Human or Machine? Exploring How Anthropomorphism, Performance and Social Intelligence Impact Trust in Human-AI Teams

Submitted for the degree of Doctor of Philosophy (PhD)

Morgan Elizabeth Bailey, MSc

School of Psychology & Neuroscience & School of Computing Science
College of Science & Engineering

University of Glasgow

December 2024

Abstract

In an era where Artificial Intelligence is becoming integral to human teams, understanding the role of trust in Human-AI Teams is essential for effective collaboration. This thesis investigates how anthropomorphism, AI system performance, and social intelligence influence trust calibration, team performance, and human perceptions of AI teammates. The research addresses significant gaps in Human-AI Team literature by drawing on interdisciplinary insights from psychology, computing science, and human-computer interaction. The work is structured into six chapters, each contributing to a comprehensive understanding of trust in Human-AI Teams.

Chapter 1 provides a literature review on the dynamics of human-agent teams, trust, and social intelligence. It explores how anthropomorphic design, AI reliability, and social intelligence contribute to trust development, highlighting the limitations of existing theories and the need for a multidisciplinary approach.

Chapter 2 presents a bibliometric analysis of trust research from 1922 to 2021. By analysing 39,628 documents, this chapter identifies key research trends, foundational contributions, and interdisciplinary intersections. The study reveals the evolving nature of trust research and underscores the importance of integrating diverse disciplinary insights to address complex trust dynamics in Human-AI Teams.

Chapter 3 explores the impact of anthropomorphism and AI system reliability on trust and performance in Human-AI Teams. Using experimental methods, it demonstrates that while anthropomorphic design can enhance trust, this effect is contingent on AI reliability. The findings highlight the risks of overtrust when anthropomorphic cues are paired with unreliable AI systems.

Chapter 4 investigates the role of emojis and AI reliability in shaping team performance and trust. Results show that AI teammates using emojis can foster a sense of social connection and trust, but this effect varies based on the system's reliability. The study emphasises the nuanced relationship between social cues and trust calibration.

Chapter 5 examines how social alignment in AI, the ability to adapt behaviour to match human social expectations, affects trust and team behaviours. Findings indicate that AI that demonstrates adaptive social alignment behaviour can benefit trust. However, misaligned social AI can lead to mistrust and reduced performance and has more impactful effects.

Chapter 6 synthesises the key findings, offering conclusions and practical recommendations. The research underscores the importance of calibrated trust, ensuring humans neither over-rely nor under-rely on AI. Effective Human-AI Teams require AI systems that balance anthropomorphic design, transparency, and social intelligence to foster sustainable trust. The chapter highlights the need for ongoing interdisciplinary research and ethical considerations to guide the development of AI teammates.

Overall, this thesis contributes to understanding trust dynamics in Human-AI Teams by demonstrating that successful collaboration hinges on the careful integration of anthropomorphic cues, system reliability, and social intelligence. The findings provide information for designing AI systems that are not only reliable but also socially intelligent, fostering more effective and ethical human-AI Teams.

Table of Contents

List of Tables

List of Figures

List of Accompanying Material

Acknowledgements

A huge thank you to Frank, who has guided, supported, and reassured me every step of the way. I am deeply grateful for the time you gave me and the invaluable lessons you taught me. Thank you for guiding me through not only the trials and tribulations of academia but also through wider life. Your calm presence made an enormous difference, and I will greatly miss our meetings and seeing your dogs.

To my undergraduate and master's supervisor, Jeremy: thank you for believing in me before I believed in myself. Your confidence and encouragement set me on this path, and I would not be here without you. I will always be grateful for the time and support you gave me, which had such a profound impact on the trajectory of my life.

Thank you to my fellow PhD students and the staff within the CDT, who have consistently been there to listen, collaborate, and offer support. Special thanks to Gordon and Ellen for sharing in the challenges and being a constant light throughout this journey. To my friends from undergraduate, Tay, Alex, and Harry, thank you for spending nine years listening to my worries and stresses about university, and for always bringing me joy, even on the toughest of days. Thanks also to Clacker, for always being my biggest superfan, bringing the loudest laughs, and reminding me that the best friendships often bloom in the most unexpected ways.

To my mum, whose ambition and drive always showed me that I could do hard things. To my dad, who has supported me unconditionally in everything I have ever done and never once doubted me. You both taught me so much and made me believe I could achieve anything I set my mind to. Your unwavering faith and love have been my foundation. To Joe and Jack, thank you for being the best big brothers anyone could wish for, for always lifting my spirits, reminding me there is more to life than my PhD, and supporting me in everything I bring your way.

To my dear Frank and Charlie, for keeping me company, snacking on popcorn with me, and staying up through the late nights. For two tiny hamsters, you brought so much joy — I miss you both.

Finally, to my partner Aaron: you have been a constant source of comfort, strength, and love. Thank you for being by my side through the highs and lows, the laughter and tears, and for never losing faith in me, even when I lost faith in myself. Whether through spontaneous trips to the West Coast, endless cups of coffee, or laughter when things felt overwhelming, your support has meant more than I can express. I am endlessly grateful to have had you with me throughout this journey.

For Laarni - Tag you're it! X

Author’s Declaration

I declare that I am the sole author of this thesis. To the best of my knowledge, this thesis contains no material previously published by any other person except where due acknowledgement has been made. This thesis contains no material which has been accepted as part of the requirements of any other academic degree or non-degree program.

Definitions/Abbreviations

AI – Artificial Intelligence

CASA – Computers Are Social Actors

EI - Emotional Intelligence

HAI – Human-AI Interaction

HAT – Human-AI Team

HCI – Human Computer Interaction

HI - Hybrid Intelligence

PtTM – Propensity to Trust Machines

SA - Situational Awareness

SI – Social Intelligence

TiA – Trust in AI

XAI – Explainable AI

Chapter 1 A Literature Review of Past Work Investigating the Dynamics of Human-Agent Teams, Trust, and Social Intelligence

1.1 Introduction

In today's world, intelligent machines are no longer just tools; they are increasingly becoming integral members of human teams across industries, from healthcare and finance to emergency response and defence. For example, AI-driven systems like IBM's Watson assist in diagnosing medical conditions, while autonomous drones collaborate with human soldiers in the military. These examples underscore a critical transformation: as Artificial Intelligence (AI) evolves from performing isolated tasks to acting as collaborative partners, understanding the dynamics of Human-Agent Teams (HATs) becomes paramount.

However, while the potential of HATs is vast, their success hinges on overcoming significant challenges. How can trust be cultivated when AI lacks the emotional cues of human teammates? What role does anthropomorphism play in HATs, and is it always beneficial? Can AI systems develop sufficient Social Intelligence (SI) to navigate complex human team dynamics effectively? These questions remain inadequately addressed in existing research, leaving gaps in our understanding of what makes HATs successful.

This literature review aims to synthesise the research on HATs, focusing on three critical factors: trust, anthropomorphism, and SI. By examining these interconnected themes, this review highlights the current state of research, identifies unresolved issues and provides a roadmap for future investigations.

1.2 Human-Agent Teams

The shift from viewing AI as mere tools to considering them as collaborative teammates represents a significant development in AIand Human-Computer Interaction (HCI). Traditionally, researchers viewed AI as a tool designed to perform specific tasks more efficiently than humans. In recent years, HATs have become a novel area of research (Rix, 2022), shifting AI's role from a tool to a

collaborative teammate (Berretta et al., 2023; McNeese et al., 2018, 2021; Rix, 2022). This shift stems from recognising that practical HATs rely on human perceptions of AI as emotionally intelligent, communicative partners rather than task-driven tools (Wynne & Lyons, 2018). To foster effective collaboration, shared mental models and communication processes are crucial for establishing collective goals and trust (Lyons et al., 2021).

Historically, complex technical systems, particularly in the mid-20th century, focused on mathematical and logical problem-solving, reinforcing their image as efficient tools. However, advancements in machine learning and natural language processing in the late 20th century facilitated a transition toward more human-like interactions. Recently, AI tools such as ChatGPT and Co-Pilot (OpenAI., 2024) have further solidified AI's role as a potential workplace collaborator. Co-Pilot (Microsoft, 2024) exemplifies this transition by acting as a coding partner, understanding code context, suggesting code completions, and even generating entire functions, moving beyond the capabilities of a simple code editor. However, there are theories that even before these recent changes in AI, humans still viewed computers as teammates, and we will discuss these ideas further in the next section.

1.2.1 Computers as Teammates

The Computers Are Social Actors (CASA) paradigm (Nass et al., 1994b) has influenced our understanding of how humans interact with computers, suggesting that we instinctively apply social norms to these interactions. For example, Nass et al. (1994) demonstrated that even experienced computer users unconsciously exhibit politeness and gender stereotypes when interacting with computers. CASA suggests that our social responses to technology are deeply ingrained and often automatic.

However, CASA has limitations. Firstly, it does not fully account for how context shapes interactions. For instance, users might readily accept suggestions from a music recommendation AI in a leisure context but might be more critical of similar suggestions from an AI financial advisor in a work context where economic security is at stake (Angerschmid et al., 2022; Bansal et al., 2021; Salimzadeh et

al., 2023). Additionally, CASA overgeneralises by assuming all users engage with computers socially, neglecting individual differences like personality traits, previous experience with AI and cultural backgrounds (Agarwal & Prasad, 1999; Yi et al., 2005). Gambino (2020) emphasises that relationships with technology are not static. Early interactions might be influenced more by novelty and general social tendencies. However, repeated use leads to a more nuanced understanding. Imagine a user initially treating a chatbot politely but, over time, learning it responds the same regardless of their tone. Learning about the AI's consistent behaviour could lead them to adopt a more direct communication style, a shift based on individual differences not accounted for in CASA's fixed script model.

Furthermore, recent evidence suggests that this effect might be waning, particularly for technologies that have become ubiquitous. Heyselaar (2023) directly challenges CASA by replicating a foundational study on politeness towards computers, finding no evidence that participants today exhibit more politeness when interacting with the same computer. We could attribute this shift in user behaviour to the increasing prevalence of technology in our lives, and CASA might occur strongest when applied to emergent technologies. Due to these changes, there is a need to consider the evolving nature of HCI and explore alternative frameworks that move beyond the assumption that all computers are inherently social actors.

Recognising these limitations, Gambino et al. (2020) proposed a refined model incorporating modern technological interactions. Gambino's (2020) paper proposes that people may mindlessly apply human-computer scripts similarly to human-human scripts during social interactions with technologies. This theoretical extension to CASA suggests that users develop distinct scripts for interacting with technology, not simply borrowed from human-human interaction. This model addresses Heysel's (2023) findings by acknowledging that users may not always treat computers as social actors, especially as technology becomes more familiar and integrated into daily life. The social affordances of the media agent and the temporal factors of the relationship with media agents influence the development of these human-computer scripts. In other words, how people interact with a specific technology changes over time and with experience, leading to unique

interaction patterns. This framework allows for a deeper understanding of human-technology interaction beyond the initial novelty phase, explaining why the CASA effect might be less pronounced with older technologies like desktop computers.

Despite these advancements, challenges persist. While Gambino's (2020) extension to the CASA paradigm offers a valuable refinement by acknowledging the development of human-computer scripts, it still leaves some areas open to critique. Primarily, the developed model lacks specificity regarding the mechanisms and timelines involved in human-computer script development. While it rightfully points to factors like social affordances and experience, it does not clarify how these factors interact or the time scales on which they influence user behaviour. For instance, how do users differentiate between the novelty of a new technology and its inherent affordances to shape their initial interactions? Similarly, the model remains vague when the mindless application of social scripts gives way to more reasoned, learned behaviour, offering no empirical grounding for this transition. Without such details, the model struggles to provide concrete predictions about user behaviour, hindering its ability to guide the design of more effective human-AI interactions.

