J Feltovich, Hyuckchul Jung, Shriniwas Kulkarni, William Taysom, and Andrzej Uszok. Dimensions of adjustable autonomy and mixed-initiative interaction. In *Agents and Computational Autonomy: Potential, Risks, and Solutions 1*, pages 17–39. Springer, 2004.

[42] Diego De Siqueira Braga, Marco Niemann, Bernd Hellingrath, and Fernando Buarque De Lima Neto. Survey on computational trust and reputation models. *ACM Comput. Surv.*, 51(5), November 2018.

[43] Matthias Brand, Kirsten Labudda, and Hans J Markowitsch. Neuropsychological correlates of decision-making in ambiguous and risky situations. *Neural Networks*, 19(8):1266–1276, 2006.

[44] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[45] Christina Breuer, Joachim Hüffmeier, Frederike Hibben, and Guido Hertel. Trust in teams: A taxonomy of perceived trustworthiness factors and risk-taking behaviors in face-to-face and virtual teams. *Human relations*, 73(1):3–34, 2020.

[46] Ann L Brown. Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The journal of the learning sciences*, 2(2):141–178, 1992.

[47] Edgar Brunner and Ullrich Munzel. Nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Planning and Inference*, 82(1-2):163–181, 2000.

[48] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.

[49] Chris Burnett, Timothy J Norman, and Katia Sycara. Trust decision-making in multi-agent systems. In *Proceedings of the twenty-second international joint conference on artificial intelligence-volume volume one*, pages 115–120, 2011.

[50] Chris Burnett, Timothy J. Norman, and Katia Sycara. Stereotypical trust and bias in dynamic multiagent systems. *ACM Transactions on Intelligent Systems and Technology*, 4, 3 2013.

[51] Min Cai, Gen Wang, Xinggang Luo, and Xueqi Xu. Task allocation of human-robot collaborative assembly line considering assembly complexity and workload balance. *International Journal of Production Research*, pages 1–27, 2025.

[52] Lucile Callebert, Domitile Lourdeaux, and Jean-Paul Barthès. A trust-based decision-making approach applied to agents in collaborative environments. In *8th International Conference on Agents and Artificial Intelligence (ICAART 2016)*, pages 287–295, 2016.

[53] Michela Carraro, Andrea Furlan, and Torbjørn Netland. Unlocking team performance: How shared mental models drive proactive problem-solving. *human relations*, 78(4):407–437, 2025.

[54] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. *On the utility of learning about humans for human-AI coordination*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[55] Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca D. Dragan. On the utility of learning about humans for human-ai coordination. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5175–5186, 2019.

[56] Oliver Carsten and Marieke H Martens. How can humans understand their automated cars? hmi principles, problems and solutions. *Cognition, Technology & Work*, 21(1):3–20, 2019.

[57] Christiano Castelfranchi and Rino Falcone. *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons, 2010.

[58] Cristiano Castelfranchi and Rino Falcone. Trust is much more than subjective probability: Mental components and sources of trust. In *Proceedings of the 33rd annual Hawaii international conference on system sciences*. IEEE, 2000.

[59] Cristiano Castelfranchi and Rino Falcone. *Trust & Self-Organising Socio-technical Systems*. Springer International Publishing, 2010.

[60] C. Centeio Jorge. Effects of task allocation using human's willingness in trust and teamwork. `https://osf.io/pfxbq`, June 2025. Retrieved from osf.io/pfxbq.

[61] Carolina Centeio Jorge, Nikki H. Bouman, Catholijn M. Jonker, and Myrthe L. Tielman. Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork. *Frontiers Robotics AI*, 10, 2023.

[62] Carolina Centeio Jorge, Elena Dumitrescu, Catholijn M. Jonker, Razvan Loghin, Sahar Marossi, Elena Uleia, and Myrthe L. Tielman. How should your artificial teammate tell you how much it trusts you? In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA '25, New York, NY, USA, 2025. Association for Computing Machinery.

[63] Carolina Centeio Jorge, Catholijn Jonker, and Myrthe Tielman. Data underlying the study on the effects of task allocation using human's willingness in trust and teamwork., 2025.

[64] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. Artificial trust for decision-making in human-ai teamwork: Steps and challenges. In *CEUR Workshop Proceedings*, volume 3456, 2023.

[65] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. Artificial trust for decision-making in human-ai teamwork: Steps and challenges (short paper). In Pradeep K. Murukannaiah and Teresa Hirzle, editors, *Proceedings of the Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence co-located with (HHAI 2023), Munich, Germany, June 26-27, 2023*, volume 3456 of *CEUR Workshop Proceedings*, pages 150–156. CEUR-WS.org, 2023.

[66] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. Data for interdependence and trust analysis (ITA): a framework for human-machine team design, 2024.

[67] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. How should an AI trust its human teammates? exploring possible cues of artificial trust. *ACM Trans. Interact. Intell. Syst.*, 14(1):5:1–5:26, 2024.

[68] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. How should an ai trust its human teammates? exploring possible cues of artificial trust. *ACM Trans. Interact. Intell. Syst.*, 14(1), jan 2024.

[69] Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. Interdependence and trust analysis (ita): a framework for human–machine team design. *Behaviour & Information Technology*, pages 1–21, 2024.

[70] Carolina Centeio Jorge, Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams. In Dongxia Wang, Rino Falcone, and Jie Zhang, editors, *Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST 2021) Co-located with the 20th International Conferences on Autonomous Agents*

and Multiagent Systems (AAMAS 2021), London, UK, May 3-7, 2021, volume 3022 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[71] Carolina Centeio Jorge, Siddharth Mehrotra, Myrthe L Tielman, and Catholijn M Jonker. Trust should correspond to trustworthiness: A formalization of appropriate mutual trust in human-agent teams. In *22nd International Trust Workshop 2021*, 2021.

[72] Carolina Centeio Jorge, Myrthe L. Tielman, and Catholijn M. Jonker. Artificial trust as a tool in human-ai teams. In Daisuke Sakamoto, Astrid Weiss, Laura M. Hiatt, and Masahiro Shiomi, editors, *ACM/IEEE International Conference on Human-Robot Interaction, HRI 2022, Sapporo, Hokkaido, Japan, March 7 - 10, 2022*, pages 1155–1157. IEEE / ACM, 2022.

[73] Carolina Centeio Jorge, Myrthe L. Tielman, and Catholijn M. Jonker. Assessing artificial trust in human-agent teams: a conceptual model. In Carlos Martinho, João Dias, Joana Campos, and Dirk Heylen, editors, *IVA '22: ACM International Conference on Intelligent Virtual Agents, Faro, Portugal, September 6 - 9, 2022*, pages 24:1–24:3. ACM, 2022.

[74] Carolina Centeio Jorge and Anna Sophie Ulfert-Blank. Multittrust-multidisciplinary perspectives on human-ai team trust. In *CEUR Workshop Proceedings*, volume 3456, pages 132–136. CEUR-WS, 2023.

[75] Christopher P Cerasoli, Jessica M Nicklin, and Alexander S Nassrelgrgawi. Performance, incentives, and needs for autonomy, competence, and relatedness: A meta-analysis. *Motivation and emotion*, 40(6):781–813, 2016.

[76] José-Antonio Cervantes, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. Artificial moral agents: A survey of the current status. *Sci. Eng. Ethics*, 26(2):501–532, 2020.

[77] Wesley P. Chan, Morgan Crouch, Khoa Cong Hoang, Charlie Chen, Nicole L. Robinson, and Elizabeth A. Croft. Improving human-robot collaboration through augmented reality and eye gaze. *ACM Trans. Hum. Robot Interact.*, 14(3):42:1–42:19, 2025.

[78] Mai Lee Chang, Greg Trafton, J Malcolm McCurry, and Andrea Lockerd Thomaz. Unfair! perceptions of fairness in human-robot teams. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 905–912. IEEE, 2021.

[79] R.A. Chechile. *Bayesian Statistics for Experimental Scientists: A General Introduction Using Distribution-Free Methods.* MIT Press, 2020.

[80] Cheng Chen, Mengqi Liao, and S. Shyam Sundar. When to explain? exploring the effects of explanation timing on user perceptions and trust in ai systems. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, TAS '24, New York, NY, USA, 2024. Association for Computing Machinery.

[81] Jessie Y. C. Chen, Frank Ole Flemisch, Joseph B. Lyons, and Mark A. Neerincx. Guest editorial: Agent and system transparency. *IEEE Transactions on Human-Machine Systems*, 50(3):189–193, 2020.

[82] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha S. Srinivasa. Planning with trust for human-robot collaboration. In Takayuki Kanda, Selma Sabanovic, Guy Hoffman, and Adriana Tapus, editors, *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2018, Chicago, IL, USA, March 05-08, 2018*, pages 307–315. ACM, 2018.

[83] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha S. Srinivasa. Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Trans. Hum. Robot Interact.*, 9(2):9:1–9:23, 2020.

[84] Lu Cheng, Kush R Varshney, and Huan Liu. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71:1137–1181, 2021.

[85] Yujiao Cheng, Liting Sun, and Masayoshi Tomizuka. Human-aware robot task planning based on a hierarchical task model. *IEEE Robotics and Automation Letters*, 6(2):1136–1143, 2021.

[86] Jiaee Cheong, Nikhil Churamani, Luke Guerdan, Tabitha Edith Lee, Zhao Han, and Hatice Gunes. Causal-hri: Causal learning for human-robot interaction. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1311–1313, 2024.

[87] Kinzang Chhogyal, Abhaya C. Nayak, A. Ghose, and Khanh Hoa Dam. A value-based trust assessment model for multi-agent systems. *28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019.

[88] Erin K. Chiou, Mustafa Demir, Verica Buchanan, Christopher C. Corral, Mica R. Endsley, Glenn J. Lematta, Nancy J. Cooke, and Nathan J. McNeese. Towards human–robot teaming: Tradeoffs of explanation-based communication strategies in a virtual search and rescue task. *International Journal of Social Robotics*, 14(5):1117–1136, 2022.

[89] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, 127:107018, 2022.

[90] Jiska Cohen-Mansfield, Marcia S. Marx, Laurence S Freedman, Havi Murad, Natalie G. Regier, Khin Thein, and Maha Dakheel-Ali. The comprehensive process model of engagement. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry*, 19 10:859–70, 2011.

[91] Allan Collins, Diana Joseph, and Katerine Bielaczyc. Design research: Theoretical and methodological issues. *Journal of the Learning Sciences*, 13(1):15 – 42, 2004. Cited by: 1090.

[92] Filipa Correia, Joana Campos, Francisco S Melo, and Ana Paiva. Robotic gaze responsiveness in multiparty teamwork. *International Journal of Social Robotics*, 15(1):27–36, 2023.

[93] Filipa Correia, Sofia Petisca, Patrícia Alves-Oliveira, Tiago Ribeiro, Francisco S Melo, and Ana Paiva. "i choose... you!" membership preferences in human–robot teams. *Autonomous Robots*, 43(2):359–373, 2019.

[94] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951.

[95] Caterina Cruciani, Anna Moretti, and Paolo Pellizzari. Dynamic patterns in similarity-based cooperation: An agent-based investigation. *Journal of Economic Interaction and Coordination*, 12(1), 2017.

[96] Mario A. Cypko, Lea Timmermann, Igor M. Sauer, and Claudia Müller-Birn. Towards human-robotic collaboration: Observing teamwork of experienced surgeons in robotic-assisted surgery. In Max Mühlhäuser, Christian Reuter, Bastian Pfleging, Thomas Kosch, Andrii Matviienko, Kathrin Gerling, Sven Mayer, Wilko Heuten, Tanja Döring, Florian Müller, and Martin Schmitz, editors, *MuC '22: Mensch und Computer 2022, Darmstadt Germany, September 4 - 7, 2022*, pages 566–571. ACM, 2022.

[97] Abhinav Dahiya, Alexander M Aroyo, Kerstin Dautenhahn, and Stephen L Smith. A survey of multi-agent human–robot interaction systems. *Robotics and Autonomous Systems*, 161:104335, 2023.

[98] DARPA Strategic Technology Office. ASIMOV: Autonomy standards and ideals with military operational values. DARPA Program Summary, 2025. Accessed: 2025-09-15.

[99] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

[100] Joachim De Greeff, Tina Mioch, Willeke Van Vught, Koen Hindriks, Mark A Neerincx, and Ivana Kruijff-Korbayová. Persistent robot-assisted disaster response. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 99–100, 2018.

[101] Bart A De Jong, Kurt T Dirks, and Nicole Gillespie. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of applied psychology*, 101(8):1134, 2016.

[102] Filippo Santoni de Sio and Jeroen van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers Robotics AI*, 5:15, 2018.

[103] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478, 2020.

[104] Leslie A DeChurch and Jessica R Mesmer-Magnus. Measuring shared team mental models: A meta-analysis. *Group dynamics: Theory, research, and practice*, 14(1):1, 2010.

[105] Edward L Deci and Richard M Ryan. *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media, 2013.

[106] Diede P. M. Van der Hoorn, Anouk Neerincx, and Maartje M. A. de Graaf. "i think you are doing a bad job!": The effect of blame attribution by a robot in human-robot collaboration. In Cindy L. Bethel, Ana Paiva, Elizabeth Broadbent, David Feil-Seifer, and Daniel Szafir, editors, *ACM/IEEE International Conference on Human-Robot Interaction, HRI 2021, Boulder, CO, USA, March 8-11, 2021*, pages 140–148. ACM, 2021.

[107] Neel Dhanaraj, Minseok Jeon, Jeon Ho Kang, Stefanos Nikolaidis, and Satyandra K Gupta. Preference elicitation and incorporation for human-robot task scheduling. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 3103–3110. IEEE, 2024.

[108] Chau Thi Diem Le, Miklos Pakurar, Istvan Andras Kun, and Judit Olah. The impact of factors on information sharing: An application of meta-analysis. *Plos one*, 16(12):e0260653, 2021.

[109] Hasra Dodampegama and Mohan Sridharan. Back to the future: toward a hybrid architecture for ad hoc teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3–10, 2023.

[110] Junliang Du, Sifeng Liu, and Yong Liu. A limited cost consensus approach with fairness concern and its application. *European Journal of Operational Research*, 298(1):261–275, 2022.

[111] Barbara Dunin-Keplicz and Rineke Verbrugge. *Teamwork in multi-agent systems: A formal approach*. John Wiley & Sons, 2011.

[112] Pablo D'Este and Nicolás Robinson-García. Interdisciplinary research and the societal visibility of science: The advantages of spanning multiple and distant scientific fields. *Research Policy*, 52(2):104609, 2023.

[113] Maximilian Eder and Helle Sjøvaag. Artificial intelligence and the dawn of an algorithmic divide. *Frontiers in Communication*, 9:1453251, 2024.

[114] ELSA Lab Defence. ELSA Lab Defence: Ethical, legal, and societal aspects of military ai. `https://elsalabdefence.nl/`, 2025. Accessed: 2025-09-15.

[115] Emre Erdogan, Frank Dignum, Rineke Verbrugge, and Pinar Yolum. Toma: computational theory of mind with abstractions for hybrid intelligence. *Journal of Artificial Intelligence Research*, 82:285–311, 2025.

[116] Mohammadreza Esfandiari, Senjuti Basu Roy, and Sihem Amer-Yahia. Explicit preference elicitation for task completion time. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1233–1242, 2018.

[117] Connor Esterwood, Ruijian Hannah Guan, Xin Ye, and Lionel P Robert. Virtually the same or realistically different?: A meta-analysis of real vs.'not so real'robots. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 559–568. IEEE, 2025.

[118] Connor Esterwood and Lionel P. Robert Jr. Three strikes and you are out!: The impacts of multiple human-robot trust violations and repairs on robot trustworthiness. *Comput. Hum. Behav.*, 142:107658, 2023.

[119] Connor Esterwood and Lionel P Robert. The theory of mind and human-robot trust repair. *Scientific Reports*, 13(1):9877, 2023.

[120] Cedric Faas, Richard Bergs, Sarah Sterz, Markus Langer, and Anna Maria Feit. Give me a choice: The consequences of restricting choices through ai-support for perceived autonomy, motivational variables, and decision performance. *arXiv preprint arXiv:2410.07728*, 2024.

[121] Hannah Fahnenstich, Tobias Rieger, and Eileen Roesler. Trusting under risk–comparing human to ai decision support agents. *Computers in Human Behavior*, page 108107, 2023.

[122] Rino Falcone and Cristiano Castelfranchi. Trust dynamics: How trust is influenced by direct experiences and by trust itself hybrid cognitive architectures for artificial agents view project. 2004.

[123] Rino Falcone, Giovanni Pezzulo, and Cristiano Castelfranchi. A fuzzy approach to a belief-based trust computation. In Rino Falcone, K. Suzanne Barber, Larry Korba, and Munindar P. Singh, editors, *Trust, Reputation, and Security: Theories and Practice, AAMAS 2002 International Workshop, Bologna, Italy, July 15, 2002, Selected and Invited Papers*, volume 2631 of *Lecture Notes in Computer Science*, pages 73–86. Springer, 2002.

[124] Rino Falcone, Giovanni Pezzulo, and Cristiano Castelfranchi. A fuzzy approach to a belief-based trust computation. volume 2631, pages 73–86. Springer Verlag, 2003.

[125] Rino Falcone, Giovanni Pezzulo, and Cristiano Castelfranchi. A fuzzy approach to a belief-based trust computation. In *Trust, Reputation, and Security: Theories and Practice: AAMAS 2002 International Workshop, Bologna, Italy, July 15, 2002. Selected and Invited Papers 5*, pages 73–86. Springer, 2003.

[126] Rino Falcone, Michele Piunti, Matteo Venanzi, and Cristiano Castelfranchi. From manifesta to krypta: The relevance of categories for trusting others. *ACM Trans. Intell. Syst. Technol.*, 4(2):27:1–27:24, 2013.

[127] Rino Falcone, Michele Piunti, Matteo Venanzi, and Cristiano Castelfranchi. From manifesta to krypta: The relevance of categories for trusting others. *ACM Transactions on Intelligent Systems and Technology*, 4, 3 2013.

[128] Alfred J. Farina, George R. Wheaton, and Edwin A. Fleishman. Development of a taxonomy of human performance: The task characteristics approach to performance prediction. 1971.

[129] Shaheen Fatima, Nicholas R Jennings, and Michael Wooldridge. Learning to resolve social dilemmas: A survey. *Journal of Artificial Intelligence Research*, 79:895–969, 2024.

[130] Luis F. C. Figueredo, Rafael Castro Aguiar, Lipeng Chen, Thomas C. Richards, Samit Chakrabarty, and Mehmet Remzi Dogar. Planning to minimize the human muscular effort during forceful human-robot collaboration. *ACM Trans. Hum. Robot Interact.*, 11(1):10:1–10:27, 2022.

[131] Gerhard Fischer. A research framework focused on ai and humans instead of ai versus humans. *Proceedings http://ceur-ws. org ISSN*, 1613:0073, 2022.

[132] Donald L Fisher, William J Horrey, John D Lee, and Michael A Regan. *Handbook of human factors for automated, connected, and intelligent vehicles*. CRC press, 2020.

[133] Susan T Fiske, Amy JC Cuddy, and Peter Glick. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83, 2007.

[134] Paul M Fitts. *Human engineering for an effective air-navigation and traffic-control system*. National Research Council, Div. of, 1951.

[135] Qianrao Fu, Herbert Hoijtink, and Mirjam Moerbeek. Sample-size determination for the bayesian t test and welch's test using the approximate adjusted fractional bayes factor. *Behavior Research Methods*, 53(1):139–152, 2021.

[136] Toshio Fukuda, Paolo Dario, and Guang-Zhong Yang. Humanoid robotics—history, current state of the art, and challenges. *Science Robotics*, 2(13):eaar4043, 2017.

[137] Susan R Fussell, Robert E Kraut, F Javier Lerch, William L Scherlis, Matthew M McNally, and Jonathan J Cadiz. Coordination, overload and team performance: effects of team communication strategies. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 275–284, 1998.

[138] Victor Galaz, Miguel A Centeno, Peter W Callahan, Amar Causevic, Thayer Patterson, Irina Brass, Seth Baum, Darryl Farber, Joern Fischer, David Garcia, et al. Artificial intelligence, systemic risks, and sustainability. *Technology in society*, 67:101741, 2021.

[139] Matthias Gamer, Jim Lemon, and Ian Fellows Puspendra Singh <puspendra.pusp22@gmail.com>. *irr: Various Coefficients of Interrater Reliability and Agreement*, 2012. R package version 0.84.1.

[140] Joseph L. Gastwirth, Yulia R. Gel, Wei Miao, Kayo Noguchi, Wendy L. W. Hui, Vyacheslav Lyubchich, Radhakrishnan Ramachandran, and Min Xu. *lawstat: Tools for Biostatistics, Public Policy, and Law*, 2022. R package version 3.4.

[141] David Gefen, Elena Karahanna, and Detmar W Straub. Trust and tam in online shopping: An integrated model. *MIS quarterly*, pages 51–90, 2003.

[142] Baocheng Geng and Pramod K. Varshney. Human-machine collaboration for smart decision making: Current trends and future opportunities. In *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*, pages 61–67, 2022.

[143] Eleni Georganta and Anna-Sophie Ulfert. My colleague is an ai! trust differences between ai and human teammates. *Team Performance Management: An International Journal*, 30(1/2):23–37, 2024.

[144] Eleni Georganta and Anna-Sophie Ulfert. Would you trust an ai team member? team trust in human–ai teams. *Journal of occupational and organizational psychology*, 97(3):1212–1241, 2024.

[145] Michael P. Georgeff, Barney Pell, Martha E. Pollack, Milind Tambe, and Michael J. Wooldridge. The belief-desire-intention model of agency. In Jörg P. Müller, Munindar P. Singh, and Anand S. Rao, editors, *Intelligent Agents V, Agent Theories, Architectures, and Languages, 5th International Workshop, ATAL '98, Paris, France, July 4-7, 1998, Proceedings*, volume 1555 of *Lecture Notes in Computer Science*, pages 1–10. Springer, 1998.

[146] Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 429–437, 2020.

[147] Vijai N Giri and B Pavan Kumar. Assessing the impact of organizational communication on job satisfaction and job performance. *Psychological Studies*, 55(2):137–143, 2010.

[148] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.

[149] Matthew C. Gombolay, Reymundo A. Gutierrez, Shanelle G. Clarke, Giancarlo F. Sturla, and Julie A. Shah. Decision-making authority, team efficiency and human worker satisfaction in mixed human-robot teams. *Auton. Robots*, 39(3):293–312, 2015.

[150] O Can Görür, Benjamin Rosman, Fikret Sivrikaya, and Sahin Albayrak. Fabric: A framework for the design and evaluation of collaborative robots with extended human adaptation. *ACM Transactions on Human-Robot Interaction*, 12(3):1–54, 2023.

[151] Cedric Goubard and Yiannis Demiris. Cooking up trust: Eye gaze and posture for trust-aware action selection in human-robot collaboration. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, TAS '23, New York, NY, USA, 2023. Association for Computing Machinery.

[152] Andrew Griffin. How amazon is trying to make the world fall in love with its robots. *The Independent*, 2024. Accessed: 2025-09-13.

[153] Nathan Griffiths. Task delegation using experience-based multi-dimensional trust. In Frank Dignum, Virginia Dignum, Sven Koenig, Sarit Kraus, Munindar P. Singh, and Michael J. Wooldridge, editors, *4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), July 25-29, 2005, Utrecht, The Netherlands*, pages 489–496. ACM, 2005.

[154] Rebecca Grossman and Jennifer Feitosa. Team trust over time: Modeling reciprocal and contextual influences in action teams. *Human Resource Management Review*, 28:395–410, 12 2018.

[155] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. Trusting artificial agents: Communication trumps performance. In Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh, editors, *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, pages 299–306. ACM, 2023.

[156] Manju Gundumogula and M Gundumogula. Importance of focus groups in qualitative research. *International Journal of Humanities and Social Science (IJHSS)*, 8(11):299–302, 2020.

[157] Yaohui Guo and X Jessie Yang. Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, 13(8):1899–1909, 2021.

[158] Beth Gutelius and Sanjay Pinto. Pain Points: Data on Work Intensity, Monitoring, and Health at Amazon Warehouses, 10 2023.

[159] Zachary Guyton, Richard Pak, and Ericka Rovira. The role of automation etiquette and task-criticality on performance, workload, automation reliance, and user confidence. *Applied Ergonomics*, 125:104430, 2025.

[160] Andrea L Guzman and Seth C Lewis. Artificial intelligence and communication: A human–machine communication research agenda. *New media & society*, 22(1):70–86, 2020.

[161] J Richard Hackman and Greg R Oldham. Motivation through the design of work: Test of a theory. *Organizational behavior and human performance*, 16(2):250–279, 1976.

[162] Vera Hagemann, Michèle Rieth, Amrita Suresh, and Frank Kirchner. Human-ai teams—challenges for a team-centered ai at work. *Frontiers in Artificial Intelligence*, 6:1252897, 2023.

[163] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart de Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors*, 53(5):517–527, 2011.

[164] Maaike Harbers, Catholijn M. Jonker, and M. Birna van Riemsdijk. Context-sensitive sharedness criteria for teamwork. In Ana L. C. Bazzan, Michael N. Huhns, Alessio Lomuscio, and Paul Scerri, editors, *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, Paris, France, May 5-9, 2014*, pages 1507–1508. IFAAMAS/ACM, 2014.

[165] Maaike Harbers and Mark A. Neerincx. Value sensitive design of a virtual assistant for workload harmonization in teams. *Cogn. Technol. Work.*, 19(2-3):329–343, 2017.