In conclusion, the evolution of AI from a tool to a collaborative teammate marks a significant shift in how humans interact with technology. While early frameworks such as CASA have been instrumental in explaining social interactions with AI, recent advancements reveal the limitations of these models in accounting for the dynamic and context-dependent nature of human-AI relationships. The transition from novelty-driven interactions to more complex, learned behaviours requires a deeper understanding of how people perceive and engage with AI over time. Nuanced research is particularly crucial in the context of HATs, where the development of trust, shared goals, and communication is essential for effective collaboration. The following section will delve into the complexities of defining and understanding HATs, exploring how these teams operate and their unique challenges in balancing human-AI dynamics.

1.2.2 Defining and Understanding Human-AI Teams

As AI becomes more deeply embedded in daily life, from personalised recommendations to complex decision support systems, the need to understand and define HATs grows increasingly critical. While early research attempted to apply established Organizational Psychology theories to HATs, assuming a simple translation of team dynamics, this approach has proven inadequate (Berretta et al., 2023; McNeese et al., 2018). The key distinction lies in whether AI agents are perceived and function as true "teammates" rather than mere tools (Hauptman et al., 2023; Peeters et al., 2021; Rix, 2022; Schelble et al., 2022). This shift from tool to teammate hinges on factors like the AI agent's ability to exhibit qualities associated with human team members, such as predictability, directability, and a suitable level of autonomy (Hauptman et al., 2023; McNeese et al., 2018; Zhang et al., 2021). This realisation has sparked a growing consensus that HATs require specialised research approaches tailored to their unique challenges, moving beyond adapting existing theories and developing frameworks that account for the distinct dynamics of HATs.

Rix (2022) conducted a meta-analysis and proposed a framework that outlines four essential drivers for forming practical HATs: a minimum of two individuals, shared goals, interdependence among team members, and clearly defined roles and functions for both human and AI teammates. Rix (2022) also argues that for these teams to be truly successful, they must function as cohesive social entities, moving beyond a purely transactional relationship. To become a social entity, there needs to be a solid team identity where AI agents are seen as "teammates" rather than just tools, establishing a trust foundation between humans and AI teammates.

1.2.2.1 The Importance of Shared Goals

The concept of shared goals is crucial for fostering a sense of "teamness", which occurs through a shared purpose and a spirit of collaboration that binds team members together (Musick et al., 2021; Schelble et al., 2022). A shared understanding of the team's goals helps to align individual efforts and promotes a more cohesive and coordinated approach to problem-solving (C. Liang et al.,

2019; Sanneman & Shah, 2022). While the importance of shared goals is widely acknowledged, shared goals are associated with complexities and potential challenges in the context of HATs.

Firstly, there is a distinction between "overall shared goals" and "local goals" that might exist at the individual level (Rix, 2022). This distinction is crucial because while a team might have an overarching objective, individual team members, including AI agents, might have specific objectives that could conflict with the team's overall goal. For example, in a resource allocation task, the overall goal might be to optimise resource distribution for the entire team, but individual agents might be programmed to prioritise maximising their own resources (Chiou et al., 2019; Oh et al., 2018).

Secondly, genuine "teamness" requires the presence of shared goals and the perception that all team members, human and AI, benefit somewhat from achieving those goals. If the attainment of a shared goal disproportionately benefits one team member, whether human or AI, it can lead to resentment, distrust, and reduced collaboration among team members (Flathmann et al., 2023; Ong et al., 2012; Schelble et al., 2022).

To effectively implement shared goals, it is important to have explicit goal communication; designers should create AI agents to communicate their goals clearly and explicitly to their human teammates. This transparency can help to alleviate concerns about hidden agendas and foster a sense of shared understanding within the team (Schelble et al., 2022). Rather than assigning rigidly defined roles, HATs should be designed to encourage co-creation, where both human and AI teammates contribute their unique capabilities towards achieving shared goals (Lawton et al., 2023; Merritt & McGee, 2012; Oh et al., 2018).

1.2.2.2 Interdependence and Collaboration

Interdependency is essential for successful HATs because it directly influences team cohesion and, by extension, team performance (Wiethof et al., 2021). When team members perceive their success as intertwined, it strengthens their sense of

shared responsibility and promotes a collaborative spirit. Interdependency is consistent with the concept of "teamness", where a sense of cohesion and belonging within the team is crucial for effective collaboration. However, AI can both enhance and disrupt interdependency within HATs. On the one hand, AI can foster interdependency by shouldering some of the cognitive load typically carried by human team members, allowing for more efficient collaboration and leading to better outcomes (Döppner et al., 2019; Zhou et al., 2017). On the other hand, increasing interdependence can cause issues with transparency and lead to a breakdown in understanding of AI behaviour.

However, while potentially beneficial for efficiency, the increasing autonomy of AI systems can also introduce challenges to interdependency. Johnson et al. (2012) caution that highly autonomous AI systems risk reducing team transparency. If human team members do not understand why an AI agent is taking specific actions, it can lead to mistrust and a breakdown in collaboration (Johnson et al., 2012). This lack of transparency can undermine situation awareness, hindering the team's ability to adapt effectively to changing circumstances.

Therefore, finding the balance between AI autonomy and communication is vital for maintaining interdependency in HATs. As Schelble et al. (2022) highlighted, it can enhance team cognition and trust if AI agents can clearly articulate their goals and align them with the team's objectives. This transparency ensures that human team members feel confident in the AI's actions and understand its contribution to the team, ultimately increasing interdependency.

1.2.2.3 Role Definition and Specialisation

The assignment of unique roles and functions is also critical for successful HAT formation. Derrick and Elson (2019) suggest that roles should be assigned based on each team member's abilities, similar to human teams. This approach implies a division of labour where each team member, whether human or AI, is responsible for tasks best suited to their capabilities. However, Siemon (2022) argues that AI agents should not possess a wide range of skills but instead focus on excelling in one area. This specialised role for AI agents could involve tasks like data analysis, pattern recognition, or task automation, freeing human teammates to focus on

tasks requiring creativity, critical thinking, or interpersonal skills. Oh et al. (2018) found that successful co-creation in HATs does not necessarily require completely distinct roles. Their findings suggest that AI agents and human teammates can effectively collaborate even with some overlap in skills and responsibilities.

Whether roles are unique or shared, research agrees on the importance of clearly defining those roles. Oh et al. (2018) emphasise that even when roles are not entirely distinct, they must still be well-defined to ensure that all team members clearly understand their contributions and how they support the team's goals. This clarity helps to minimise confusion, facilitate coordination, and foster a shared understanding of responsibilities.

1.2.2.4 Social Dynamics

Rix (2022) argues that to function as more than just tools within HATs, AI systems must be designed with social behaviours that encourage the team to act as a cohesive social entity. Rix (2022) suggests that AI should be designed to embody similar social characteristics to those traditionally found in human teams, which includes concepts like team spirit, group cohesion, and a sense of shared identity (Chiocchio & Essiembre, 2009; Hackman, 1987; Kozlowski & Ilgen, 2006). The team's effectiveness is significantly enhanced when AI systems exhibit these social behaviours, suggesting that the success of HATs relies on task-based efficiency and social dynamics within the team (Oh et al., 2018).

Rix (2022) proposes that AI systems should exhibit human-like qualities and engage in relationship-building behaviours within the team to reinforce the idea that it is a part of the social entity of the team rather than a separate entity. Focusing on social behaviours in AI highlights the importance of moving past AI as simply a tool to complete tasks. By designing AI with social behaviours that enable it to function as part of a cohesive team unit, HATs may achieve success and provide a more positive and productive experience for human team members.

Rix (2022) also highlights that researchers often underestimate the complexity of creating practical HATs, particularly in ensuring that machines can perform tasks typically expected of human team members, such as building relationships and

providing understandable explanations for their actions. Rix (2022) suggests that machines require definite configurations to function effectively as teammates.

1.2.2.5 Limitations of Rix's Framework

While Rix (2022) provides a valuable framework for understanding HATs, limitations exist. A primary concern is the lack of a clear and consistent definition of what constitutes a HAT. This ambiguity in defining HATs makes it difficult to compare findings across studies and develop a cohesive understanding of the factors influencing team effectiveness. Finally, Rix (2022) acknowledges the challenge of bridging the gap between research conducted in controlled environments and the complexities of real-world HAT applications. This gap raises questions about the ecological validity of current research findings and how these findings can be generalised to real-world settings where AI systems must operate in dynamic and unpredictable contexts.

Following this, Rix's (2022) paper also has methodological limitations. The paper primarily relies on existing literature, particularly from Information Systems, and lacks original empirical validation of its proposed framework. This reliance on a limited scope of literature without empirical testing raises concerns about the generalizability of the findings and the framework's applicability to real-world settings. While Rix's (2022) work offers foundational insights into the methodological challenges of understanding HATs, more intricate dynamics unfold in these interactions. In the next section, we delve into the complexities of designing AI systems for HATs, focusing on error tendencies, mental models, and team dynamics.

1.2.3 Accuracy and Confidence in HATs

When designing AI systems for HATs, presenting information to facilitate the human teammate's understanding of the AI's behaviour, particularly its error tendencies, can be effective for building accurate mental models. As previously discussed, this understanding allows humans to develop an accurate mental model of the AI's capabilities and limitations, enabling informed decisions about reliance and collaboration with the AI system (Bansal et al., 2019; Grimes et al., 2021).

This process of developing an understanding of the AI partner connects to the concept of situation awareness (SA) from human factors research (Sanneman & Shah, 2022). SA involves an individual's perception and comprehension of their surroundings, including how information might unfold. In HATs, sufficient SA about the AI's behaviour, particularly around its error tendencies, is crucial for making informed decisions about when to trust or override the AI (Sanneman & Shah, 2022).

Understanding the nuances of AI properties, like accuracy and confidence, in HATs is crucial for effective teaming (Bansal et al., 2021). While increased AI accuracy might seem intuitively linked to better HAT performance, this is not always true, and the relationship is complex. Despite exhibiting lower individual accuracy, an AI that presented low-confidence in its output, can improve team performance with increased accuracy in specific situations (Bansal et al., 2021). These findings occur because lower AI confidence allows for a more accurate mental model of the AI teammate's behaviour to be formed by the human teammate. This improved mental model gives the human teammate a better understanding of the AI's error rate and tendencies, ultimately allowing for more informed decisions about when to trust or override the AI's recommendations. These findings assist in understanding the importance of accuracy and confidence during the AI design phase, prioritising team-based utility and outcomes over individual AI performance metrics.

Bansal et al., (2019a) emphasised that having a highly accurate AI might not be enough; the human teammate needs to understand how the AI arrives at its conclusions and where its potential pitfalls lie (Bansal et al., 2019b). Providing the human teammate with clear and concise information about the AI's error boundaries, particularly as they relate to parsimony and stochasticity, can contribute to developing this shared understanding and allow for more effective human-AI collaboration.

In this context, parsimony refers to the simplicity of representing areas where the AI is prone to making mistakes. A more parsimonious error boundary is more manageable for humans to understand and remember (Bansal et al., 2019a). For

instance, an error boundary that can be explained with fewer features or more straightforward rules is more parsimonious and, therefore, more accessible for humans to integrate into their mental models.