[166] Maaike Harbers, Rineke Verbrugge, Carles Sierra, and John Debenham. The examination of an information-based approach to trust. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 71–82. Springer, 2007.

[167] Kerstin S Haring, Elizabeth Phillips, Elizabeth H Lazzara, Daniel Ullman, Anthony L Baker, and Joseph R Keebler. Applying the swift trust model to human-robot teaming. In *Trust in Human-Robot Interaction*, pages 407–427. Elsevier, 2021.

[168] Bharat Hazari and Vijay Mohan. Exclusion and the growth of ai technology: a trade-theoretic analysis. *Frontiers in Human Dynamics*, 6:1203664, 2024.

[169] Sebastian Herrera. Amazon is on the cusp of using more robots than humans in its warehouses. *The Wall Street Journal*, 2025. Accessed: 2025-09-13.

[170] Sarita Herse, Jonathan Vitale, Benjamin Johnston, and Mary-Anne Williams. Using trust to determine user decision making & task outcome during a human-agent collaborative task. In Cindy L. Bethel, Ana Paiva, Elizabeth Broadbent, David Feil-Seifer, and Daniel Szafir, editors, *ACM/IEEE International Conference on Human-Robot Interaction, HRI 2021, Boulder, CO, USA, March 8-11, 2021*, pages 73–82. ACM, 2021.

[171] Andreas Herzig, Emiliano Lorini, Jomi F. Hübner, and Laurent Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL*, 18:214–244, 12 2009.

[172] Teresa Heyder, Nina Passlack, and Oliver Posegga. Ethical management of human-ai interaction: Theory development review. *The Journal of Strategic Information Systems*, 32(3):101772, 2023.

[173] Steffen Hoesterey and Linda Onnasch. The effect of risk on trust attitude and trust behavior in interaction with information and decision automation. *Cognition, Technology & Work*, 25(1):15–29, 2023.

[174] Guy Hoffman and Xuan Zhao. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(1):1–31, 2020.

[175] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[176] Koen Hogenelst, Roos Schelvis, Tanja Krone, Marylene Gagné, Matti Heino, Keegan Knittle, and Nelli Hankonen. A within-person approach to the relation between quality of task motivation, performance and job satisfaction in everyday working life. *Motivation and Emotion*, 46(5):588–600, 2022.

[177] Tiffany Matej Hrkalovic, Aria Li, Magnus Bopp, Yingling Li, and Daniel Balliet. Task affordances affect partner preferences. *Journal of Experimental Social Psychology*, 119:104751, 2025.

[178] Sheng-Jen Hsieh, Andy R. Wang, Anna Madison, Chad Tossell, and Ewart de Visser. Adaptive driving assistant model (ADAM) for advising drivers of autonomous vehicles. *ACM Trans. Interact. Intell. Syst.*, 12(3):21:1–21:28, 2022.

[179] Lixiao Huang, Nancy J Cooke, Robert S Gutzwiller, Spring Berman, Erin K Chiou, Mustafa Demir, and Wenlong Zhang. Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In *Trust in human-robot interaction*, pages 301–319. Elsevier, 2021.

[180] Xiaowei Huang, Marta Kwiatkowska, and Maciej Olejnik. Reasoning about cognitive trust in stochastic multiagent systems. *ACM Transactions on Computational Logic (TOCL)*, 20(4):1–64, 2019.

[181] Clark Leonard Hull. *Principles of behavior: An introduction to behavior theory*. Appleton-Century-Crofts, 1959.

[182] Human Rights Watch. A hazard to human rights: Autonomous weapons systems and digital decision-making. Human Rights Watch Report, April 28 2025. Accessed: 2025-09-15.

[183] Rehan Iftikhar, Yi-Te Chiu, Mohammad Saud Khan, and Catherine Barbara Caudwell. Human-agent team dynamics: A review and future research opportunities. *IEEE Trans. Engineering Management*, 71:10139–10154, 2024.

[184] Michael Inzlicht, Amitai Shenhav, and Christopher Y Olivola. The effort paradox: Effort is both costly and valued. *Trends in cognitive sciences*, 22(4):337–349, 2018.

[185] Bahar Irfan, Aditi Ramachandran, Mariacarla Staffa, and Hatice Gunes. Lifelong learning and personalization in long-term human-robot interaction (leap-hri) adaptivity for all. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 929–931, 2023.

[186] Mohammad Hossein Jarrahi. Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586, 2018.

[187] Tjalling Haije Jasper van der Waa. Matrx: Human agent teaming rapid experimentation software, July 2023.

[188] Theodore Jensen, Yusuf Albayram, Mohammad M.H. Khan, Ross Buck, Emil Coman, and Md A. Al Fahim. Initial trustworthiness perceptions of a drone system based on performance and process information. In *Proceedings of 6th International Conference on Human-Agent Interaction*, 2018.

[189] Junli Jiang and Pavel Naumov. A logic of trust-based beliefs. *Synthese*, 204(2):46, 2024.

[190] Craig J. Johnson, Mustafa Demir, Nathan J. McNeese, Jamie C. Gorman, Alexandra T. Wolff, and Nancy J. Cooke. The impact of training on human-autonomy team communications and trust calibration. *Hum. Factors*, 65(7):1554–1570, 2023.

[191] Matt Johnson, J. Bradshaw, R. Hoffman, P. Feltovich, and D. Woods. Seven cardinal virtues of human-machine teamwork: Examples from the darpa robotic challenge. *IEEE Intelligent Systems*, 29:74–80, 2014.

[192] Matthew Johnson and Jeffrey M. Bradshaw. Chapter 16 - the role of interdependence in trust. In Chang S. Nam and Joseph B. Lyons, editors, *Trust in Human-Robot Interaction*, pages 379–403. Academic Press, 2021.

[193] Matthew Johnson and Jeffrey M Bradshaw. The role of interdependence in trust. In *Trust in human-robot interaction*, pages 379–403. Elsevier, 2021.

[194] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1):43–69, 2014.

[195] Matthew Johnson, Catholijn M. Jonker, M. Birna van Riemsdijk, Paul J. Feltovich, and Jeffrey M. Bradshaw. Joint activity testbed: Blocks world for teams (BW4T). In Huib Aldewereld, Virginia Dignum, and Gauthier Picard, editors, *Engineering Societies in the Agents World X, 10th International Workshop, ESAW 2009, Utrecht, The Netherlands, November 18-20, 2009. Proceedings*, volume 5881 of *Lecture Notes in Computer Science*, pages 254–256. Springer, 2009.

[196] Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.

[197] Catholijn M Jonker, Luciano Cavalcante Siebert, and Pradeep K Murukannaiah. Reflective hybrid intelligence for meaningful human control in decision-support systems. In *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, pages 188–204. Edward Elgar Publishing, 2024.

[198] Catholijn M. Jonker, M. Birna van Riemsdijk, Iris van de Kieft, and Maria L. Gini. Compositionality of team mental models in relation to sharedness and team performance. In He Jiang, Wei Ding, Moonis Ali, and Xindong Wu, editors, *Advanced Research in Applied Artificial Intelligence - 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2012, Dalian, China, June 9-12, 2012. Proceedings*, volume 7345 of *Lecture Notes in Computer Science*, pages 242–251. Springer, 2012.

[199] Catholijn M Jonker, M Birna Van Riemsdijk, and Bas Vermeulen. Shared mental models: A conceptual analysis. In *International workshop on coordination, organizations, institutions, and norms in agent systems*, pages 132–151. Springer, 2010.

[200] Carolina Centeio Jorge, Emma M van Zoelen, Ruben Verhagen, Siddharth Mehrotra, Catholijn M Jonker, and Myrthe L Tielman. Appropriate context-dependent artificial trust in human-machine teamwork. In *Putting AI in the Critical Loop*, pages 41–60. Elsevier, 2024.

[201] Lionel P. Robert Jr., Alan R. Dennis, and Yu-Ting Caisy Hung. Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *J. Manag. Inf. Syst.*, 26(2), 2009.

[202] Julakha Jahan Jui, Imali T Hettiarachchi, Asim Bhatti, Mohamed Ragab Mahmoud Farghaly, and Douglas Creighton. A recent review on subjective and objective assessment of trust in human autonomy teaming. *IEEE Transactions on Human-Machine Systems*, 2025.

[203] Malte F Jung, Dominic DiFranzo, Solace Shen, Brett Stoll, Houston Claure, and Austin Lawrence. Robot-assisted tower construction—a method to study the impact of a robot's allocation behavior on interpersonal dynamics and collaboration in groups. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(1):1–23, 2020.

[204] Eija Kaasinen, Anu-Hanna Anttila, Päivi Heikkilä, Jari Laarni, Hanna Koskinen, and Antti Väätänen. Smooth and resilient human–machine teamwork as an industry 5.0 design challenge. *Sustainability*, 14(5):2773, 2022.

[205] Monika Kaczorowska, Paweł Karczmarek, Małgorzata Plechawska-Wójcik, and Mikhail Tokovarov. On the improvement of eye tracking-based cognitive workload estimation using aggregation functions. *Sensors*, 21(13):4542, 2021.

[206] Heather M Kharouba. Now is the time for academics to think and act beyond academia, 2024.

[207] Aman Khullar, Nikhil Nalin, Abhishek Prasad, Ann John Mampilli, and Neha Kumar. *Nurturing Capabilities: Unpacking the Gap in Human-Centered Evaluations of AI-Based Systems*. Association for Computing Machinery, New York, NY, USA, 2025.

[208] Noona Kiuru, Birgit Spinath, Anna-Leena Clem, Kenneth Eklund, Timo Ahonen, and Riikka Hirvonen. The dynamics of motivation, emotion, and task performance in simulated achievement situations. *Learning and Individual Differences*, 80:101873, 2020.

[209] Glen Klien, David D Woods, Jeffrey M Bradshaw, Robert R Hoffman, and Paul J Feltovich. Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95, 2004.

[210] Bran Knowles, Jason D'Cruz, John T. Richards, and Kush R. Varshney. Humble ai. *Commun. ACM*, 66(9):73–79, August 2023.

[211] Bing Cai Kok and Harold Soh. Trust in robots: Challenges and opportunities. *Current Robotics Reports*, 1(4):297–309, 2020.

[212] Wouter Kool, Joseph T McGuire, Zev B Rosen, and Matthew M Botvinick. Decision making and the avoidance of cognitive demand. *Journal of experimental psychology: general*, 139(4):665, 2010.

[213] Jonas Koreis. Human–robot vs. human–manual teams: Understanding the dynamics of experience and performance variability in picker-to-parts order picking. *Computers & Industrial Engineering*, 200:110750, 2025.

[214] Angelos Kostis, Maria Bengtsson, and Malin H Näsholm. Mechanisms and dynamics in the interplay of trust and distrust: Insights from project-based collaboration. *Organization Studies*, 43(8):1173–1196, 2022.

[215] E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. de Vries. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems*, 35(2):1–20, 2021.

[216] Esther S Kox, Milou Hennekens, Jason S Metcalfe, and José H Kerstholt. Trust violations due to error or choice: The differential effects on trust repair in human-human and human-robot interaction. *ACM Transactions on Human-Robot Interaction*, 2025.

[217] Richard A Krueger and Mary Anne Casey. *Designing and conducting focus group interviews*, volume 18. Citeseer, 2002.

[218] John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.

[219] Lars Kunze, Nick Hawes, Tom Duckett, Marc Hanheide, and Tomáš Krajník. Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters*, 3(4):4023–4030, 2018.

[220] Takashi Kuremoto, Tetsuya Tsurusaki, Kunikazu Kobayashi, Shingo Mabu, and Masanao Obayashi. An improved reinforcement learning system using affective factors. *Robotics*, 2(3):149–164, 2013.

[221] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[222] Benjamin H Lang, Sven Nyholm, and Jennifer Blumenthal-Barby. Responsibility gaps and black box healthcare ai: shared responsibilization as a solution. *Digital Society*, 2(3):52, 2023.

[223] Claus W Langfred. Too much of a good thing? negative effects of high trust and individual autonomy in self-managing teams. *Academy of management journal*, 47(3):385–399, 2004.

[224] Daniella Laureiro-Martínez, Stefano Brusoni, and Maurizio Zollo. The neuroscientific foundations of the exploration- exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics*, 3(2):95, 2010.

[225] Theresa Law and Matthias Scheutz. Trust: Recent concepts and evaluations in human-robot interaction. *Trust in human-robot interaction*, pages 27–57, 2021.

[226] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. Trusting artificial agents: Communication trumps performance. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 299–306, 2023.

[227] J. D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 46:50 – 80, 2004.

[228] Michael D Lee, Shunan Zhang, Miles Munro, and Mark Steyvers. Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, 12(2):164–174, 2011.

[229] Sun Kyong Lee and Juhyung Sun. Testing a theoretical model of trust in human-machine communication: emotional experience and social presence. *Behaviour & Information Technology*, 42(16):2754–2767, 2023.

[230] Won-Hyong Lee and Jong-Hwan Kim. Hierarchical emotional episodic memory for social human robot collaboration. *Autonomous Robots*, 42(5):1087–1102, 2018.

[231] Michael Lewis, Huao Li, and Katia Sycara. Deep learning, transparency, and trust in human robot teamwork. In *Trust in Human-Robot Interaction*, pages 321–352. Elsevier, 2020.

[232] Michael Lewis, Katia Sycara, and Phillip Walker. *The Role of Trust in Human-Robot Interaction*. Springer International Publishing, 2018.

[233] Chunxia Li, Yunmei Ding, Dongdong Wang, and Chuanyao Deng. The effect of nurses' perceived social support on job burnout: The mediating role of psychological detachment. *Journal of Advanced Nursing*, 2025.

[234] Jamy Li. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37, 2015.

[235] Min Li, Xiaoxun Sun, Hua Wang, and Yanchun Zhang. Multi-level delegations with trust management in access control systems. *Journal of Intelligent Information Systems*, 39(3):611–626, 2012.

[236] Clare Lohrmann, Maria P. Stull, Alessandro Roncone, and Bradley Hayes. Generating pattern-based conventions for predictable planning in human-robot collaboration. *ACM Trans. Hum. Robot Interact.*, 13(4):53:1–53:23, 2024.

[237] Paul Benjamin Lowry, Ryan M Schuetzler, Justin Scott Giboney, and Thomas A Gregory. Is trust always better than distrust? the potential value of distrust in newer virtual teams engaged in short-term decision-making. *Group Decision and Negotiation*, 24(4):723–752, 2015.

[238] Michael Luck and Mark D'Inverno. Engagement and cooperation in motivated agent modelling. In Chengqi Zhang and Dickson Lukose, editors, *Distributed Artificial Intelligence Architecture and Modelling*, pages 70–84, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg.

[239] Matthew B. Luebbers, Aaquib Tabrez, Kyler Ruvane, and Bradley Hayes. Autonomous justification for enabling explainable decision support in human-robot teaming. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu, editors, *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023.

[240] Lanssie Mingyue Ma, Martijn Ijtsma, Karen M Feigh, and Amy R Pritchett. Metrics for human-robot team design: A teamwork perspective on evaluation of human-robot teams. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(3):1–36, 2022.

[241] Patrick Mair and Rand Wilcox. Robust statistical methods using wrs2. *The WRS2 package*, 2019.

[242] Ilias El Makrini, Kelly Merckaert, Joris De Winter, Dirk Lefeber, and Bram Vanderborght. Task allocation for improved ergonomics in human-robot collaborative assembly. *Interaction Studies*, 20(1):102–133, 2019.

[243] Ali Ahmad Malik and Arne Bilberg. Complexity-based task allocation in human-robot collaborative assembly. *Industrial Robot: the international journal of robotics research and application*, 46(4):471–480, 2019.

[244] Hsiao-Yen Mao, An-Tien Hsieh, and Chien-Yu Chen. The relationship between workplace friendship and perceived job significance. *Journal of Management & Organization*, 18(2):247–262, 2012.

[245] Marcos Maroto-Gómez, Álvaro Castro-González, José Carlos Castillo, María Malfaz, and Miguel Ángel Salichs. An adaptive decision-making system supported on user preference predictions for human–robot interactive communication. *User Modeling and User-Adapted Interaction*, 33(2):359–403, 2023.

[246] Magoroh Maruyama. Managerial and administrative accidents. *Management Dynamics*, 6(1):133–137, 2006.

[247] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2):273, 2000.

[248] Juliette Mattioli, Gabriel Pedroza, Souhaiel Khalfaoui, and Bertrand Leroy. Combining data-driven and knowledge-based ai paradigms for engineering ai-based safety-critical systems. In *Workshop on Artificial Intelligence Safety (SafeAI)*, 2022.

[249] Roger C Mayer and James H Davis. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology*, 84(1):123, 1999.

[250] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Source: The Academy of Management Review*, 20:709–734, 1995.

[251] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2020.

[252] Joseph Edward McGrath. *Groups: Interaction and performance*, volume 14. Prentice-Hall Englewood Cliffs, NJ, 1984.

[253] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[254] Kevin R. McKee, Xuechunzi Bai, and Susan T. Fiske. Warmth and competence in human-agent cooperation. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 898–907, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems.

[255] D. Harrison McKnight, Vivek Choudhury, and Charles J. Kacmar. Developing and validating trust measures for e-commerce: An integrative typology. *Inf. Syst. Res.*, 13(3):334–359, 2002.

[256] Nathan J McNeese, Mustafa Demir, Erin K Chiou, and Nancy J Cooke. Trust and team performance in human–autonomy teaming. *International Journal of Electronic Commerce*, 25(1):51–72, 2021.

[257] Kieron J Meagher and Andrew Wait. Worker trust in management and delegation in organizations. *The Journal of Law, Economics, and Organization*, 36(3):495–536, 2020.

[258] Giulio Mecacci and Filippo Santoni de Sio. Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics Inf. Technol.*, 22(2):103–115, 2020.

[259] Malek Mechergui and Sarath Sreedharan. Goal alignment: Re-analyzing value alignment problems using human-aware AI. In Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh, editors, *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, pages 2331–2333. ACM, 2023.

[260] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. Integrity-based explanations for fostering appropriate trust in ai agents. *ACM Transactions on Interactive Intelligent Systems*, 14(1):1–36, 2024.

[261] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM J. Responsib. Comput.*, 1(4), November 2024.

[262] Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. More similar values, more trust? - the effect of value similarity on trust in human-agent interaction. In Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan, editors,

AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021, pages 777–783. ACM, 2021.

[263] Elena Merlo, Edoardo Lamon, Fabio Fusaro, Marta Lorenzini, Alessandro Carfì, Fulvio Mastrogiovanni, and Arash Ajoudani. An ergonomic role allocation framework for dynamic human–robot collaborative tasks. *Journal of Manufacturing Systems*, 67:111–121, 2023.

[264] Susan Michie, Maartje M Van Stralen, and Robert West. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science*, 6(1):1–12, 2011.

[265] Suzanne P. Mikawa, Sharon K. Cunnington, and Scott A. Gaskins. Removing barriers to trust in distributed teams: understanding cultural differences and strengthening social ties. In Susan R. Fussell, Pamela J. Hinds, and Toru Ishida, editors, *Proceedings of the 2009 international workshop on Intercultural collaboration, IWIC '09, Palo Alto, California, USA, February 20-21, 2009*, pages 273–276. ACM, 2009.

[266] Abdulla M. Mohamed and M. Huhns. Multiagent benevolence as a societal norm. 2001.

[267] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. Metaphor no more: A 15-year review of the team mental model construct. *Journal of management*, 36(4):876–910, 2010.

[268] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Trans. Interact. Intell. Syst.*, 11(3–4), September 2021.

[269] Moley Robotics. X-AiR Kitchen. `https://www.moley.com/x-air-kitchen/`, 2025. Accessed June 2025.

[270] David L Morgan. *Focus groups as qualitative research*, volume 16. Sage publications, 1996.

[271] David L Morgan, Jutta Ataie, Paula Carder, and Kim Hoffman. Introducing dyadic interviews as a method for collecting qualitative data. *Qualitative health research*, 23(9):1276–1284, 2013.

[272] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

[273] Tanja Müller, Masud Husain, and Matthew AJ Apps. Preferences for seeking effort or reward information bias the willingness to work. *Scientific Reports*, 12(1):19486, 2022.

[274] Michael J Munje, Lylybell K Teran, Bradon Thymes, and Joseph P Salisbury. Team3 challenge: Tasks for multi-human and multi-robot collaboration with voice and gestures. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 91–96, 2023.

[275] Robin R Murphy. *Introduction to AI robotics*. MIT press, 2019.

[276] Changjoo Nam, Huao Li, Shen Li, Michael Lewis, and Katia Sycara. Trust of humans in supervisory control of swarm robots with varied levels of autonomy. pages 825–830. Institute of Electrical and Electronics Engineers Inc., 1 2019.

[277] Changjoo Nam, Phillip Walker, Huao Li, Michael Lewis, and Katia Sycara. Models of trust in human control of swarms with varied levels of autonomy. *IEEE Transactions on Human-Machine Systems*, 50:194–204, 6 2020.

[278] C. Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56:81–103, 2000.

[279] Manisha Natarajan and Matthew C. Gombolay. Effects of anthropomorphism and accountability on trust in human robot interaction. In Tony Belpaeme, James E. Young, Hatice Gunes, and Laurel D. Riek, editors, *HRI '20: ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, United Kingdom, March 23-26, 2020*, pages 33–42. ACM, 2020.

[280] Manisha Natarajan, Esmaeil Seraj, Batuhan Altundas, Rohan Paleja, Sean Ye, Letian Chen, Reed Jensen, Kimberlee Chestnut Chang, and Matthew Gombolay. Human-robot teaming: grand challenges. *Current Robotics Reports*, 4(3):81–100, 2023.

[281] Mark A Neerincx et al. Cognitive task load analysis: allocating tasks and designing support. *Handbook of cognitive task design*, 2003:283–305, 2003.

[282] Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *International conference on engineering psychology and cognitive ergonomics*, pages 204–214. Springer, 2018.

[283] Federica Nenna, Davide Zanardi, Egle Maria Orlando, Michele Mingardi, Giulia Buodo, and Luciano Gamberini. Addressing trust and negative attitudes toward robots in human-robot collaborative scenarios: Insights from the industrial work setting. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2024, Crete, Greece, June 26-28, 2024*. ACM, 2024.

[284] Joshua Newman. Promoting interdisciplinary research collaboration: A systematic review, a critical literature review, and a pathway forward. *Social Epistemology*, 38(2):135–151, 2024.

[285] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha S. Srinivasa. Formalizing human-robot mutual adaptation: A bounded memory model. In Christoph Bartneck, Yukie Nagai, Ana Paiva, and Selma Sabanovic, editors, *The Eleventh ACM/IEEE International Conference on Human Robot Interation, HRI 2016, Christchurch, New Zealand, March 7-10, 2016*, pages 75–82. IEEE/ACM, 2016.

[286] Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha S. Srinivasa. Planning with verbal communication for human-robot collaboration. *ACM Trans. Hum. Robot Interact.*, 7(3):22:1–22:21, 2018.

[287] Nikolaos Nikolakis, Kostantinos Sipsas, Panagiota Tsarouchi, and Sotirios Makris. On a shared human-robot task scheduling and online re-scheduling. *Procedia CIRP*, 78:237–242, 2018.

[288] Ali Noormohammadi-Asl, Ali Ayub, Stephen L Smith, and Kerstin Dautenhahn. Task selection and planning in human-robot collaborative processes: To be a leader or a follower? In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1244–1251. IEEE, 2022.

[289] Ali Noormohammadi-Asl, Ali Ayub, Stephen L. Smith, and Kerstin Dautenhahn. Adapting to human preferences to lead or follow in human-robot collaboration: A system evaluation. In *32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, August 28-31, 2023*, pages 1851–1858. IEEE, 2023.

[290] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. Accountability in artificial intelligence: What it is and how it works. *Ai & Society*, 39(4):1871–1882, 2024.

[291] Shofiyati Nur Karimah, Teruhiko Unoki, and Shinobu Hasegawa. Implementation of long short-term memory (lstm) models for engagement estimation in online learning. In *2021 IEEE International Conference on Engineering, Technology Education (TALE)*, pages 283–289, 2021.

[292] AM de Oliveira, Marcus Tolentino Silva, Tais Freire Galvao, and Luciane Cruz Lopes. The relationship between job satisfaction, burnout syndrome and depressive symptoms: An analysis of professionals in a teaching hospital in brazil. *Medicine*, 97(49):e13364–e13364, 2018.

[293] James Onken, Reid Hastie, and William Revelle. Individual differences in the use of simplification strategies in a complex decision-making task. *Journal of Experimental Psychology: Human Perception and Performance*, 11(1):14, 1985.

[294] Scott Ososky, David Schuster, Elizabeth Phillips, and F. Jentsch. Building appropriate trust in human-robot teams. In *AAAI Spring Symposium: Trust and Autonomous Systems*, 2013.

[295] Julia Ostheimer, Soumitra Chowdhury, and Sarfraz Iqbal. An alliance of humans and machines for machine learning: Hybrid intelligent systems and their design principles. *Technology in Society*, 66:101647, 2021.

[296] Katherine O'Toole and Emőke-Ágnes Horvát. Extending human creativity with ai. *Journal of Creativity*, 34(2):100080, 2024.

[297] Thomas O'neill, Nathan McNeese, Amy Barron, and Beau Schelble. Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors*, 64(5):904–938, 2022.

[298] Fabio Paglieri, Cristiano Castelfranchi, Célia Costa Pereira, Rino Falcone, Andrea Tettamanzi, and Serena Villata. Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation. *Comput. Math. Organ. Theory*, 20(2):176–194, June 2014.

[299] Michael E Palanski and Francis J Yammarino. Integrity and leadership:: clearing the conceptual confusion. *European Management Journal*, 25(3):171–184, 2007.