Stochasticity refers to the consistency of the AI's errors within its error boundary. A non-stochastic error boundary implies that the AI's errors are predictable and occur consistently for specific types of inputs, making it easier for humans to learn the pattern. On the other hand, a stochastic error boundary makes the AI's errors less predictable and, therefore, more difficult for the human teammate to anticipate. This unpredictability can hinder the human's ability to adjust their treatment of the AI accordingly.

Bansal et al., (2019a) highlighted that AI models with parsimonious and non-stochastic error boundaries are more accessible for humans to understand, leading to more accurate mental models. This enhanced understanding can lead to more effective collaboration and improved team performance in HATs.

1.2.3.1 AI Teammate Vs Human Teammate

Another complex aspect of HAT dynamics is the differences in how human teammates perceive and interact with AI teammates compared to human teammates. This difference in perception can significantly impact team decisions and outcomes, particularly in situations requiring trust and collaboration. For example, research indicates that when faced with the choice of saving either an AI or a human teammate in defensive team games, participants often prioritise the "best outcome" when saving the AI, as opposed to a "protect the teammate" rationale when saving a human (Ong et al., 2012). These rationales suggest that humans might not instinctively afford the same level of care or value to AI teammates compared to human teammates.

Further illustrating this point, research points to the tendency for AI teammates to be unfairly blamed for team failures, a phenomenon not typically observed with human counterparts (Merritt et al., 2011). The unfair blame suggests a bias against AI teammates, where humans are more likely to attribute blame to the AI even when it is unjustified (Jones-Jang & Park, 2023). This difference in treatment

could be attributed to the challenges humans face in forming accurate and robust mental models of AI behaviour or to the social pressures of pleasing human team members.

1.2.3.2 User-Centric HATs

Maintaining a user-centric approach is crucial when considering the integration of AI teammates within human teams. While human teams organically adapt and perceive changes in roles or responsibilities through explicit communication and implicit cues, these subtle signals are often lost in translation when AI teammates are involved. Bansal et al. (2019b) explain the potential problems of neglecting user-centricity, particularly when AI teammates undergo software updates without proper communication with their human teammates. AI updates might improve the AI's performance, but they can negatively impact overall team performance if these updates are not transparently communicated to the human team members. This performance/compatibility trade-off stems from human teammates developing mental models of their AI counterparts' capabilities and limitations through experience. When an AI teammate's behaviour changes due to an uncommunicated update, it disrupts the established mental model, leading to confusion, mistrust, and, ultimately, a decline in team performance (Bansal et al., 2019b).

Finally, Berretta et al., (2023) conducted a scoping review highlighting the necessity for a human-centric approach to HATs. The review emphasises the importance of a sociotechnical approach to successfully developing AI agents from tools to rounded teammates. Berretta et al. (2023) argues that, as AI systems become more sophisticated and integrated into collaborative work environments, there is a growing need to shift the focus from a purely technology-centric perspective to one that prioritises the human element in these teams. They propose that successfully developing AI agents into true teammates, rather than just tools, requires a sociotechnical approach, which acknowledges the interconnected nature of social and technical systems.

Berretta et al. (2023) also emphasised the need for joint optimisation, where both the AI system's capabilities and the human teammate's needs and experiences are

considered and designed in tandem. This resonates with Bansal et al's. (2019b) work on the importance of transparent communication surrounding AI updates. By keeping human teammates informed about changes in their AI counterparts, developers can ensure that the HAT's social (human) and technical (AI) elements are aligned, fostering a more successful collaborative partnership.

While research offers valuable insights into human-AI teammate relationships, it also exhibits certain limitations that warrant consideration. One notable limitation concerns the generalizability of findings. Many studies, including those investigating decision-making in defensive team games, employ specific and potentially artificial task environments. This specificity raises concerns about whether the observed behaviours and dynamics are generalisable to more complex, real-world collaborative settings.

Furthermore, there are limitations to measuring and operationalising key concepts. While research emphasises the importance of mental models in shaping HAT interactions, accurately assessing these internal representations poses a significant challenge. Relying solely on behavioural observations or self-reported data, as is common in the discussed research, might not fully capture the complexity and nuance of how humans mentally represent and interact with their AI counterparts.

1.2.4 Conclusions of Human-AI Teams

In conclusion, exploring HATs reveals promising insights and significant challenges. While the initial application of organisational psychology theories provided a helpful starting point, it has become evident that these theories alone cannot fully address the complexities of HATs. Rix's (2022) framework highlights critical drivers such as shared goals, interdependency, unique roles, and social dynamics, which are essential for forming practical HATs. However, the field still grapples with defining HATs, integrating insights from human team dynamics, and ensuring research translates to real-world applications.

Further research must focus on understanding the more sophisticated variables that affect HAT performance, including the impact of AI autonomy, accuracy, confidence, and user-centric design. By addressing these issues and adopting a

more human-centric approach, we can advance the development of AI systems that function as truly effective and integrated team members. Despite the extensive literature on HATs, specific gaps persist, such as a lack of focus on more complex team dynamics, the impact of intricate variables such as performance, the adaptability of AI and a human-centred approach. Among these challenges, trust emerges as a pivotal factor influencing the success of HATs, underpinning both interpersonal interactions and the integration of AI into team settings. The following section delves into the concept of trust, exploring its complexity and significance within the context of HATs.

1.3 Trust and AI

1.3.1 Trust

One critical aspect of HATs that requires deeper exploration is trust. Trust, a complex concept, has been extensively studied across various disciplines and situations. Interpersonal trust, for instance, involves confidence in an individual's integrity, reliability, and fairness (Rotter, 1980). It is crucial for maintaining healthy personal and professional relationships, typically developed through consistent, honest, and supportive interactions. Interpersonal trust facilitates effective communication and conflict resolution, making it a foundational element of successful human interactions (Rotter, 1980).

In contrast, organisational trust pertains to confidence in an organisation's fairness, integrity, and fulfilment of commitments (Mayer et al., 1995). Organisational trust influences factors such as reputation, employee engagement, and performance. It is built through transparent communication, ethical practices, and consistent actions, leading to increased loyalty, lower turnover, and a motivated workforce (Bornstein et al., 2016; McAllister, 1995; Shockley-Zalabak et al., 2000). Within organisations, team trust refers to the confidence and reliance among team members. Organisational team trust is essential for effective teamwork and collaboration and for creating a safe environment for sharing and risk-taking, typically nurtured through shared experiences, mutual respect, and open communication (Costa et al., 2018). Higher team trust is associated with improved problem-solving, creativity, and overall team performance (Costa et al.,

2018). Although these types of trust share common elements, each is distinct and requires precise definitions and measurements (Ulfert et al., 2023).

From a psychological perspective, trust operates not just as a behavioural act but also as a psychological state. Krueger et al., (2007) explores the neural mechanisms behind conditional (earned) and unconditional (passive) trust using hyper fMRI. Their findings highlight that the paracingulate cortex is crucial for building trust by inferring intentions to predict behaviour. Conditional trust activates the ventral tegmental area, associated with reward evaluation, while unconditional trust activates the septal area, linked to social attachment. Additionally, Krueger et al., (2007) and Dimoka (2010) demonstrate that trust and distrust involve distinct neural mechanisms. Trust is related to brain areas involved in reward prediction and social attachment, while distrust correlates to areas associated with intense emotions and fear of loss. These insights suggest that our brain differentiates between trust and distrust, influencing social interactions and responses to others' behaviours.

1.3.1.1 Defining Trust

Given the varied understandings of trust, the concept often falls prey to the jangle fallacy (Freeman & Kelley, 1928), where similar terms might have different meanings across disciplines. The multidisciplinary nature of trust adds to the complexity of consistently defining it. Despite these challenges, research across multiple fields has identified some recurring themes.

Rousseau et al., (1998) conducted a meta-analysis across diverse disciplines and found several common elements of trust. One key theme is the willingness to be vulnerable based on expectations of positive outcomes from others' actions. A willingness to be vulnerable underscores that trust fundamentally involves taking risks while anticipating favourable results. Additionally, trust is not just a behavioural act or decision but a psychological state, a mental and emotional readiness to expose oneself to potential risks due to positive expectations about another's intentions or behaviour. Trust also requires specific conditions to exist, such as the presence of risk and interdependence. Risk introduces the possibility of loss or harm, making trust significant. At the same time, interdependence

means that one party's goals cannot be achieved without relying on another, thus making trust essential for cooperation.

Overall, Rousseau et al., (1998) argue that, despite different disciplinary perspectives, there is broad agreement on the core components of trust. This consensus suggests that the fundamental aspects of willing vulnerability and positive expectations are consistent in economic transactions, social relationships, or institutional settings. The proposed definition is:

> "Trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another". - (Rousseau et al., 1998)

While Rousseau's definition of trust is valuable and consistent across various fields, it is essential to recognise its limitations. The paper, while comprehensive, may not fully account for recent developments or changes in how trust is perceived, particularly in the context of AI and technology. As trust dynamics in HATs may differ from traditional scenarios, exploring contemporary literature on trust in technology and AI is crucial.

1.3.1.2 AI Agents and Trust

The concept becomes increasingly intricate when examining trust in AI due to its unique characteristics compared to trust in human relationships. Trust in technology often intersects with the concept of reliability, creating a dynamic where the two are sometimes conflated. For instance, users may not explicitly state that they "trust" their mobile phones, yet their heavy reliance on them for critical daily functions suggests an implicit form of trust. However, this reliance can quickly erode if the technology fails or performs inconsistently, as reliability is a cornerstone of trust in technological contexts (Mcknight et al., 2011). Unlike interpersonal trust, which develops through mutual interactions and shared experiences, trust in AI often hinges on performance metrics, transparency, and the system's ability to meet user expectations reliably over time (Hoff & Bashir, 2015).

The relationship between trust and technology and automation becomes more intricate in high-stakes environments. For example, pilots trusting autopilot systems must navigate a delicate balance between reliance and vigilance, which parallels interpersonal trust, where complete reliance can coexist with moments of doubt (Trösterer et al., 2017). This complexity arises because trust in automation systems is shaped by their functional capabilities and user perceptions of the system's autonomy, transparency, and ability to adapt to unforeseen circumstances. Users may simultaneously "trust" an AI system for its precision while "distrusting" it in situations requiring contextual judgment or ethical reasoning, highlighting parallels with the complicated nature of human trust (Hancock et al., 2011; Lee & See, 2004).

Building on this, Lee and See (2004) emphasised the importance of transparently designing interfaces that communicate functionality, reliability, and limitations. This fosters trust by equipping users with clear expectations of system performance. These principles are crucial in ensuring that trust is calibrated, neither overestimated nor underestimated, based on the AI's capabilities and the context in which it operates.

The relationship between trust and AI can also change depending on the context of the situation it is used in. For instance, trust may hinge on a medical diagnostic AI system's perceived accuracy and adherence to professional standards. In contrast, a customer service chatbot may rely more on conversational fluidity and responsiveness (Lee & See, 2004). These varying expectations demonstrate that trust in AI is not monolithic but deeply context-dependent, influenced by the task, environment, and user perception (Glikson & Woolley, 2020). The integration of transparent communication and practical design elements, as advocated by Lee and See (2004), becomes vital for fostering trust in these diverse applications. By understanding the interplay of these factors, we can move closer to designing AI systems that create appropriate and productive human-AI trust dynamics.

As AI becomes prevalent, we will likely see a growing emphasis on trust in AI as a unique type of trust. The following section will delve into the various aspects of trust in technology, including the ideal levels and how team dynamics can affect

this trust. Understanding these concepts will be crucial for navigating the complexities of human-AI interactions and ensuring effective collaboration.