[300] Priyam Parashar, Lindsay M. Sanneman, Julie A. Shah, and Henrik I. Christensen. A taxonomy for characterizing modes of interactions in goal-driven, human-robot teams. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, Macau, SAR, China, November 3-8, 2019*, pages 2213–2220. IEEE, 2019.

[301] Raja Parasuraman and Christopher A Miller. Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, 47(4):51–55, 2004.

[302] Christoph Petzoldt, Dario Niermann, Emily Maack, Marius Sontopski, Burak Vur, and Michael Freitag. Implementation and evaluation of dynamic task allocation for human–robot collaboration in assembly. *Applied Sciences*, 12(24):12645, 2022.

[303] Rosalind W. Picard. *Affective Computing*. The MIT Press, 09 1997.

[304] Karen Putzeys and Bram De Wever. "pick a partner!": A qualitative study on university students' motives to choose a specific partner for collaboration. In *Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning-CSCL 2023, pp. 376-377*. International Society of the Learning Sciences, 2023.

[305] Python Software Foundation. *Python Language Reference*, 2024. Available at `https://www.python.org`.

[306] Daniel S Quintana and Donald R Williams. Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using jasp. *BMC psychiatry*, 18(1):1–8, 2018.

[307] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018.

[308] Ram Rachum, Yonatan Nakar, and Reuth Mirsky. Stubborn: An environment for evaluating stubbornness between agents with aligned incentives. *CoRR*, abs/2304.12280, 2023.

[309] Robin R Radtke, Roland F Speklé, and Sally K Widener. Flourish or flounder: Do trust-centric management controls encourage knowledge sharing and team performance? *Accounting, Organizations and Society*, 107:101429, 2023.

[310] Kartik Ramachandruni, Cassandra Kent, and Sonia Chernova. UHTP: A user-aware hierarchical task planning framework for communication-free, mutually-adaptive human-robot collaboration. *ACM Trans. Hum. Robot Interact.*, 13(3):44:1–44:27, 2024.

[311] Anand Srinivasa Rao and M. Georgeff. Bdi agents: From theory to practice. In *ICMAS*, 1995.

[312] Minglun Ren, Nengying Chen, and Hui Qiu. Human-machine collaborative decision-making: An evolutionary roadmap based on cognitive intelligence. *International Journal of Social Robotics*, 15(7):1101–1114, 2023.

[313] Carl Orge Retzlaff, Srijita Das, Christabel Wayllace, Payam Mousavi, Mohammad Afshari, Tianpei Yang, Anna Saranti, Alessa Angerschmid, Matthew E Taylor, and Andreas Holzinger. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. *Journal of Artificial Intelligence Research*, 79:359–415, 2024.

[314] Zahra Rezaei Khavas, Monish Reddy Kotturu, S Reza Ahmadzadeh, and Paul Robinette. Do humans trust robots that violate moral trust? *ACM Transactions on Human-Robot Interaction*, 13(2):1–30, 2024.

[315] Elaine Rich, Kevin Knight, and Shivashankar B Nair. Artificial intelligence, 2010.

[316] Dale Richards, Ian Griffiths, Kelvin Yeung, and Jennifer Cowell-Butler. Designing for human-machine teams: A methodological enquiry. In *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*, pages 1–4. IEEE, 2022.

[317] Virginia Richardson. The role of attitudes and beliefs in learning to teach. *Handbook of research on teacher education/Macmillan*, 1996.

[318] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. Understanding the role of explanation modality in ai-assisted decision-making. In Alejandro Bellogín, Ludovico Boratto, Olga C. Santos, Liliana Ardissono, and Bart P. Knijnenburg, editors, *UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4 - 7, 2022*, pages 223–233. ACM, 2022.

[319] Rowena Rodrigues. Legal and human rights issues of ai: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, 4:100005, 2020.

[320] Alessandro Roncone, Olivier Mangin, and Brian Scassellati. Transparent role assignment and task allocation in human robot collaboration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1014–1021. IEEE, 2017.

[321] William B Rouse and Nancy M Morris. On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, 100(3):349, 1986.

[322] Ellen Rusman, Jan Van Bruggen, Peter Sloep, and Rob Koper. Fostering trust in virtual project teams: Towards a design framework grounded in a trustworthiness antecedents (twan) schema. *International Journal of Human-Computer Studies*, 68(11):834–850, 2010.

[323] SJ Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 4th edition, 2021.

[324] Mark Ryan. In ai we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5):2749–2767, 2020.

[325] Richard M Ryan and Edward L Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67, 2000.

[326] Elie Saad, Koen V. Hindriks, and Mark A. Neerincx. Ontology design for task allocation and management in urban search and rescue missions. In Ana Paula Rocha and H. Jaap van den Herik, editors, *Proceedings of the 10th International Conference on Agents and Artificial Intelligence, ICAART 2018, Volume 2, Funchal, Madeira, Portugal, January 16-18, 2018*, pages 622–629. SciTePress, 2018.

[327] Jordi Sabater and Carles Sierra. Review on computational trust and reputation models. *Artif. Intell. Rev.*, 24(1):33–60, 2005.

[328] Jordi Sabater-Mir and Laurent Vercouter. Trust and reputation in multiagent systems. *Multiagent systems*, page 381, 2013.

[329] Eduardo Salas, Dana E Sims, and C Shawn Burke. Is there a "big five" in teamwork? *Small group research*, 36(5):555–599, 2005.

[330] Raquel Salcedo Gil, Anna-Sophie Ulfert, Sonja Rispens, and Pascale Le Blanc. Optimizing training for human-robot collaboration in learning factories: An employee-centered perspective. In Sebastian Thiede and Eric Lutters, editors, *Learning Factories of the Future*, pages 258–265, Cham, 2024. Springer Nature Switzerland.

[331] Filippo Santoni de Sio and Jeroen Van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:323836, 2018.

[332] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 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, 2016.

[333] Kristin E. Schaefer, Brandon S. Perelman, Gregory M. Gremillion, Amar R. Marathe, and Jason S. Metcalfe. A roadmap for developing team trust metrics for human-autonomy teams. In *Trust in Human-Robot Interaction*. Academic Press, 2021.

[334] Paul Scharre. *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton & Company, New York, 2018.

[335] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. Let's think together! assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–29, 2022.

[336] Matthias Scheutz, Scott A DeLoach, and Julie A Adams. A framework for developing and using shared mental models in human-agent teams. *Journal of Cognitive Engineering and Decision Making*, 11(3):203–224, 2017.

[337] Johannes Schiebener and Matthias Brand. Self-reported strategies in decisions under risk: role of feedback, reasoning abilities, executive functions, short-term-memory, and working memory. *Cognitive processing*, 16(4):401–416, 2015.

[338] Nadine Schlicker, Kevin Baum, Alarith Uhde, Sarah Sterz, Martin C Hirsch, and Markus Langer. How do we assess the trustworthiness of ai? introducing the trustworthiness assessment model (tram). *Computers in Human Behavior*, 170:108671, 2025.

[339] Analia Schlosser, Zvika Neeman, and Yigal Attali. Differential performance in high versus low stakes tests: Evidence from the gre test*. *The Economic Journal*, 129(623):2916–2948, 05 2019.

[340] Tjeerd A. J. Schoonderwoerd, Emma M. van Zoelen, Karel van den Bosch, and Mark A. Neerincx. Design patterns for human-ai co-learning: A wizard-of-oz evaluation in an urban-search-and-rescue task. *Int. J. Hum. Comput. Stud.*, 164:102831, 2022.

[341] F. Schoorman, Roger Mayer, and J. Davis. An integrative model of organizational trust: Past, present, and future. *Academy of Management Review*, 32:344–354, 2007.

[342] Isabella Seeber, Eva Bittner, Robert O Briggs, Triparna De Vreede, Gert-Jan De Vreede, Aaron Elkins, Ronald Maier, Alexander B Merz, Sarah Oeste-Reiß, Nils Randrup, et al. Machines as teammates: A research agenda on ai in team collaboration. *Information & management*, 57(2):103174, 2020.

[343] Isabella Seeber, Lena Waizenegger, Stefan Seidel, Stefan Morana, Izak Benbasat, and Paul Benjamin Lowry. Collaborating with technology-based autonomous agents: Issues and research opportunities. *Internet Research*, 30(1):1–18, 2020.

[344] Andreas Seidler, Marleen Thinschmidt, Stefanie Deckert, Francisca Then, Janice Hegewald, Karen Nieuwenhuijsen, and Steffi G Riedel-Heller. The role of psychosocial working conditions on burnout and its core component emotional exhaustion–a systematic review. *Journal of occupational medicine and toxicology*, 9(1):1–13, 2014.

[345] Francesco Semeraro, Alexander Griffiths, and Angelo Cangelosi. Human–robot collaboration and machine learning: A systematic review of recent research. *Robotics and Computer-Integrated Manufacturing*, 79:102432, 2023.

[346] Rajendra Kumar Shah. Teachers' belief: an overview. *International Journal of Creative Research Thoughts (IJCRT)*, 9(1):3890–3909, 2021.

[347] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[348] Thomas B Sheridan. Extending three existing models to analysis of trust in automation: signal detection, statistical parameter estimation, and model-based control. *Human factors*, 61(7):1162–1170, 2019.

[349] Aron Wolf Siegel and Jan Maarten Schraagen. Team reflection makes resilience-related knowledge explicit through collaborative sensemaking: observation study at a rail post. *Cognition, technology & work*, 19(1):127–142, 2017.

[350] Dominik Siemon. Elaborating team roles for artificial intelligence-based teammates in human-ai collaboration. *Group Decision and Negotiation*, pages 1–42, 2022.

[351] Maarten Sierhuis, Jeffrey Bradshaw, Alessandro Acquisti, Ron van Hoof, Renia Jeffers, and Andrzej Uszok. Human-agent teamwork and adjustable autonomy in practice. In *Proceeding of the 7th International Symposium on Artificial Intelligence, Robotics and Automation in Space*, pages 1–8, 2003.

[352] P. W. Singer. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. Penguin Press, New York, 2009.

[353] Michael L Slepian and Daniel R Ames. Internalized impressions: The link between apparent facial trustworthiness and deceptive behavior is mediated by targets' expectations of how they will be judged. *Psychological science*, 27(2):282–288, 2016.

[354] Enoch Oluwademilade Sodiya, Uchenna Joseph Umoga, Olukunle Oladipupo Amoo, and Akoh Atadoga. Ai-driven warehouse automation: A comprehensive review of systems. *GSC Advanced Research and Reviews*, 18(2):272–282, 2024.

[355] Richard L Solomon. The influence of work on behavior. *Psychological Bulletin*, 45(1), 1948.

[356] Kavyaa Somasundaram, Andrey Kiselev, and Amy Loutfi. Intelligent disobedience: A novel approach for preventing human induced interaction failures in robot teleoperation. In Ginevra Castellano, Laurel D. Riek, Maya Cakmak, and Iolanda Leite, editors, *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2023, Stockholm, Sweden, March 13-16, 2023*, pages 142–145. ACM, 2023.

[357] Jinzhu Song and Hengyu Lin. Exploring the effect of artificial intelligence intellect on consumer decision delegation: The role of trust, task objectivity, and anthropomorphism. *Journal of Consumer Behaviour*, 23(2):727–747, 2024.

[358] Yao Song and Yan Luximon. When trustworthiness meets face: Facial design for social robots. *Sensors*, 24(13):4215, 2024.

[359] D. Sandy Staples and Pauline Ratnasingham. Trust: the panacea of virtual management? In Janice I. DeGross, Rudy Hirschheim, and Michael Newman, editors, *Proceedings of the Nineteenth International Conference on Information Systems, ICIS 1998, Helsinki, Finland, December 13-16, 1998*, pages 128–144. Association for Information Systems, 1998.

[360] Marc J Stern and Kimberly J Coleman. The multidimensionality of trust: Applications in collaborative natural resource management. *Society & Natural Resources*, 28(2):117–132, 2015.

[361] Rachel E Stuck, Brittany E Holthausen, and Bruce N Walker. The role of risk in human-robot trust. In *Trust in human-robot interaction*, pages 179–194. Elsevier, 2021.

[362] Rachel E Stuck, Brianna J Tomlinson, and Bruce N Walker. The importance of incorporating risk into human-automation trust. *Theoretical Issues in Ergonomics Science*, 23(4):500–516, 2022.

[363] Vidullan Surendran and A. Wagner. Your robot is watching: Using surface cues to evaluate the trustworthiness of human actions. *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–8, 2019.

[364] Vidullan Surendran and Alan R. Wagner. Your robot is watching: Using surface cues to evaluate the trustworthiness of human actions. In *28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019, New Delhi, India, October 14-18, 2019*, pages 1–8. IEEE, 2019.

[365] John Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier, 2011.

[366] Bence Márk Szeszák, István Gergely Kerékjártó, László Soltész, and Péter Galambos. Industrial revolutions and automation: Tracing economic and social transformations of manufacturing. *Societies*, 15(4):88, 2025.

[367] Joanna Szulc and Nigel King. The practice of dyadic interviewing: Strengths, limitations and key decisions. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, volume 23. DEU, 2022.

[368] Aaquib Tabrez. Autonomous policy explanations for effective human-machine teaming. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 23423–23424. AAAI Press, 2024.

[369] Maxime Taquet, Jordi Quoidbach, Yves-Alexandre De Montjoye, Martin Desseilles, and James J Gross. Hedonism and the choice of everyday activities. *Proceedings of the national Academy of Sciences*, 113(35):9769–9773, 2016.

[370] Alina Tausch and Annette Kluge. The best task allocation process is to decide on one's own: effects of the allocation agent in human–robot interaction on perceived work characteristics and satisfaction. *Cognition, Technology & Work*, 24(1):39–55, 2022.

[371] Alina Tausch, Corinna Peifer, Britta Marleen Kirchhoff, and Annette Kluge. Human–robot interaction: how worker influence in task allocation improves autonomy. *Ergonomics*, 65(9):1230–1244, 2022.

[372] Safety & Security TNO Defence. Flyers – defence programmes 2022. Technical Report TNO 2022 P11443, TNO - Netherlands Organisation for Applied Scientific Research, July 2021. Accessed: 2025-09-15.

[373] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. Taxonomy of trust-relevant failures and mitigation strategies. In Tony Belpaeme, James E. Young, Hatice Gunes, and Laurel D. Riek, editors, *HRI '20: ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, United Kingdom, March 23-26, 2020*, pages 3–12. ACM, 2020.

[374] Patrick Tucker. Defense department budget request goes hard on ai, autonomy. Defense One, July 1 2025. Accessed: 2025-09-15.

[375] Anna-Sophie Ulfert and Eleni Georganta. A model of team trust in human-agent teams. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 171–176, New York, NY, USA, 2020. Association for Computing Machinery.

[376] Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, Siddharth Mehrotra, and Myrthe Tielman. Shaping a multidisciplinary understanding of team trust in human-ai teams: a theoretical framework. *European Journal of Work and Organizational Psychology*, pages 1–14, 2023.

[377] Daniel Ullman and Bertram F. Malle. Measuring gains and losses in human-robot trust: evidence for differentiable components of trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19, page 618–619. IEEE Press, 2020.

[378] Vaibhav V. Unhelkar, Shen Li, and Julie A. Shah. Decision-making for bidirectional communication in sequential human-robot collaborative tasks. In Tony Belpaeme, James E. Young, Hatice Gunes, and Laurel D. Riek, editors, *HRI '20: ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, United Kingdom, March 23-26, 2020*, pages 329–341. ACM, 2020.

[379] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. The impact of benevolence in computational trust. In *Agreement Technologies*, pages 210–224. Springer, 2013.

[380] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. A socio-cognitive perspective of trust. In *Agreement Technologies*, pages 419–429. Springer, 2013.

[381] Joana Urbano, Ana Paula Rocha, and Eugénio C. Oliveira. A dynamic agents' behavior model for computational trust. In Luis Antunes and Helena Sofia Pinto, editors, *Progress in Artificial Intelligence, 15th Portuguese Conference on Artificial Intelligence, EPIA 2011, Lisbon, Portugal, October 10-13, 2011. Proceedings*, volume 7026 of *Lecture Notes in Computer Science*, pages 536–550. Springer, 2011.

[382] Alena Valtonen, Minna Saunila, Juhani Ukko, Luke Treves, and Paavo Ritala. Ai and employee wellbeing in the workplace: An empirical study. *Journal of Business Research*, 199:115584, 2025.

[383] Iris van de Kieft, Catholijn M Jonker, and M Birna van Riemsdijk. Shared mental models for decision support systems and their users. In *3rd International Workshop on Collaborative Agents-REsearch and development (CARE 2011)*, pages 54–63, 2011.

[384] Karel van den Bosch, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx. Six challenges for human-ai co-learning. In *Adaptive Instructional Systems: First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*, pages 572–589. Springer, 2019.

[385] Piet Van den Bossche, Wim Gijselaers, Mien Segers, Geert Woltjer, and Paul Kirschner. Team learning: building shared mental models. *Instructional science*, 39(3):283–301, 2011.

[386] Jasper van der Waa, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, and Ioana Cocu. Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI*, 8:640647, 2021.

[387] Jurriaan van Diggelen, Jonathan Barnhoorn, Ruben Post, Joris Sijs, Nanda van der Stap, and Jasper van der Waa. Delegation in human-machine teaming: Progress, challenges and prospects. In *Advances in Intelligent Systems and Computing*, volume 1254, pages 10–16. Springer, 2021.

[388] Jurriaan van Diggelen and Matthew Johnson. Team design patterns. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 118–126, 2019.

[389] Jurriaan van Diggelen, Karel van den Bosch, Mark Neerincx, and Marc Steen. Designing for meaningful human control in military human-machine teams. In *Research handbook on meaningful human control of artificial intelligence systems*, pages 232–252. Edward Elgar Publishing, 2024.

[390] Emma Van Zoelen, Tina Mioch, Mani Tajaddini, Christian Fleiner, Stefani Tsaneva, Pietro Camin, Thiago S Gouvêa, Kim Baraka, Maaike HT De Boer, and Mark A Neerincx. Developing team design patterns for hybrid intelligence systems. In *HHAI 2023: Augmenting Human Intellect*, pages 3–16. IOS Press, 2023.

[391] Emma M van Zoelen, Anita Cremers, Frank PM Dignum, Jurriaan van Diggelen, and Marieke M Peeters. Learning to communicate proactively in human-agent teaming. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 238–249. Springer, 2020.

[392] Mascha Van't Wout and Alan G Sanfey. Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, 108(3):796–803, 2008.

[393] Jose Vargas-Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. Impact of annotation modality on label quality and model performance in the automatic assessment of laughter in-the-wild. *IEEE Transactions on Affective Computing*, 2023.

[394] Rineke Verbrugge, Ben Meijering, Stefan Wierda, Hedderik Van Rijn, and Niels Taatgen. Stepwise training supports strategic second-order theory of mind in turn-taking games. *Judgment and Decision Making*, 13(1):79–98, 2018.

[395] Rineke Verbrugge and Lisette Mol. Learning to apply theory of mind. *Journal of Logic, Language and Information*, 17(4):489–511, 2008.

[396] Ruben S Verhagen, Alexandra Marcu, Mark A Neerincx, and Myrthe L Tielman. The influence of interdependence on trust calibration in human-machine teams. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 300–314. IOS Press, 2024.

[397] Ruben S Verhagen, Siddharth Mehrotra, Mark A Neerincx, Catholijn M Jonker, and Myrthe L Tielman. Exploring effectiveness of explanations for appropriate trust: Lessons from cognitive psychology. *arXiv preprint arXiv:2210.03737*, 2022.

[398] Ruben S. Verhagen, Mark A. Neerincx, Can Parlar, Marin Vogel, and Myrthe L. Tielman. Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance. In Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh, editors, *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, pages 2316–2318. ACM, 2023.

[399] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. Agent allocation of moral decisions in human-agent teams: Raise human involvement and explain potential consequences. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2302–2317, 2025.

[400] Ruben S Verhagen, Mark A Neerincx, X Jessie Yang, and Myrthe L Tielman. Advancing human-machine teaming: Definitions, challenges, future directions. In *HHAI 2025*, pages 49–59. IOS Press, 2025.

[401] Samuele Vinanzi and Angelo Cangelosi. CASPER: cognitive architecture for social perception and engagement in robots. *CoRR*, abs/2209.01012, 2022.

[402] Samuele Vinanzi, Angelo Cangelosi, and Christian Goerick. The collaborative mind: intention reading and trust in human-robot interaction. *Iscience*, 24(2), 2021.

[403] Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. Would a robot trust you? developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374, 4 2019.

[404] Ewart J De Visser, Marieke, M M Peeters, Malte, F Jung, Spencer Kohn, Tyler, H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, 12:459–478, 2020.

[405] Alan R Wagner, Paul Robinette, and Ayanna Howard. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(4):1–24, 2018.

[406] ME Walton, Steven W Kennerley, DM Bannerman, PEM Phillips, and Matthew FS Rushworth. Weighing up the benefits of work: behavioral and neural analyses of effort-related decision making. *Neural networks*, 19(8):1302–1314, 2006.

[407] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. "brilliant ai doctor" in rural clinics: Challenges in ai-powered clinical decision support system deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[408] Jie Wang, Wuji Lin, Xu Fang, and Lei Mo. The influence of emotional visual context on the judgment of face trustworthiness. *Psychology Research and Behavior Management*, pages 963–976, 2020.

[409] Jingfei Wang, Yan Yan, Yaoguang Hu, Xiaonan Yang, and Lixiang Zhang. A transfer reinforcement learning and digital-twin based task allocation method for human-robot collaboration assembly. *Engineering Applications of Artificial Intelligence*, 144:110064, 2025.

[410] Zhenting Wang, Takuya Kiyokawa, Natsuki Yamanobe, Weiwei Wan, and Kensuke Harada. Assembly task allocation for human-robot collaboration considering stability and assembly complexity. *IEEE Access*, 12, 2024.

[411] Daniel M. Wegner. *Transactive Memory: A Contemporary Analysis of the Group Mind*, pages 185–208. Springer New York, New York, NY, 1987.

[412] Christopher W Wiese, Christian Dormann, Hoda Vaziri, Louis Tay, Bart Wille, Job Chen, Lauren H Moran, and Yuhua Li. Happy work, happy life? a replication and comparison of the longitudinal effects between job and life satisfaction using continuous time meta-analysis. *Journal of Organizational Behavior*, 46(4):487–511, 2025.

[413] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[414] Jessica L Wildman, Marissa L Shuffler, Elizabeth H Lazzara, Stephen M Fiore, C Shawn Burke, Eduardo Salas, and Sena Garven. Trust development in swift starting action teams: A multilevel framework. *Group & organization management*, 37(2):137–170, 2012.

[415] Jessica L Wildman, Amanda L Thayer, Michael A Rosen, Eduardo Salas, John E Mathieu, and Sara R Rayne. Task types and team-level attributes: Synthesis of team classification literature. *Human Resource Development Review*, 11(1):97–129, 2012.

[416] Eirwen Williams. "mission impossible, now possible": These high-tech robots to heroically clear 2,850 radioactive sandbags from fukushima plant, March 2025.

[417] Travis J Wiltshire, Kyana Van Eijndhoven, Elwira Halgas, and Josette MP Gevers. Prospects for augmenting team interactions with real-time coordination-based measures in human-autonomy teams. *Topics in Cognitive Science*, 16(3):391–429, 2024.

[418] Michael Winikoff. Towards trusting autonomous systems. In Amal El Fallah Seghrouchni, Alessandro Ricci, and Tran Cao Son, editors, *Engineering Multi-Agent Systems - 5th International Workshop, EMAS 2017, Sao Paulo, Brazil, May 8-9, 2017, Revised Selected Papers*, volume 10738 of *Lecture Notes in Computer Science*, pages 3–20. Springer, 2017.

[419] Patrick Henry Winston. *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc., USA, 1992.

[420] NB Yeswanth et al. An ai-powered interactive assistant: Integrating multimodal interaction for enhanced user experience. In *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, pages 1–7. IEEE, 2024.

[421] Michael Yip, Septimiu Salcudean, Ken Goldberg, Kaspar Althoefer, Arianna Menciassi, Justin D Opfermann, Axel Krieger, Krithika Swaminathan, Conor J Walsh, He Huang, et al. Artificial intelligence meets medical robotics. *Science*, 381(6654):141–146, 2023.

[422] Mifrah Youssef, En-Nouaary Abdeslam, and Dahchour Mohamed. A jade based testbed for evaluating computational trust models. In *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pages 1–7, 2015.

[423] Mike Zajko. Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, 16(3):e12962, 2022.

[424] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. Investigating ai teammate communication strategies and their impact in human-ai teams for effective teamwork. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–31, 2023.

[425] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Bart Knijnenburg, and Guo Freeman. Verbal vs. visual: How humans perceive and collaborate with ai teammates using different communication modalities in various human-ai team compositions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–34, 2024.

[426] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[427] Michelle D Zhao, Reid Simmons, and Henny Admoni. Learning human contribution preferences in collaborative human-robot tasks. In *Conference on Robot Learning*, pages 3597–3618. PMLR, 2023.