1.3.1.3 Trust in AI vs Human Trust

Trust in AI systems differs fundamentally from trust between humans, particularly in its development, emotional components, and reliance on system reliability. Unlike human trust, which grows through emotional connections, social norms, and shared experiences (Riegelsberger et al., 2005; Schniter et al., 2020; Weiss et al., 2021), trust in AI often hinges on the system's performance, predictability, and user expectations. Lee and See (2004) emphasise that trust in automation is a dynamic process influenced by past experiences, current performance, and perceived reliability. This means that while reliability is critical, trust depends on the user's interpretation of the AI's behaviour and its alignment with expectations.

Human trust tends to recover from violations through emotional appeals, apologies, and corrective actions (Lewicki & Brinsfield, 2017; Sharma et al., 2023). In contrast, trust in AI lacks these interpersonal mechanisms and is more transactional. For instance, Glikson and Woolley (2020) found that trust in AI usually begins at a low baseline and increases with hands-on experience. However, this trajectory varies depending on the form of the AI. While virtual AI often sees trust decline over time, robotic AI may evoke mixed or negative emotions due to its anthropomorphic features. Emotional trust in AI is thus shaped by its representation, with anthropomorphism enhancing trust in virtual settings but potentially causing discomfort in robotic systems when their capabilities fail to match their human-like appearance.

Research on brain imaging and trust in AI is sparse and often miscited. Contrary to some claims, studies do not show that trust in AI replicates the brain activation patterns seen with trust in humans (Krop et al., 2024; Wienrich et al., 2021). Montag et al., (2023) found no significant neurostructural correlations between trust in humans and AI. Their study, involving self-reports and MRI brain imaging, revealed that trust in humans was associated with specific brain regions (striatal-thalamic and prefrontal areas), indicating a neurostructural basis for human trust.

In contrast, trust in AI did not correspond to specific brain regions, suggesting that trust in AI may not have a direct neurostructural basis.

Montag et al., (2024) further investigated the neurocognitive mechanisms underlying trust in humans and AI across different cultures. They found that trust in humans and AI are primarily distinct constructs, with cultural context influencing trust levels. In Germany, trust in humans was higher than trust in AI, and the gap was more pronounced compared to Singapore, where there was a moderate correlation between the two types of trust. Personality traits also influenced trust; neuroticism was associated with greater fear of AI, while conscientiousness and agreeableness correlated with lower fear and higher acceptance of AI.

These studies suggest that trust in AI and humans is processed differently, behaviourally and neurostructurally. While human trust appears to have a neurostructural basis, trust in AI does not show a similar linkage. Limitations of these studies include reliance on self-reports, cross-sectional design, and cultural differences in interpreting trust. Moreover, framing AI interactions, such as through anthropomorphism or embodiment, might influence trust levels, potentially mirroring human-human interactions. Future research should explore these aspects further.

Jung et al., (2019) examined human trust in machine agents through behaviour and EEG activity. Their study revealed that the AI agent's appearance, voice, movements, and risk-taking traits influence trust. External cues drove explicit judgments of human likeness, while implicit trust, measured by intervention frequency in agent decisions, was affected by the agent's risk-taking personality. EEG data showed significant changes in theta band power in the frontal-central region of the brain following an agent's decision, corresponding with trust fluctuations. The study highlighted that trust develops dynamically through interaction and that AI agent performance and human-like characteristics influence it. However, its controlled experimental setting may limit ecological validity, and the specific traits studied might not fully capture real-world complexities in HATs. To further understand how these characteristics influence

trust, it is essential to explore the concept of anthropomorphism, particularly how human-like features in AI systems shape user perceptions and interactions.

1.3.1.4 Anthropomorphism and Trust

Anthropomorphism is the process of attributing human-like qualities to non-human agents (Guthrie, 1997). Anthropomorphism can influence how users perceive AI systems, sometimes enhancing trust by evoking familiarity while leading to discomfort or scepticism in other cases. Understanding how anthropomorphism interacts with trust dynamics is critical for designing effective HATs. Designing AI with anthropomorphic features can be beneficial because it allows humans to project their social schemas onto AI systems (Fussell et al., 2008). If designers created an AI agent with very human features that look realistically trustworthy, a human would likely feel more trust toward this AI agent than toward one lacking anthropomorphism (Glikson & Woolley, 2020).

There are two main ways in which anthropomorphism is presented: through physical attributes and social behaviours (Duffy, 2003). Researchers recommend that for robots to encourage an anthropomorphic projection, they should possess not only human-like facial features but also limbs similar to those of a human and other human-like features, including movement (Złotowski et al., 2015). However, it is also important for the robot to have social behaviours that are typical of a human, such as facial expressions, gestures, and engagement (Duffy, 2003).

Although anthropomorphism can have powerful effects, it is essential to note that there can be limitations to its success. The first of these was developed by Mori (1970) (Mori, 2012) and is called The Uncanny Valley effect. Mori posed that anthropomorphism initially leads to increased empathy and affinity among humans. However, there is a dramatic drop in affinity when robots look and act almost human but are not entirely convincing, resulting in eeriness and distrust. Furthermore, it is crucial to understand the implications of using anthropomorphic AI due to its priming effects on other AI technology and how this may impact trust calibration. Anthropomorphic priming can shape user expectations and interactions with AI, influencing trust levels across different AI applications (Zanatto et al., 2016).

Research has examined how anthropomorphism can impact trust in AI (Glikson & Woolley, 2020; Roy & Naidoo, 2021; Seymour & Van Kleek, 2021; Troshani et al., 2021; Waytz et al., 2014), with findings indicating AI representation (robots, virtual agents, embedded systems) and perceived machine intelligence are key antecedents to trust development. In addition, cognitive trust is influenced by AI's tangibility, transparency, reliability, and immediacy behaviours. Emotional trust is notably affected by anthropomorphism, and when there is little anthropomorphism, there is little emotional trust compared to when anthropomorphised AI. It is further enhanced by AI's capability to exhibit human-like qualities and behaviours (Glikson & Woolley, 2020).

However, anthropomorphism may not be the most important feature when improving HAI in specific contexts. Pelau et al., (2021) found that anthropomorphic traits alone do not ensure acceptance; empathy and interaction quality play crucial mediating roles. Empathy in AI significantly influences consumer acceptance. AI devices that show understanding and care are more readily accepted. The impact of empathy and understanding in AI design highlights the need for AI to mimic human-like emotional responses. These findings suggest a shift towards developing AI with advanced emotional intelligence capabilities and not just basic levels of anthropomorphism.

Troshani et al., (2021) also explored the role of anthropomorphism in relation to trust in AI. The main findings suggest that AI's human-like features can enhance and undermine trust, depending on their implementation and user perceptions. Positive experiences with AI's human-like interactions can boost trust, while excessive human likeness may trigger discomfort, which could be related to the Uncanny Valley (Mori, 2012). Troshani et al., (2021) also found contextual influences, where the context of AI use (e.g., healthcare vs. customer service) significantly affects trust, so designing AI for specific roles is essential. This work highlights the need for AI systems that balance human-like interaction with transparency and user control.

These findings on anthropomorphism highlight its dual potential to enhance or undermine trust, underscoring the importance of careful design in calibrating trust.

By aligning human-like features with system capabilities and contextual needs, AI designers can help users appropriately balance their reliance on these systems. The following section will delve deeper into trust calibration, examining strategies for managing overtrust and undertrust in AI interactions to promote effective and safe human-AI collaborations.

1.3.1.5 Trust Calibration

Understanding trust in AI is crucial due to its complexity and the need for calibrated trust in AI systems. Calibrated trust is essential to address the problems of excessive trust (over-reliance) and insufficient trust (under-reliance) (de Visser et al., 2020; Ingram et al., 2021; Wang et al., 2016). Excessive trust occurs when users rely too heavily on AI recommendations without adequate scepticism or verification, potentially leading to failures. In contrast, insufficient trust arises when users disregard or undervalue AI's capabilities and recommendations, which can result in the AI being underutilised or ignored. Effective trust calibration ensures users can appropriately balance their trust in AI, leveraging its capabilities while maintaining necessary scrutiny.

Robinette et al., (2016) conducted an experiment highlighting overt trust issues when using embodied AI (robots) in an emergency. The experiment involved participants who followed a robot's guidance during a simulated emergency despite the robot displaying unreliable behaviour in previous non-emergency tasks. Surprisingly, all participants followed the robot's emergency instructions, even those who had observed the robot's poor performance in navigation tasks immediately prior. This phenomenon occurred across various conditions, including when the robot malfunctioned or provided no rational guidance. A significant portion of participants rationalised their trust in the robot based on its designated role as an "emergency guide", despite witnessing its earlier failures. Robinette et al's., (2016) work highlights the critical importance of designing robots that can communicate their operational status and limitations to prevent overtrust. Robinette et al., (2016) suggest that AI needs mechanisms enabling robots to decline trust or redirect humans to more reliable sources of assistance when they are not functioning optimally. The study highlights a potentially dangerous level of

human overtrust in robots during emergencies, emphasising the need to consider human-robot interaction dynamics in designing emergency response robots.

We can use meaningful explainability, adaptive communication, and continuous trust repair to address the overtrust issue highlighted by Robinette et al., (2016). Meaningful explainability goes beyond simple transparency and necessitates that AI systems actively address potential over-reliance, particularly in high-stake scenarios where users might rationalise trust based on the AI's role (Bansal et al., 2021; Lopez et al., 2023; Ulfert et al., 2023). For instance, AI could explicitly acknowledge its limitations by stating its inconsistent performance in similar tasks and encourage human confirmation with another source. Meaningful explainability would promote a more critical evaluation from humans. Adaptive trust calibration involves AI dynamically adjusting communication based on the user's perceived trust (Chen et al., 2023). If an AI agent senses hesitation, it could increase explanation granularity, quantify its confidence level, or even proactively defer trust by suggesting alternative courses of action or human consultation with other, more reliable sources.

Lastly, trust repair should be an ongoing dialogue where AI demonstrates continuous learning (Kim & Song, 2021; Schelble et al., 2024). Trust repair can be achieved through proactive self-evaluation and communication of performance, highlighting improvements based on past experiences, and actively seeking feedback from human teammates. By incorporating these principles, AI can shift from being mindlessly followed to becoming trusted partners that earn and maintain calibrated trust through dynamic, transparent interaction.

1.3.2 Explainability and Trust

Transparency emerges as a key factor in enhancing trust calibration in AI systems. As AI technology advances, there is a growing need to shift from traditional black-box methods, where the decision-making process remains opaque, to more transparent and explainable AI systems (XAI) (Adadi & Berrada, 2018). This shift stems from the recognition that the rapid adoption of AI, especially in sensitive areas like healthcare, finance, and legal applications, requires robust, trustworthy, and understandable systems for human users.

Research emphasises that achieving calibrated trust becomes a significant challenge without clearly understanding how AI systems arrive at their decisions. These concerns have driven efforts to develop methods that provide insights into AI decision-making processes. Experts recognise that achieving this requires an interdisciplinary approach, combining AI with cognitive and social sciences. (Adadi & Berrada, 2018; Endsley, 2023; Guidotti et al., 2018; Kim et al., 2023).