List of SIKS dissertations

2016
01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UvA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UvA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017
01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UvA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UvA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VUA), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VUA), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TU/e), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VUA), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OU), Learning visually grounded and multilingual representations

2020
01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TU/e), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
31 Gongjin Lan (VUA), Learning better – From Baby to Better
32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
07 Armel Lefebvre (UU), Research data management for open science
08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (UL), Robust rules for prediction and description
24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
04 Mike Huisman (UL), Understanding Deep Meta-Learning
05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach

10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
19 Fang Hou (UU), Trust in Software Ecosystems
20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
23 Roderick van der Weerdt (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning
29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering

35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment
36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing
37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval
38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity
39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence
40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans
41 Mozhgan Vazifehdoostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights
42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data
43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education
44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning
45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings
46 Ya Song (TU/e), Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making
47 Tom Kouwenhoven (UL), Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions
48 Evy van Weelden (TiU), Integrating Virtual Reality and Neurophysiology in Flight Training
49 Selene Báez Santamaría (VUA), Knowledge-centered conversational agents with a drive to learn
50 Lea Krause (VUA), Contextualising Conversational AI
51 Jiaxu Zhao (TU/e), Understanding and Mitigating Unwanted Biases in Generative Language Models
52 Qiao Xiao (TU/e), Model, Data and Communication Sparsity for Efficient Training of Neural Networks
53 Gaole He (TUD), Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems
54 Go Sugimoto (VUA), MISSING LINKS Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities
55 Sietze Kai Kuilman (TUD), AI that Glitters is Not Gold: Requirements for Meaningful Control of AI Systems
56 Wijnand van Woerkom (UU), A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law

57 Syeda Amna Sohail (UT), Privacy-Utility Trade-Off in Healthcare Metadata Sharing and Beyond: A Normative and Empirical Evaluation at Inter and Intra Organizational Levels
58 Junhan Wen (TUD), "From iMage to Market": Machine-Learning-Empowered Fruit Supply
59 Mohsen Abbaspour Onari (TU/e), From Explanation to Trust: Modeling and Measuring Trust in Explainable Decision Support
60 Marcel Jurriaan Robeer (UU), Beyond Trust: A Causal Approach to Explainable AI in Law Enforcement
61 Shuai Wang (VUA), Links in Large Integrated Knowledge Graphs: Analysis, Refinement, and Domain Applications
62 Khaleel Asyraaf Mat Sanusi (OU), Augmenting a learning model within immersive learning environments for psychomotor skills
63 Rashid Zaman (TU/e), Online Conformance Checking on Degraded Data
64 Jens d'Hondt (TU/e), Effective and Efficient Multivariate Similarity Search
65 Aswin Balasubramaniam (UT), Disentangling Runner Drone Interaction Potentialities

2026 01 Pei-Yu Chen (TUD), Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support
02 Hezha Hassan Mohammedkhan (TiU), Estimating Body Measurements of Children from 2D Images: Towards the Automatic Detection of Malnutrition
03 Kyriakos Psarakis (TUD), Democratizing Scalable Cloud Applications: Transactional Stateful Functions on Streaming Dataflows
04 Boyu Xu (UU), Exploring Indirect Relations Between Topics in Neuroscience Literature Using Augmented Reality to Inform Experimental Design
05 Koen Minartz (TU/e), Stochastic Simulation with Geometric Deep Generative Models
06 Azim Afroozeh (CWI, VUA), FastLanes: A Next-Gen File Format
07 Inès Blin (VUA), Narrative Understanding with Knowledge Graphs
08 Paul van Vulpen (UU), Debating Digital Dominance: Decentralized Technology Governance For Strategic Autonomy
09 Afrizal Doewes (TU/e), Rethinking Automated Essay Scoring: Agreement, Fairness, and Feedback
10 Nikolaos Delapaschos Kondylidis (VUA), Establishing Task-Oriented Understanding between Agents
11 Işıl Baysal Erez (UT), Handling Missing Data with Meta-Learning and Large Language Models
12 Xue Li (UvA), From Fine-tuning to Prompting: A Paradigm Shift in Knowledge Graph Construction
13 Isaac da Silva Torres (VUA), Guidelines To Flux Between Conceptual Models: Understanding Complex Digital Business Ecosystems
14 Philip Lippmann (TUD), Synthetic Data for Robust Language Modelling
15 Rashmi Khazanchi (OU), Artificial Intelligence in Education: Impact of AI-Based Systems on Mathematics Achievement

16 Carolina Ferreira Gomes Centeio Jorge (TUD), Modelling Artificial Trust for Effective Human-AI Teamwork
17 Maria Tsfasman (TUD), Towards Predicting Memory in Multimodal Group Interactions
18 Riccardo Lo Bianco (TU/e), Deep Reinforcement Learning for Automated Decision-Making in Process Management Systems
19 Israel Campero Jurado (TU/e), Innovations in Optimization and Applications in Healthcare
20 Iftitahu Ni'mah (TU/e), Contrastive Learning and Evaluation in Low Resource Scenario of Natural Language Processing
21 Francisco N.F.Q. Simoes (UU), Causality, Information, and Decision-Making
22 Ruben Verhagen (TUD), Transparent and Explainable Agents for Human-Agent Teaming

Acknowledgments

Throughout my PhD, I often thought about what I would write in this section. I have much to be grateful for, and I appreciate the opportunity to put it on paper and share it with those who supported me along the way. This PhD title is the result of a five-year journey, but that journey did not begin five years ago. I would like to start by broadly acknowledging the advantages life gave me before those years began. From here onwards, I will move on to the people who sustained me, through their support, their love, or simply by being there. As my father once told me, *we can only aim as high as our own ceilings*. My ceilings were high, so I aimed high, and here I am today.

The family I was born into is, of course, the main factor shaping how high my ceilings are. As such, I must start by thanking my parents, Elsa and Alípio, whose priority was always to make those ceilings as high as possible. They did so by ensuring we received the best education they could provide, both in and out of school, by showing us the world, debating societal issues, giving us access to books without limits, and taking us often to the cinema and the park. Both of my parents completed a PhD and remain involved in academia. I experienced this world through their eyes long before I arrived here, which made my own PhD trajectory feel feasible and come with few surprises. Thank you, mum, for teaching me that independence should be a priority and that gender should not define careers or lifestyles. Thank you for your constant care, and for making me feel that you are always ready to drop everything and come to my side when I need you. And dad, thank you for teaching me how to find joy in learning, turning even apparently boring things into exciting puzzles or moments of quiet contemplation. Thank you for your affection, and for all the long conversations about everything and nothing. Adoro-vos. Thank you to my grandparents, Tinda, Nando, Lita, and Tó, for creating the opportunities my parents had, for fighting for their goals, and for supporting us. Special thanks to my grandmothers, who walked so I could run.

Turning now to what academia gave me, I would like to thank my supervisors, Myrthe and Catholijn. Myrthe, thank you for managing my expectations, for teaching me how to set boundaries, and for everything you do to promote a healthier academic environment. Thank you for your quick thinking, creativity, and remarkable organisational skills. Thank you for allowing me to try and fail, for encouraging me not to shy away from difficult topics, and for being there to redirect me when needed. Catholijn, thank you for your mentorship and for helping me think beyond my PhD and specific research questions. Thank you also for your leadership of the Interactive Intelligence group, which you worked hard to make diverse, engaging, and welcoming. It is a group that brought me a great deal of joy, friendship, and learning, and for that I am very grateful. Thank you both for choosing me, for trusting me, and for seeing me as more than my academic output. Thank you to Mark, my most recent promotor, for reading my work, suggesting improvements, posing relevant questions, and for helping me navigate ethical challenges in collaborations.

To my committee, Max Mulder, Josette Gevers, Rineke Verbrugge, and Rino Falcone, thank you for your availability and for your willingness to read and oppose my work. Thank you for paving the way, within your respective fields, for the research questions I pursue. Max, thank you for joining us at my second yearly meeting, for offering an honest perspective on my progress, and for helping us redirect and refocus. I was feeling somewhat lost at that point, and that meeting was far more helpful than you may realise. Josette, thank you for helping build the bridge from organisational psychology to this multidisciplinary world of human–agent collaboration. Rino and Rineke, thank you for your foundational computational work on teamwork, trust, and decision-making. It is always a pleasure to read your work, and I feel very fortunate to have the opportunity to meet you and discuss my research with you.

Doing a PhD can be a lonely walk, but it never felt that way to me, since I was incredibly lucky to be part of the Interactive Intelligence group. Over these five years, I met many interesting people who taught me so much. I discussed an amazing range of topics with so many of you, and through those conversations I learned to see the world through many different eyes. Thank you all II members for the help, for dropping by my office for small interactions, and for the meals we shared together.

In particular, thank you to my paranymphs, Linyun and Micha, for once again promptly accepting to help me and for always being there for me. I love you two individually, but when I share a room with both of you, a whole new level of fun is unlocked. Thank you for all the laughter. Micha, somehow we are very similar, and it has always been so easy to spend hours together. Thank you for walking with me so many times and for turning those walks into a seminar series. Thank you for sharing your thoughts and feelings, and for helping me process mine. Linyun, thank you for giving me such consistent presence and love from the moment you entered my life. I had no idea how much I needed it. Thank you for all the random food you brought me and for making it so comfortable to be myself around you. Thank you for making Thursday dinners with Pei-Yu a tradition and for helping me finally adopt her after years of trying.

Thank you to the other colleagues who also became friends. Masha, thank you for all the fun, and not so fun, moments we shared. From crochet to co-writing, you were there for me throughout. Pei-Yu, thank you for listening to so many of my stories in the office and for generously laughing often enough to make me feel funny. For all the dinners, gym sessions, and help - thank you. Ruben, thank you for making me a welcome kit when I arrived in the Netherlands in the middle of a pandemic. We started and finished together, progressing steadily side by side, and that gave me structure and familiarity. Emma, thank you for persisting with online socials during the pandemic, for welcoming any discussion, and for having me as your paranymph. You are probably the first person I connected with in II and it always a pleasure to chat away with you. Mani, thank you for your kindness, warmth, heart-to-heart conversations, and all the delicious dishes you have made for us. Enrico, don't be jealous, your cooking is great too. Thank you also for organising all the group activities that created great memories, for checking on me and making sure I kept it together. Mo, thank you for your enthusiasm at every bi-weekly meeting and game night, and for being the great colleague you are. Nele, thank you for helping me at times when I felt at bay, whether with Bayesian statistics or with keeping my experiments simple. Sid, thank you for showing the way and for suggesting co-authoring from my very first day.

Rolf, thank you for your care and kindness, and for inviting me over on Sinterklaas - it brought me great joy.

To Morita, Amir, Paul, Zuzanna, Davide, Elena, Deborah, Shubhalaxmi, Shambhawi, and Sietze, thank you for the chats by the coffee machine or elsewhere. To Fran, Ilir, and Bernd, thank you for reviewing my papers and for always being available to give advice. To Miguel, Agnes, Antonio, Urja, Ruben, Joanna, Eric, Yu-Wen, Alex, Bram, Michiel, Yanzhe, Edgar, Koray, and Jinke, thank you for making this group special. Thank you to Anita, Ruud, Bart, and Wouter for the support and all the backstage magic. Thank you to Luciano, Pradeep, Willem-Paul, Catha, Stephanie, Luuk, Frans, and Frank for all the interesting topics you have brought to the group. Thank you for embracing the multidisciplinarity they create, and for choosing a careful approach to technology development.

At TU Delft, I met exceptional people across its various layers and initiatives. My PhD was part of the AI Lab AI*MAN, which is part of the Delft AI Initiative. Thank you to everyone who was part of the lab at some point and interacted with my work, including Anahita, Guido, Max, Mirko, and Shanza. Shanza, thank you for your friendship and for all the great moments together. Thank you to the staff who handle communications and social events, particularly Charlotte. I also want to thank others I met through this initiative, such as Karin and Dina.

When I arrived at TU Delft, in the middle of a lockdown, I looked for activities (online, of course) and found the Improv group. Although it gave me a lot of anxiety, it also allowed me to discover a new side of myself and to feel that I belonged in a community. I want to thank the whole DIG group for organising it and for being so welcoming and kind, and to everyone I had the chance to talk to one-on-one. I am especially grateful to Carlos and Siert. Later, I met Chirag, who at the time was Chair of the faculty's PhD Council and invited me to join. Thank you for bringing me in and for all the great discussions. It is always a pleasure to meet you in the hallway. I also want to thank all other members of the council while I was part of it, particularly Willem, Michael, Ids, Youri, Imara, Lorena, Cristian, Henry, Xiaoling, and Aysun. Thank you, and thanks to all other members before and after me, for your contributions to improving PhD life. Thank you to Geurt, Ada, and Sanne for supporting me and believing that my voice could make a difference. After the PhD Council, I was part of the Integrity Board, where I had the pleasure of participating in difficult but necessary discussions. Thanks for having me, especially Ibo and Remco, for giving me that opportunity. Thank you to Jens Kober, who agreed to be my mentor and spent time discussing my journey with me. I am also grateful to Hayley Hung for inspiring me at different moments throughout my PhD, and to Tiffany, for her insights and easy smile.

While I was supervised during my PhD, I had the privilege of supervising some great students, and I want to thank them all. In particular, those I supervised during their MSc thesis, Nikki, Jan-Willem, and Charlotte, and those I had the chance to publish with, Elena, Elena, Sahar, and Razvan.