While potentially highly accurate, Black-box AI models offer limited insights into their internal workings, making it difficult for human users to assess their reliability and make informed judgments about when to trust their outputs. This opacity becomes particularly problematic in high-stakes scenarios where AI decisions can have significant consequences. XAI addresses these concerns by employing various methods to provide insights into AI decision-making processes (Adadi & Berrada, 2018). Adadi & Berrada (2018) broadly categorise these methods as intrinsic, which features built-in explainability, or post-hoc, which provides explanations generated after the AI has made a decision. There is also growing interest in model-agnostic XAI methods, which aim to understand the predictive responses of various AI models, regardless of their specific architectures, to broaden their applicability (Adadi & Berrada, 2018).

The success of XAI hinges on aligning AI explanations with human cognitive processes (Bansal et al., 2021). In other words, XAI must present information in a way that is understandable and meaningful to human users, considering their cognitive limitations and biases (Förster et al., 2020; Ingram et al., 2021; Sanneman & Shah, 2022). Simply providing technical details about an AI model's internal workings is unlikely to foster trust or understanding. Instead, XAI must strive to bridge the gap between AI technology and human cognition by offering relevant, interpretable, and actionable explanations from a human perspective. More research is needed on how humans perceive, understand, and trust AI explanations and how these explanations could impact trust calibration in HATs.

1.3.3 Conclusions About Trust in AI

Trust is a pivotal element in the success of HATs and the broader integration of AI into human decision-making processes. This review has demonstrated that trust in

AI systems is complicated, involving reliability alongside emotional, psychological, and contextual factors. Unlike trust in humans, trust in AI often hinges on transparency, explainability, and calibrated interactions, which are critical for preventing over-reliance or under-reliance on AI systems. The dynamic nature of trust, shaped by past experiences and evolving perceptions of AI's performance, further emphasises the importance of designing AI systems that are adaptable, explainable, and aligned with user expectations.

Ultimately, this section underscores the need for a nuanced, interdisciplinary approach to understanding and building trust in AI. By focusing on transparency, user-centred design, and the continuous calibration of trust, AI can evolve from being perceived as a tool to becoming reliable and trustworthy teammates in routine and high-stakes environments. Further research is essential to explore the long-term dynamics of trust in AI, particularly in real-world applications with the highest stakes.

1.4 Social Intelligence

1.4.1 Social Intelligence in Humans

One facet of human behaviour that could be useful for successfully anthropomorphising is SI. The concept of SI traces back to 1920, originating with Thorndike's classification of intelligence, which posited three types: abstract, mechanical, and social (Thorndike, 1920). Thorndike defined social intelligence as

> "the ability to understand and manage men and women, boys and girls — to act wisely in human relations" (p. 228).

Nevertheless, the most widely recognised definition hails from Vernon (Vernon, 1933), encapsulating it as

> "The ability to get along with people in general, social technique or ease in society, knowledge of social matters, susceptibility to stimuli from other members of a group, as well as insight into the temporary moods or underlying personality traits of strangers" (p. 44).

Currently, differing theories persist regarding the accurate definition and measurement of SI (Weis & Süß, 2005), but a consensus generally divides SI into five core categories: social understanding, social memory, social perception, social creativity, and social knowledge (Kihlstrom & Cantor, 2000).

One of the reasons SI could be promising for improving HATs is due to its impact on human teams. Woolley et al., (2010) found that higher scores on an SI scale positively correlated with higher scores and levels of collective intelligence, suggesting that SI improves a team's overall ability to perform various tasks and positively impacts team performance.

1.4.2 Social Intelligence In AI

SI is fundamental to interpreting and responding to social phenomena, a skill underpinning meaningful human interactions. In AI, SI involves creating computational systems capable of sensing, perceiving, reasoning about, learning from, and responding to other human or artificial agents' affective, behavioural, and cognitive constructs. To become socially intelligent, AI must achieve social perception, which involves extracting relevant information from sensory stimuli, social knowledge encompassing explicit and procedural norms, and social memory to maintain consistency. Furthermore, social reasoning enables AI to interpret stimuli and infer intentions, while social creativity allows counterfactual reasoning about social situations, akin to the human capacity for "theory of mind". Lastly, social interaction entails engaging dynamically with others in co-regulated patterns, a core requirement for collaborative settings (Lee et al., 2024; Mathur et al., 2024).

The context in which SI operates shapes its application. Social settings such as homes or hospitals dictate norms for behaviour, while the roles and attributes of actors, whether human or machine, influence interaction patterns. Embodiment and anthropomorphism further affect the dynamics, as AI systems range from disembodied virtual agents to physically embodied robots, each eliciting different user responses (Mathur et al., 2024). Interaction structures involving individual agents, pairs, or groups add complexity to these social exchanges, which unfold across diverse periods, from split-second decisions to relationships that evolve

over the years (Sufyan et al., 2024). These factors highlight the importance of multidimensionality and contextual adaptability in implementing SI.

Research into Social-AI has advanced significantly in recent years, driven by natural language processing, machine learning, and robotics progress. While early studies relied on rule-based systems for modelling social behaviours, modern approaches leverage machine learning and deep learning to predict and generate social phenomena using large datasets annotated with ground truth labels (Satyanarayana et al., 2018). Substantial focus has been placed on modelling affective phenomena such as emotions and sentiments and social behaviours like cooperation and competition. Recent developments have also explored the use of game-theoretic and probabilistic frameworks to enhance social reasoning. Notably, large language models (LLMs) have been assessed for their ability to replicate SI competencies, showing promise in linguistic understanding but revealing limitations in adapting to complex real-world contexts (Bainbridge et al., 2011; Mathur et al., 2024; Satyanarayana et al., 2018). Although significant progress has been made in modelling social phenomena in controlled environments, real-world social interactions' inherent ambiguity and richness remain challenging.

Developing socially intelligent AI is fraught with technical challenges. One key issue lies in the ambiguity of social constructs, which are inherently subjective and context-dependent (Mathur et al., 2024; Mirnig et al., 2017). For example, constructs like trust and empathy often have no clear-cut measurements, leading to interpretive misalignments between users and AI systems (Ulfert et al., 2023).

Another challenge is the subtlety of social signals, often expressed through nuanced, multimodal cues such as gestures, tone, and facial expressions. The complexity of interactions is further amplified by the need to account for multiple perspectives as each actor's perceptions, roles, and experiences evolve dynamically over time. Finally, socially intelligent agents must demonstrate adaptability, learning from implicit and explicit social signals to build a shared social reality with their human counterparts. Addressing these challenges requires robust frameworks that integrate ethical considerations to ensure socially

intelligent AI aligns with human values and promotes trust (Bainbridge et al., 1994; Sterelny, 2007; Sufyan et al., 2024).

The foundational aspects of SI in humans, such as social understanding, perception, and memory, provide a roadmap for developing socially intelligent AI. However, while humans intuitively grasp social norms and adapt based on context, programming AI to mimic these intricate behaviours remains a tough challenge. For example, while humans can adjust their behaviour based on situational appropriateness, such as avoiding humour in serious contexts, AI struggles to replicate such adaptability (Mathur et al., 2024).

The need for AI and robots to contain a level of SI is not a new concept; Dautenhahn (1995) discussed how there needed to be a shift from technological intelligence, domain-specific technical abilities that robots or AI possess, to a more general SI where robots/AI effectively communicate and cooperate with humans and other robots. Dautenhahn (1995), highlights the need for SI to provide AI with the skills necessary for interaction and collaboration, especially in scenarios that expect robots to support humans in roles involving significant social contact.

Another intricate facet of SI is its context-sensitive character. Humans typically possess a well-developed grasp of appropriateness, exemplified by behaviours like refraining from laughter during solemn occasions, and may not react favourably to a humorous AI in a serious context (Syrdal et al., 2006) AI's deficiency in social behaviour could lead to heightened scrutiny of its performance by human agents, especially when compared to the performance of their human counterparts. Previous research suggests that SI in AI can influence trust calibration (Williams et al., 2022). The mimicry of human behaviour is a fundamental element of SI (Chartrand & Bargh, 1999).

Although socially intelligent agents have the potential to improve team dynamics, providing AI with the skills necessary to appear socially intelligent is a complicated process. In addition, there is not an overwhelming amount of research into how humans will respond to these AI agents. There is a gap in the literature when

investigating the impacts of social agents in HATs, and we must understand the impact of artificial social agents before implementing them in the workplace.

1.5 Summary and Proposed Research

Substantial research exists on HATs, Trust, Anthropomorphism, and SI. However, there is a limited exploration of how these elements interact. This thesis argues that fostering successful HATs requires a human-centric approach emphasising calibrated trust, appropriate anthropomorphism, and developing socially intelligent AI agents. This argument is supported by the lack of consensus surrounding optimal implementation strategies for these factors despite their acknowledged importance.

While the HAT literature significantly emphasises defining team dynamics, the complex ways trust functions in these collaborations are often overlooked. Though many studies recognise the importance of anthropomorphism, there is no consensus on its ideal implementation or potential drawbacks. Additionally, research on the impacts of socially intelligent agents, particularly in workplace settings, remains limited, highlighting a crucial gap in our understanding before widespread implementation.

Future research should investigate the impacts of AI teammate reliability and behaviour on trust calibration. Studies could examine how initial trust formation and potential trust breaches influence the trajectory of human-AI collaboration. This research could employ methods like measuring changes in trust levels and collaboration quality over time, using subjective measures (e.g., questionnaires) and objective measures (e.g., task performance metrics). Understanding these dynamics is essential for designing AI systems that foster sustainable and robust trust relationships with human teammates.

Furthermore, the ethical implications of anthropomorphism and the potential for AI to manipulate or deceive users necessitate careful consideration. As AI systems take on increasingly social roles, future research should explore guidelines and safeguards to ensure the responsible development and deployment of HATs. This could involve establishing ethical frameworks for designing AI interactions,

particularly concerning transparency and user autonomy. Addressing these issues is crucial for fostering public trust and acceptance of AI in collaborative settings.

This thesis will first explore trust by conducting a bibliometric analysis of trust literature across time, revealing how trust research has evolved and current research trends. It will then transition to experimental chapters which examine different elements of AI design, these include anthropomorphism, the use of Emojis as a form of Emotional Intelligence and socially aligned adaptive AI to examining how different variables impact trust and performance in HATs.

Chapter 2 Understanding Trust Research and the Need for a Multidisciplinary Approach - A bibliometric analysis of trust research from 1922-2021.

In this chapter, we aim to build upon the theoretical exploration of trust in Chapter 1. Chapter 2 conducts a comprehensive bibliometric analysis to monitor the evolution of trust research across academic disciplines from 1922 to 2021. By employing bibliographic coupling and keyword co-occurrence methods, this chapter seeks to reveal the foundational contributions, research trends, and interdisciplinary intersections within the field. Bibliometric analysis is a set of methods used to quantitatively analyse academic literature, providing insights into the patterns, impacts, and trends within a specific field or across multiple disciplines. This type of analysis uses various statistical and mathematical techniques to measure and evaluate the research output and its influence based on bibliographic data, such as publication counts, citation counts, and journal relationships. Using bibliometric analysis, we can identify research trends and map scientific fields. We decided to conduct a bibliometric analysis of trust research to gain new insights into trust research and develop an understanding of this unique research area.

This chapter discusses the ideas and definitions of trust, the importance of different disciplines collaborating to understand trust, and the idea of bibliometric analysis. We then analyse 39,628 documents spanning the years 1922 to 2021. We focus on the most cited papers on trust, a Bibliographic coupling of journals and a keyword co-occurrence to provide different insights of the fields. From the analysis, we draw conclusions about the scientific mapping of trust research. This foundational understanding is integral as we transition to the experimental components of the thesis, where we will assess trust within applied experimental contexts. Currently, this work is under review at the Journal of Trust Research and we are hopeful that by the time of publishing this thesis, it will be published. I also presented this research at a University of Glasgow Workshop Morgan Bailey & Frank Pollick, Can We Trust 'Trust'? An Overview of Trust Concepts and

Definitions, Multidisciplinary Workshop on Cyber-Physical Systems (CPS) at The University of Glasgow. Glasgow, Scotland. April 16, 2024.

2.1 Introduction

Trust is a fundamental concept in human interactions, and its study has garnered increasing attention from researchers across various disciplines over the years (Botsman, 2015; Bottery, 2003; Glikson & Woolley, 2020; Hendriks et al., 2021; Hoff & Bashir, 2015; Jacovi et al., 2021; Rompf, 2015; Weiss et al., 2021). As interest in trust grows, it is essential to understand how researchers from diverse disciplines have delved into trust and how they examine trust's presence and implications in a wide array of contexts. When talking about trust, psychologists are likely to refer to trust in interpersonal relationships (Rotter, 1980), where they explore the dynamics of trust between individuals. Interpersonal trust refers to the confidence one individual has in another person's reliability, integrity, and benevolence. It encompasses the belief that the other person will act in a supportive, honest, and considerate way in one's best interests (Jing et al., 2020; McAllister, 1995; Rotter, 1980).

In contrast, we can examine how business and organisational psychology treat trust differently to Rotter (1980). Trust is a critical component in the Business and organisational sphere (Rousseau et al., 1998), and within these settings, there are elements of trust among colleagues, in leadership, and with the overall integrity of institutions (Mayer et al., 1995; Schoorman et al., 2007). When investigating trust in these areas, trust can be referred to as a willingness to be vulnerable and accept risk (Rousseau et al., 1998). For political science, trust pertains to citizens' confidence in government institutions, politicians, and political decision-making (Hetherington 1998; Weymouth et al. 2020).

More recently, trust in technology investigates individuals' reliance on and confidence in digital platforms, online transactions, data security, and the reliability of technological systems (Dodgson, 1993; Mcknight et al., 2011). Finally, academics can define trust in more complex systems, such as trusted computing, which involves the integration of hardware and software mechanisms to ensure that a computer behaves in expected ways, even when under attack (Gallery &

Mitchell, 2009; Shen et al., 2010). The primary goal of trusted computing is to provide a foundation for secure computing environments by protecting data integrity, confidentiality, and system integrity from various threats, which, although referred to as trust, is a very different concept than interpersonal or organisational trust.

These are only a few examples of the different areas which study trust and highlight how researchers interact with trust in different academic disciplines (Ulfert et al., 2023). One issue is that the research conducted within these disciplines is often self-contained within that academic silo, leading to issues when researching a concept which is shared across many disciplines and is actively researched in these disciplines simultaneously, such as trust. To address these issues, we will discuss the importance of moving away from unidisciplinary research when involving trust.

The extensive body of research on trust has provided valuable insights but can also present challenges. On one hand, it can significantly contribute to our understanding of trust by combining different perspectives and approaches. However, suppose there are discrepancies in trust definitions and diverse research approaches. In that case, it has the potential to result in perplexing and conflicting findings, which occurs from the jingle-jangle fallacy, which occurs when using the same term for different concepts (jingle) or when using different terms for the same concept (jangle) (Casper et al., 2017; Dang et al., 2020; Larsen & Bong, 2016; Marsh et al., 2019). These issues are complicated as trust lacks a universally accepted definition, which makes it more complex to research across disciplinary divergences (Rompf, 2015).

To reduce issues with defining and understanding trust, it becomes vital to understand the disciplinary differences in studying trust within individual academic silos. When within one academic silo, scholars become highly specialised in their fields and remain unidisciplinary (Hendriks et al., 2021; Mead et al., 2021). Unidisciplinary researchers can employ expert methodologies highly relevant to their research area. However, this can lead to challenges when employing this

data or methodology in a different discipline or combining data to create theoretical frameworks.

These silos can hinder effective communication and collaboration, highlighted by the difference between micro-trust studies, such as those in psychology that focus on trust at the individual or interpersonal level (Jing et al., 2020; Uslaner, 2008). This research contrasts meso trust investigations, which focus on trust within and between groups, organisations, and communities (Bottery, 2003; Grimmelikhuijsen & Knies, 2017) and macro trust examinations, focusing on broader, societal or institutional trust (Lu et al., 2016; Uslaner, 2008). Researchers working in these areas of trust research learn to understand these different concepts of trust. However, when other disciplines are using this work, the differences are less clear, which can lead to issues with measures of trust.

Furthermore, it is noteworthy that researchers encounter disparities in the understanding of trust not only across disciplines but also within individual fields, as seen in computer science. Within computer science, the study of trust encompasses a diverse spectrum of subjects, including trust modelling (Gulati, Sousa, and Lamas 2017; Gulati, Sousa, and Lamas 2018), trust in Human-Robot Interaction (HRI) (van Pinxteren et al., 2019), trust evaluation (Tang et al., 2012), and trust-based decision-making (Döppner et al., 2019; Ma et al., 2023; Zhou et al., 2017). As computer systems continue to increase across various parts of modern life, such as the workplace, social media, and the Internet of Things (IoT) devices (Khan et al., 2019), trust becomes increasingly important in determining the success and security of these systems. There is a call for trust researchers to be explicit when defining trust; they aim to measure and choose specific measures to align with these choices (Ulfert et al., 2023). However, there needs to be an understanding of the scientific landscape of trust research to see where research fits and which scales can be relevant to the research in question.

The examples provided above are on a case-by-case basis. Currently, no study aims to quantify the network of trust research to understand which disciplines are researching trust and what interactions exist between these disciplines. To fully grasp the intricacies of trust, it is essential to transition from using single-discipline

research methods. Embracing both multidisciplinary, where various fields contribute independently (Dalton et al., 2021), and interdisciplinarity, involving collaborative integration of insights from different disciplines (Dalton et al., 2021), is crucial. A transdisciplinary approach, which breaks down disciplinary boundaries to form a unified framework (Lang et al., 2012), becomes necessary for a comprehensive understanding of trust. To address this gap in the literature, we are conducting a bibliometric analysis to explore the network of trust research, focusing on trends, foundational papers, prevalent research themes and the organisation and interrelationships within the field.

In academic research, bibliometric analysis has emerged as an invaluable tool for quantitatively visualising trends and patterns across diverse fields. Recent years have witnessed the increasing utilisation of bibliometrics to gain deeper insights into various research domains. Notable examples include studies exploring the Ethics of Big Data (Kuc-Czarnecka & Olczyk, 2020), Cognitive Dysfunction (Chen et al., 2020), the challenges posed by COVID-19 (Hamidah et al., 2020), and Sustainable Tourism (Cavalcante et al., 2021). The complex nature of these investigations underscores the effectiveness of bibliometrics in a broad spectrum of research areas, highlighting its role as a prominent method for analysing large datasets. Within this broader context, this bibliometric analysis aims to comprehensively explain the intricate landscape of trust research within and across academic disciplines.

In this chapter we aim to provide valuable insights into trust research activities, impact, and collaboration patterns. Specifically, this analysis examines the evolution of research trends in the trust literature over time, such as the emergence of new research topics or the decline of certain research areas. In addition, we establish influential authors and publications to identify key contributors to the field and their impact on the development of the literature. Finally, we aim to identify interdisciplinary connections in the trust literature, which can help to identify the relationships between trust and other fields, such as psychology, computer science, and sociology, which can provide insights into the potential applications of trust research in other fields.

By identifying key research trends and influential papers over the past century, exploring contributions from different academic disciplines, and examining the evolution of common themes from 1922 to 2021, we aim to provide a comprehensive overview of the field. To structure the chapter, we are setting explicit research questions to guide the analysis. The literature guiding bibliometric analysis supports this approach, emphasising the importance of setting precise research questions to direct analyses, uncover significant trends, and understand the research's interdisciplinary nature and thematic development (Aria & Cuccurullo, 2017; Donthu et al., 2021; Moral-Muñoz et al., 2020; Zupic & Čater, 2015). We propose the following research questions to address:
RQ1: What are the key research trends and influential papers in trust research over the past century?

RQ2: How have different academic disciplines contributed to trust research?

RQ3: How have the most common themes in trust research evolved from 1922 to 2021?

2.2 Methods

2.2.1 Data Collection

We conducted our bibliometric analysis using the Web of Science (Classic) Core Collection due to its comprehensive document coverage and standardisation. The search included the entire collection up to December 4, 2021. The criteria for document inclusion were articles or proceedings papers written in English, with titles containing 'trust*', 'trustworthy*', or 'trustworthiness*'. Early access papers were excluded to prevent data processing errors.

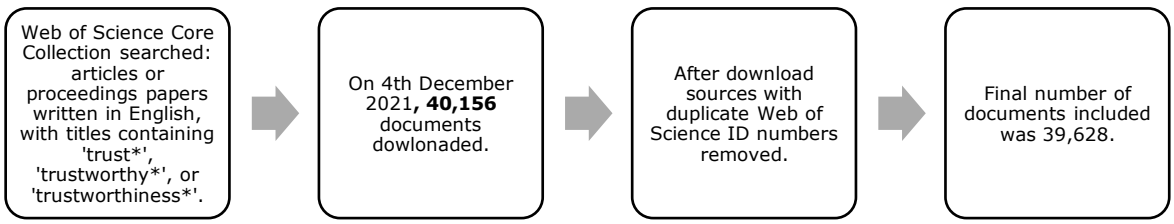
Given the objective of mapping the network of trust research, focusing on trends, foundational papers, and prevalent research themes, we deliberately decided to exclude review papers from our analysis. This decision stems from the nature of review articles, which primarily synthesise findings from a range of studies, providing valuable overviews but not contributing new empirical data. Including review papers in our dataset could potentially skew the analysis towards more

established viewpoints, thereby overshadowing emerging areas of inquiry and novel methodologies. The exclusion of review papers also addresses the methodological challenge of double-counting citations. Review articles often have extensive reference lists, which could inflate the citation counts and perceived impact of specific studies, distorting the bibliometric indicators used in our analysis. By focusing solely on empirical studies, including articles and proceedings papers, we aim to accurately represent the research dynamics within the trust literature. This approach ensures that our analysis captures the direct contributions to the field, facilitating a clearer understanding of the developmental trajectories of trust research and the interactions between different disciplines. The initial search yielded 40,156. We downloaded all results as full records, with references and abstracts in plain text format.

2.2.2 Software

We created a Python script to clean and organise the data to eliminate duplicates. We removed documents with duplicate Web of Science ID numbers and further scrutinised those with duplicate titles before removal, eliminating 528 documents. Figure 1 shows a flowchart of the overall process. For subsequent analysis of the

Figure 1. The process of collecting and cleaning data.



remaining 39,628 documents, we utilised the R Package Bibliometrix (Aria & Cuccurullo, 2017)to extract descriptive information about authors, documents, sources, and countries. We also employed VOSviewer (Van Eck & Waltman, 2007) to visualise keyword co-occurrence and bibliographic coupling of sources.

2.2.3 Data Analysis

For the analysis, we decided to perform a selection of bibliometric analysis methods: bibliographic coupling of source, keyword co-occurrence analysis, analysis of papers with the top number of citations, and the annual scientific production rate.