The nature of my research required user studies, which in turn required participants. I am deeply grateful to everyone who spent a bit of their time taking part in one (or more!) of my experiments. For reasons of anonymity, I will not list names. My work also relied on multidisciplinary collaboration, and I am thankful to have had the opportunity to sit alongside people with different perspectives and academic backgrounds. Building a

research community around a new topic is not easy, and it took me some time to find my research bubble. The first step towards finding it was undoubtedly meeting Anna-Sophie and Eleni. Thank you so much for the collaboration, for teaching me your social science perspective, and for believing that I also had something to teach you. Thank you for your kindness and friendship; I will miss working with you. Thank you, Anna, for helping and motivating me to organise the first MultiTTrust workshop. The first edition of MultiTTrust was the moment when I truly felt I had found my research community. Thank you to everyone who kept it going, particularly Nico, Morgan, Myke, Francesco, Andre, and the other organisers and programme committee members. Thank you also to the many early-career researchers who are shaping this community, and with whom I crossed paths along the way, including Esther, Tilman, Kavyaa, Samuele, Aaquib, Samin, Raquel, and many others. Thank you to all those who reviewed our multidisciplinary work with an open mind, interest and curiosity, and to all the venues that make space for new communities to form.

During my PhD, I greatly valued the opportunities to discuss my work with different audiences. Thank you to those who invited me to present my work to their research groups, particularly Karinne Ramirez-Amaro, Giovanni Misitano, Michelle Zhao, and Lionel P. Robert. Thank you as well to those who visited Delft to present their work, including Matthew Johnson, who helped me during the conceptualisation phase of chapter four.

Having the opportunity to do a research visit overseas was a long-standing goal of mine, and one that was only possible thanks to Lionel P. Robert and Dawn Tilbury. Thank you for hosting me in your research groups for two months. Research-wise, I felt a strong sense of belonging that made the experience truly worthwhile. My time at the University of Michigan was undoubtedly a highlight of my PhD. For that experience, I also owe thanks to Annette and Connor. Truly, thank you for befriending me so quickly and easily; you made it very hard for us to leave. Thank you to Xin, Samia, Arsha, Alia, Zariq, and Jo for taking me in as one of your own. Thank you to Patrícia Alves-Oliveira, Xi Jessie Yang, and Zana Buçinca for making time to meet me during that period. Thank you all for your contributions to my work. Last but not least, thank you to Ewart de Visser for the collaboration that led to chapter five.

It is almost impossible to individually mention all those who contributed to my work or to my growth over the last five years. As such, I would like to broadly thank anyone who talked to me during a coffee break, who stopped by my posters to ask a question, who sent me an email, who gave me a call, who recommended me a paper to read, who asked me a question or told me a story. To me, my PhD was really made of the aggregation of these moments. That's what kept me going and that is what made it so special.

Delft is a special place, full of fascinating people. I have met many who have shaped my life in one way or another. Thank you to Saba, Kamila, and Jay. Thank you to Jens, Akshaya, Orestes, Martin, Maria, Maria, Gabriel, and Roland. Thank you also to those I knew before but crossed paths with again in Delft, particularly Muhammad and Zé. I am also grateful for the moments I shared with fellow Portuguese PhDs I met in Delft over the last year, including João, Rúben, Leonor, and Guilherme. Thank you especially to Manel, Maria, and Riccardo for bringing a sense of home closer to me. Thank you to older friends that lived nearby and visited us, including Annelien, Luca, Tomi, Edina, and Izadi.

Now it is time to thank all the support I received from outside of academia and beyond

the places where academia took me during these five years. Those who were already there for me before I started this PhD, and still are. I want to thank my brother Miguel, whom I admire for everything I am not and whom I care for deeply. Thank you for being there for me and for growing closer to me despite the distance. Thank you to the rest of my family: Adriana, Nando Zé, Bi, Sérgio, Laia, Jorge, Joana, Tiz, Hugo, Xica, Rita, Pedrocas, and Tiago. Obrigada, a todos vocês, por me apoiarem, por quererem saber de mim e das minhas conquistas e por me receberem sempre tão bem. Gosto muito de vocês. My friends are spread across the world, and I unfortunately do not have the chance to meet them as much as I wished. However, they all stay close. Thank you so much for bearing with this stranger, for your interest in my life, for caring, and for showing up. I love you all. Thank you to Carlota, Maria, and Catarina. Thank you to Sérgio, Paulo, Rui, Marta, Cris, Tiago and Diogo. Thank you to Flóra, Laura, Gloria, Danae, and Toshi. Thank you to Trindade, Filipa, António, Ariana, and Cláudio.

Doing a PhD can also take a toll on one's mental health, especially when it involves moving and starting anew, making adult choices for the first time, or trying to find new friends and hobbies. It is a period of self-discovery: beautiful, but also daunting. It can be a time of doubt, failure, and self-sabotage. I knew this from the start, and I have always taken my mental health seriously over these five years, keeping boundaries, healthy routines, and consistent therapy. I was fortunate to stay healthy. Without naming anyone, I would like to thank the mental health professionals who supported me and several of my peers along this journey.

In this last paragraph, I would like to thank Dávid, this extraordinary person I had the absolute privilege of marrying during my PhD. When I had the (crazy?) idea of applying to this PhD position in TU Delft, we were happily living in Tokyo, and the world was in lockdown. Although that idea brought uncertainty and challenges to us, Dávid always supported me. Eventually, he also moved to Delft to live with me. For me, there is a Delft before Dávid moved and a Delft after, and the latter is far more beautiful. My research became more interesting and fruitful and my days simply happier. Thank you, Dávid, for moving, for consistently keeping us going, and for always being there for me. For reminding me of my purposes, principles, and for accepting me as a whole. For all the patience, help, dedication, companionship, and, of course, love. I love you. To finish, I want to thank Dávid's family, Ági, Zsolti, Kristóf and Patricia, all the friends that visited us from Hungary, and those who supported us throughout. Köszönöm szépen mindenkinek.

Curriculum Vitæ

Carolina Ferreira Gomes Centeio Jorge

1996/09/20 Born in Porto, Portugal

Education

2020-2025 **PhD in Computer Science**
Delft University of Technology, The Netherlands
2024 *Visiting Researcher*
University of Michigan, Ann Arbor, MI, USA

2014-2019 **BSc & MSc in Informatics and Computing Engineering**
University of Porto, Portugal
2019 *MSc Dissertation Internship*
University of Twente, The Netherlands
2018 *Exchange Semester (Erasmus)*
Polytechnic University of Catalonia (UPC), Spain

Work Experience

2025- **Senior Data Scientist Specialist**
Glintt Next, Portugal

2019-2020 **Traineeship (Vulcanus in Japan Program)**
OMRON Sinic X. Corporation, Tokyo, Japan

2017-2018 **Research Assistant**
INESCTEC, Porto, Portugal
Faculty of Engineering of University of Porto, Porto, Portugal

Institutional Governance

2024-2025 Member of the TU Delft's Integrity Board
2023 Chair of the TU Delft EEMCS PhD Council

List of Publications

Under Review and/or In Press

1. **Centeio Jorge, C.**, Jonker C. M., Robert, L. P. de Visser, E. J. & Tielman, M. L. *I know you're capable, but are you willing? Allocating tasks in human-machine teams.*

2. **Centeio Jorge, C.**, Ulfert, A. S., Georganta, E., Mehrotra, S., Tielman, M. L., & Jonker, C. M. Multidisciplinary Theory Building for Human-AI Team Trust. In *The Oxford Handbook of Computational Group and Team Dynamics.* Oxford University Press.

3. Tielman, M. L., Bailey, M., Frattolillo, F., **Centeio Jorge, C.**, Ulfert, A., & Meyer-Vitali, A. Multidisciplinary Perspectives on Human-AI Team Trust. *Interaction Studies.*

2026

1. Ning, C.W., **Centeio Jorge, C.**, Tielman, M. L. & Neerincx, M. A. (2026). "What's on your mind?": Understanding the Development of Multidimensional Trust in Social Robots. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26).* ACM, New York, NY, USA.

2025

1. **Centeio Jorge, C.**, Dumitrescu E., Jonker, C. M., Loghin, R., Marossi, S., & Tielman, M. L. (2025). How Should Your Artificial Teammate Tell You How Much It Trusts You? In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents (IVA '25).* ACM, New York, NY, USA.

2. Atzmueller, M., **Centeio Jorge, C.**, Rebelo de Sá, C., Heravi, Behzad M., Gibson, Jenny L., & Rossetti, Rosaldo J. F. (2025). Mining exceptional social behavior on attributed interaction networks. In *Machine Learning* 114 (11), 243.

2024

1. **Centeio Jorge, C.**, Jonker, C. M., & Tielman, M. L. (2024). Interdependence and trust analysis (ITA): a framework for human–machine team design. *Behaviour & Information Technology*, 1-21.

2. **Centeio Jorge**, C., de Visser, E. J., Tielman, M. L., Jonker, C. M., & Robert, L. P. (2024). Artificial Trust in Mutually Adaptive Human-Machine Teams. In *Proceedings of the AAAI Symposium Series* (Vol. 4, No. 1, pp. 18-23).

3. Ulfert, A. S., Georganta, E., **Centeio Jorge, C.**, Mehrotra, S., & Tielman, M. (2024). Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework. *European Journal of Work and Organizational Psychology*, 33(2), 158-171.

4. **Centeio Jorge, C.**, van Zoelen, E. M., Verhagen, R., Mehrotra, S., Jonker, C. M., & Tielman, M. L. (2024). Appropriate context-dependent artificial trust in human-machine teamwork. In *Putting AI in the Critical Loop* (pp. 41-60). Academic Press.

5. Brandizzi, N., **Centeio Jorge, C.**, Cipollone, R., Frattolillo, F., Iocchi, L., & Ulfert-Blank, A. S. (2023, December). 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).

6. **Centeio Jorge, C.**, Jonker, C. M., & Tielman, M. L. (2024). How should an AI trust its human teammates? Exploring possible cues of artificial trust. *ACM Transactions on Interactive Intelligent Systems*, 14(1), 1-26.

7. Mehrotra, S., **Centeio Jorge, C.**, Jonker, C. M., & Tielman, M. L. (2024). Integrity-based explanations for fostering appropriate trust in AI agents. *ACM Transactions on Interactive Intelligent Systems*, 14(1), 1-36.

2023

1. **Centeio Jorge, C.**, Bouman, N. H., Jonker, C. M., & Tielman, M. L. (2023). Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork. *Frontiers in Robotics and AI*, 10, 1143723.

2. **Centeio Jorge, C.**, Jonker, C. M., & Tielman, M. L. (2023). Artificial trust for decision-making in human-AI teamwork: Steps and challenges. In *Proceedings of the Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence co-located with HHAI 2023* (Vol. 3456, pp. 150-156). CEUR-WS.

3. **Centeio Jorge, C.**, & Ulfert-Blank, A. S. (2023). MULTITTRUST - Multidisciplinary Perspectives on Human-AI Team Trust. In *Proceedings of the Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence co-located with HHAI 2023* (Vol. 3456, pp. 132-136). CEUR-WS.

4. Mehrotra, S., **Centeio Jorge, C.**, Jonker, C. M., & Tielman, M. L. (2023). Building Appropriate Trust in AI: The Significance of Integrity-Centered Explanations. In *HHAI 2023: Augmenting Human Intellect* (pp. 436-439). IOS Press. *[best poster award]*

5. **Centeio Jorge, C.**, Atzmueller, M., Heravi, B. M., Gibson, J. L., Rossetti, R. J., & Rebelo de Sa, C. (2023). "Want to come play with me?" Outlier subgroup discovery on spatio-temporal interactions. *Expert Systems*, 40(5), e12686.

2022

1. van Rhenen, J. W., **Centeio Jorge, C.**, Matej Hrkalovic, T., & Dudzik, B. (2022). Effects of social behaviours in online video games on team trust. In *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play* (pp. 159-165).

2. **Centeio Jorge**, C., Tielman, M. L., & Jonker, C. M. (2022). Assessing artificial trust in human-agent teams: a conceptual model. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents* (pp. 1-3).

3. **Centeio Jorge, C.**, Tielman, M. L., & Jonker, C. M. (2022). Artificial trust as a tool in human-AI teams. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 1155-1157). IEEE.

2021

1. Centeio Jorge, C., Mehrotra, S., Tielman, M. L., & Jonker, C. M. (2021). Trust should correspond to trustworthiness: A formalization of appropriate mutual trust in human-agent teams. In *22nd International Trust Workshop 2021*. CEUR-WS.

2019

1. **Centeio Jorge, C.**, Atzmueller, M., Heravi, B. M., Gibson, J. L., de Sá, C. R., & Rossetti, R. J. (2019, August). Mining exceptional social behaviour. In *EPIA conference on artificial intelligence* (pp. 460-472). Springer International Publishing.

2018

1. **Centeio Jorge, C.**, & Rossetti, R. J. (2018). On Social Interactions and the Emergence of Autonomous Vehicles. In *VEHITS* (pp. 423-430).

Included in this thesis.
Won a best paper, tool demonstration, or proposal award.