Bibliographic coupling of sources involves identifying pairs of documents that cite one or more common references and mapping connections and relationships between different research works (Donthu et al., 2021; Jarneving, 2007). VOSviewer 1.6.20 (Van Eck & Waltman, 2007) was used to perform bibliographic coupling analysis, focusing on papers sharing citations to provide insights into developing research themes over time. In bibliographic coupling maps, each node represents a source, with node size indicating the total link strength of the source. Links between nodes represent sources cited together regularly, and link thickness signifies the frequency of such co-citation. Larger nodes indicate greater total link strength, thicker links represent more frequent co-citations, and thinner links indicate less frequent ones.

Keyword co-occurrence analysis examines the frequency of specific keywords appearing together in documents, helping to identify prevalent research themes and the evolution of research focus over time (Sedighi, 2016). We used VOSviewer 1.6.20 (Van Eck & Waltman, 2007)for keyword co-occurrence analysis. Each node signifies a keyword in co-occurrence maps, with node size denoting the frequency of keyword occurrence. Links between nodes represent co-occurring or frequently co-occurring keywords, and link thickness indicates the frequency of co-occurrence. Larger nodes represent more frequent keyword occurrences, while thicker links signify a higher frequency of co-occurrence between keywords.

Analysing papers with the top number of citations identifies and examines the most highly cited papers, providing insights into key contributions and influential studies that have shaped the field. The annual scientific production rate tracks the number of publications related to trust research produced each year, allowing us to assess the growth and development of the field over time (Larsen & von Ins, 2010). To complete these analyses, we use the R-Studio Bibliometrix Package

(Aria and Cuccurullo 2017) to gather and graphically represent descriptive information about authors, documents, sources, and the annual production rate.

These bibliometric methods offer a comprehensive understanding of the patterns, impacts, and trends in trust research, enabling us to draw meaningful conclusions about the evolution and interdisciplinary nature of the field.

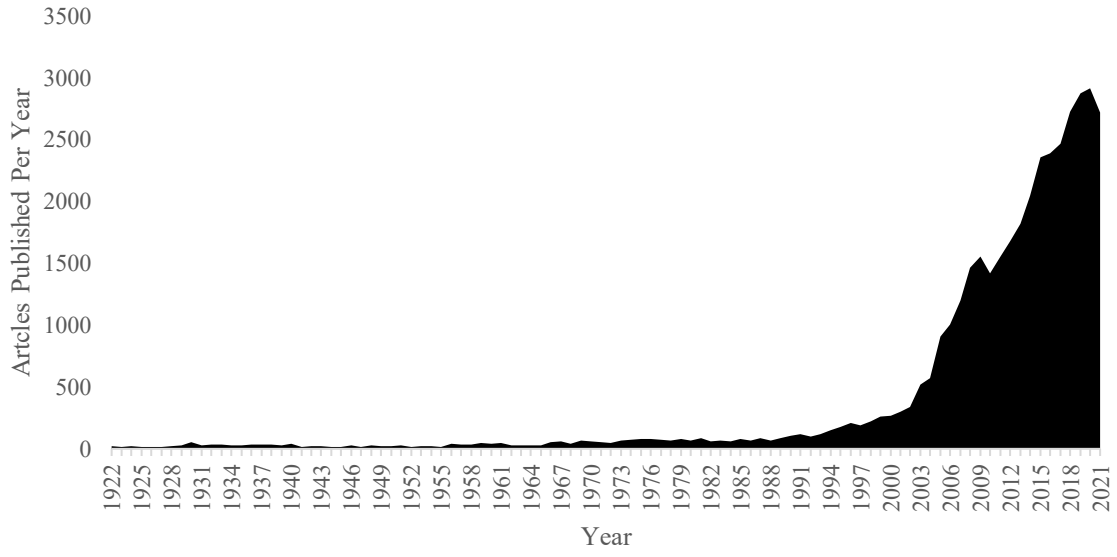
2.3 Results

We used a total of 39,628 documents spanning the years 1922 to 2021. These documents comprised 27,464 (69.4%) research articles and 12,057 (30.6%) proceedings papers. The dataset encompassed contributions from 67,358 authors, derived from 12,949 distinct publication outlets and representing 168 countries/regions.

2.3.1 Growth of Publication

Figure 2 illustrates the trajectory of annual scientific document production, which remained relatively steady until around 1991. Subsequently, a notable exponential increase in publication output commenced a pattern mirrored in various research areas as it coincides with the widespread adoption of the Internet for academic dissemination (Vakkari, 2008). It is important to note that while the overall trend reveals consistent growth, a temporary decline of 198 publications occurs in 2021.

Figure 2. The Annual Scientific Production of Documents



A delay in including articles in the core collection during our data retrieval probably causes this dip. Nevertheless, since the early 2000s, research output appears to have maintained a steady yet robust upward trajectory, underscoring the continued activity of trust-related research in academia. The exponential growth in publication output since the early 1990s highlights the increasing interest and development of fundamental research trends in trust.

2.3.2 The Top Papers Published on Trust.

From the extensive dataset of 39,628 documents, we have identified and compiled the 20 papers with the highest yearly citation rates, as presented in Table 1. We used yearly citations as a metric to mitigate the inherent bias towards older papers. Sorting by citations per year allows us to identify papers with a consistent and enduring impact, regardless of the total number of citations, highlighting papers that remain relevant and influential over time, even if they were published more recently.

The paper "Qualitative Content Analysis in Nursing Research: Concepts, Procedures and Measures to Achieve Trustworthiness" by Graneheim & Lundman (2004) tops the list with an average of 450 citations per year, mainly due to its pivotal role as a foundational work explaining the methodology for conducting trustworthy qualitative research. This paper serves as a cornerstone reference for researchers aiming to establish the trustworthiness of qualitative studies. Nowell et al's., (2017) work, focused on ensuring the trustworthiness of Thematic Analysis, also follows a similar trend with an annual citation rate of 316. While crucial for methodological robustness, these papers should be noted for their application of trust rather than direct contributions to trust research.

Morgan & Hunt's (1994) exploration of the commitment-trust theory in relationship marketing received 306 citations per year, and (Mayer et al., 1995) integrative model of organisational trust with 290 citations per year marks foundational works in organisational psychology. In business, Doney & Cannon's (1997) examination of trust and its implications for buyer-seller relationships underscores the theme's importance in commercial interactions.

The influence of technology on trust is highlighted by Ribeiro et al.'s (2016) paper on explaining predictions of machine learning models, which gathered 263 citations per year. Bertrand et al., (2004) investigated the reliability of difference-in-difference estimation in economics, which reflects trust's role in methodological rigour.

The papers span various research areas, including business, economics, psychology, computer science, and methods, illustrating trust's interdisciplinary nature. Papers like Kosfeld et al's., (2005) study on oxytocin's effect on human trust offer biological perspectives, further broadening the field's scope.

Recent papers, such as Lins et al., (2017) work on social capital, trust, and firm performance and Sicari et al., (2015) on security, privacy, and trust in Internet of Things systems, indicate growing areas of interest and the evolution of trust research to include technological and social capital dimensions.

Table 1. The 20 most cited per year articles featuring trust.

Paper Title	Citations per Year	Total Citations	Research Area
Graneheim (2004) - Qualitative Content Analysis in Nursing Research: Concepts, Procedures and Measures to Achieve Trustworthiness	450	8556	Nursing/ Methods
Nowell (2017) - Thematic Analysis: Striving to Meet the Trustworthiness Criteria	316	1893	Methods
Morgan (1994) - The Commitment-Trust Theory of Relationship Marketing	306	8861	Business
Mayer (1995) - An Integrative Model of Organizational Trust	290	8124	Organisational Psychology

Paper Title	Citations per Year	Total Citations	Research Area
Ribeiro (2016) - "Why Should I Trust You?" Explaining The Predictions of Any Classifier	263	1844	Computer Science
Bertrand (2004) - How Much Should We Trust Differences-In-Differences Estimates?	214	4061	Economics
Elo (2014) - Qualitative Content Analysis: A Focus on Trustworthiness	123	1111	Methods
Doney (1997) - An Examination of The Nature of Trust in Buyer-Seller Relationships	120	3122	Business
Kosfeld (2005) - Oxytocin Increases Trust in Humans	119	2135	Biology
Mcallister (1995) – Affect and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations	113	3174	Psychology
Josang (2007) - A Survey of Trust and Reputation Systems for Online Service Provision	106	1693	Computer Science
Chaudhuri (2001) - The Chain of Effects from Brand Trust and Brand Affect	103	2267	Business

Paper Title	Citations per Year	Total Citations	Research Area
to Brand Performance: The Role of Brand Loyalty			
Sicari (2015) - Security, Privacy and Trust in Internet of Things: The Road Ahead	101	807	Computer Science
Gosling (2004) - Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires	93	1768	Psychology/ Methods
Gulati (1995) - Does Familiarity Breed Trust - The Implications of Repeated Ties for Contractual Choice in Alliances	90	2533	Business
Lins (2017) - Social Capital, Trust, And Firm Performance: The Value of Corporate Social Responsibility During the Financial Crisis	88	529	Economics
Kim (2008) - A Trust-Based Consumer Decision-Making Model in Electronic Commerce: The Role of Trust, Perceived Risk, and Their Antecedents	87	1301	Business
Birt (2016) - Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation?	86	605	Methods

Paper Title	Citations per Year	Total Citations	Research Area
Garbarino (1999) - The Different Roles of Satisfaction, Trust, and Commitment in Customer Relationships	86	2074	Business
Pavlou (2003) - Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with The Technology Acceptance Model	86	1727	Business; e-commerce

2.3.3 Analysis of Journals and Conference Proceedings Publishing Papers on Trust

The extensive dataset of retrieved documents spans a diverse array of 12,949 journals and conference proceedings. Table 2 presents the top 20 sources that have contributed significantly to the corpus of trust-related literature. Our analysis identified the 20 most productive sources in trust research based on the number of documents published. The journal "Trust & Trustees" leads with 690 documents, followed by "IEEE Access" and "Plos One" with 239 and 196 documents, respectively. "Sustainability" and "Frontiers in Psychology" are also significant contributors, with 146 and 133 documents.

The Total Link Strength (TLS), indicative of the source's centrality in the field, the five sources with the highest TLS in trust literature are Sustainability (TLS: 5553), the Journal of Business Ethics (TLS: 4988), Plos One (TLS: 4911), Frontiers in Psychology (TLS: 4741) and Industrial Marketing Management (TLS: 4658). The average year of publication they are ranged from as early as 1939 for the "Yale Law Journal" to as recent as 2019 for "IEEE Access", "Sustainability", and "Frontiers in Psychology", reflecting both longstanding and emerging sources of scholarly output in trust research.

Table 2. The 20 most productive sources of trust literature.

Source	Number of Documents	Total Link Strength (TLS)[a]	Number of Citations	Average Year of Publications
Trusts & Trustees	690	262	464	2014
IEEE Access	239	4542	2984	2019
Plos One	196	4911	2219	2017
Sustainability	146	5553	853	2019
Frontiers In Psychology	133	4741	736	2019
Yale Law Journal	125	352	171	1939
Journal Of Business Ethics	119	4988	5306	2011
Wireless Personal Communications	109	1479	853	2017
Computers In Human Behavior	100	4393	6433	2016
Real Property Probate and Trust Journal	100	131	4	1973
Journal Of Business Research	96	4657	4800	2013
Security And Communication Networks	92	1443	687	2016
Columbia Law Review	91	260	208	1946

Source	Number of Documents	Total Link Strength (TLS)[a]	Number of Citations	Average Year of Publications
Journal Of Economic Behavior & Organization	88	2546	4221	2012
Social Science & Medicine	82	1713	3306	2012
Social Indicators Research	81	3176	1201	2016
Industrial Marketing Management	80	4658	4502	2012
Future Generation Computer Systems-The International Journal of E-science	78	1774	1840	2017
Virginia Law Review	72	153	16	1952
Psychological Reports	71	955	1478	1991

Note. [a]TLS is the sum of the strengths of all links an item has with other items in the network, indicating the item's interconnectedness and influence.

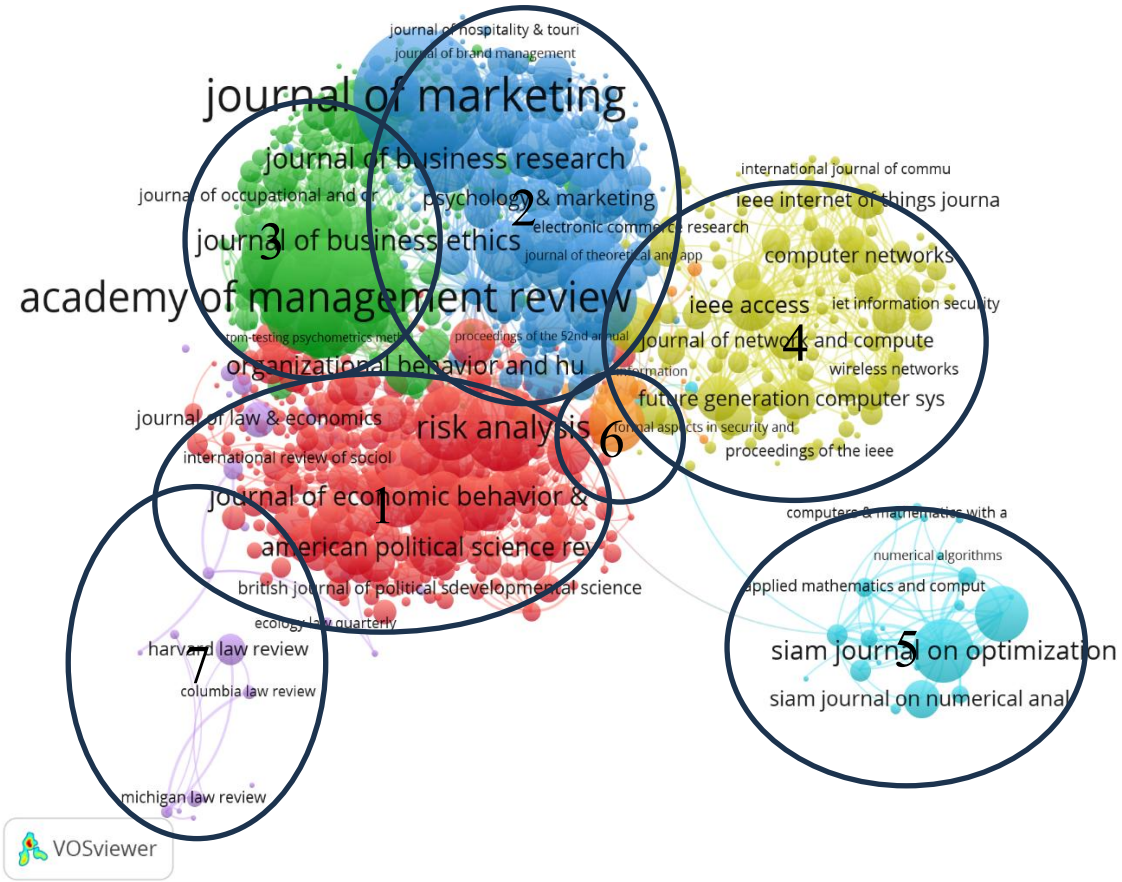
2.3.4 Bibliographic Coupling Analysis of Sources

We conducted a bibliographic coupling analysis to provide a comprehensive landscape of journals and conference proceedings publishing content on trust since 1922 and it is displayed in Figure 3.

We compiled the dataset for the analysis from an initial pool of 12,930 sources from the Web of Science. These sources included journals and conference proceedings that have published work on trust.

We employed VOSviewer to perform a bibliometric coupling analysis on the dataset. The inclusion criteria for the sources were set to a minimum of one published document and at least one citation per source, reducing the dataset to 9,308 sources. We used fractional counting to apportion citations among sources based on shared references. We selected the 1,000 sources with the highest link strength to other sources mapping. We set the network's visualisation parameters as normalisation using the Lin/Log Modularity method, layout parameters set to an attraction of 6 and a repulsion of 0, and clustering set at 1. The size of each node (circle) in the visualisation represented the number of citations for each source.

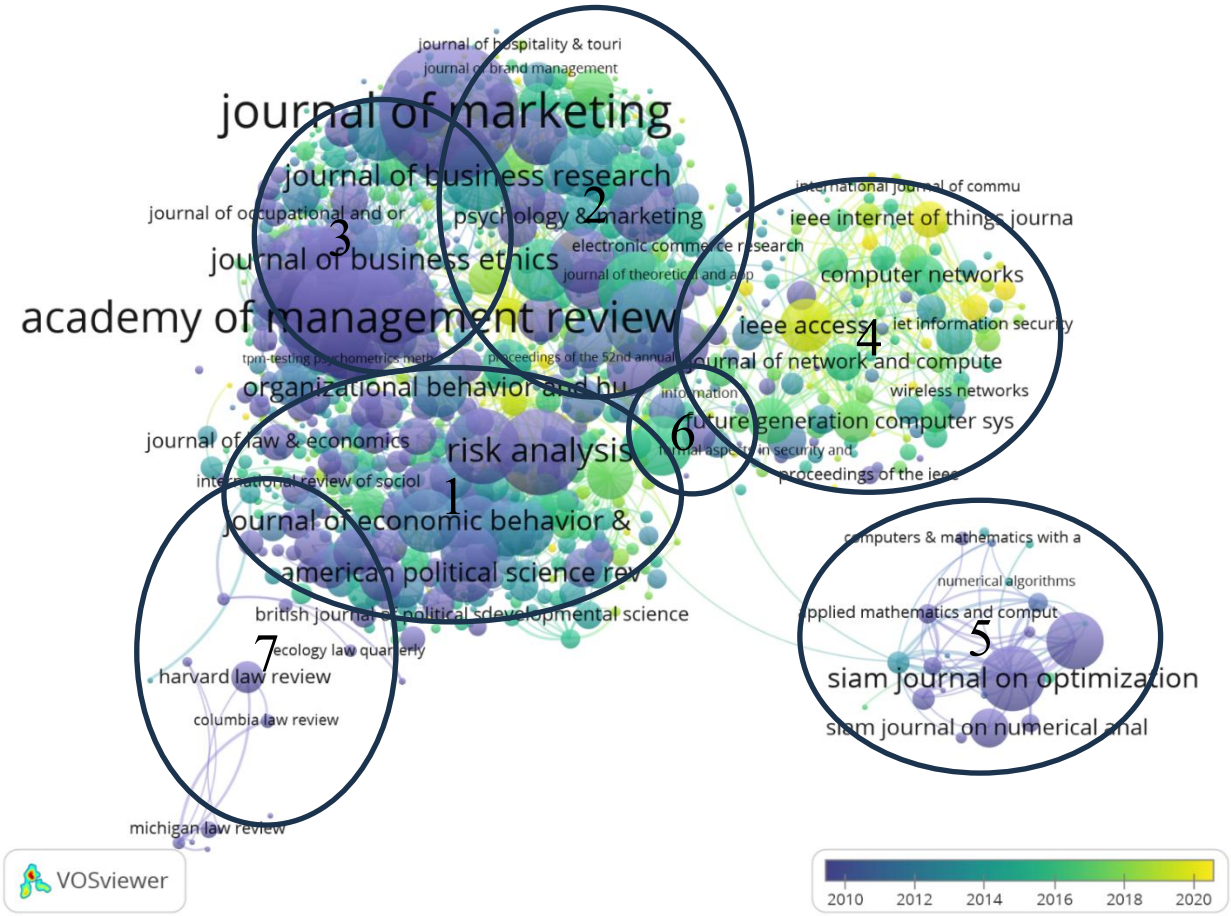
Figure 3. Bibliographic Coupling of Sources Publishing Trust Literature



To see an interactive version of this map, please visit https://tinyurl.com/2726h2tp. *This link will take you to a generic version of this map. Please enter the visualisation parameters described earlier to see the map in Figure 3.

The resulting bibliographic coupling map, presented in Figure 3, encompasses 1000 sources. This map's node size corresponds to the number of citations each source has received, while the interconnecting links depict relationships between sources. Closer nodes indicate a higher co-citation frequency, highlighting solid relationships between those sources.

Figure 4. Overlay Visualisation of the Bibliographic Coupling of Sources



To see an interactive version of this map, please visit https://tinyurl.com/2726h2tp. *This link will take you to a generic version of this map. Please enter the visualisation parameters described earlier to see the map in Figure 4.

Figure 4 is a network visualisation with an overlay that indicates the average year of publication for sources in the field of trust research. It complements this analysis with an overlay visualisation showcasing the publication periods of sources.

The colours of the nodes represent the average year of publication of the articles within each source. The colour gradient displays the timeline from 1985 to 2021, with lighter colours (like yellow) indicating more recent years and darker colours (like dark green) indicating older years. This colour coding indicates whether a source's contributions to the field are more historical or recent; a source with an average publishing year of 1955 will be dark blue, and one with an average publishing year of 2021 will be light yellow.

2.3.5 Clustering and Discipline Identification

To identify the disciplines represented within each cluster, we extracted the labels of the sources from the network visualisation. Each label typically contained the name of the journal or conference and often included terms indicative of the source's scope or focus. We conducted a manual content analysis (Hsieh & Shannon, 2005) to determine the most common words within the labels of each cluster, which served as indicators of the predominant disciplines. To focus on content-specific terminology, we excluded Common English words and specific terms related to publishing and conferences (e.g., 'journal', 'conference', 'proceedings'). Previous bibliometrics have used content analysis (Leong et al., 2021; H. Liang & Shi, 2022); we inferred the following clusters using this method.

Cluster 1 (Red) is related to social sciences, as indicated by terms such as social, public, psychology, health, economics, policy, political science, and communication. This cluster includes sources with an average publication year of 2013 and has the highest total citations. The oldest source in this cluster is "Psychological Bulletin", with an average publication year of 1991 and the newest is "Healthcare", emerging in 2021.

Cluster 2 (Green) focuses on business and management, with subjects like marketing, information systems, technology, electronic commerce, and tourism being prevalent. This cluster stands out with the recent average publication year of 2015 and a significant citation impact. The "Journal of Marketing", averaging 1997, is the oldest, while "Sustainable Production and Consumption", starting in 2021, is the newest source in this cluster.

Cluster 3 (Blue) also centres on management and business, including psychology, human resources, organisational studies, accounting, and leadership. The average publication year for this cluster is 2013. The "Journal of Psychology", dating back to 1990, is the oldest, and "International Journal of Public Leadership", which began in 2020, is the newest source.

Cluster 4 (Yellow), indicated in yellow, revolves around information systems and computing. A strong focus on security, computer science, communications, and

trust-related applications characterises it—sources in this cluster average 2015. "Trust and Deception in Virtual Societies", from 2001, is the oldest source, and the newest is "IEEE Transactions on Network Science and Engineering", which started in 2021.

Cluster 5 (Purple) encompasses mathematical and computational disciplines, including optimisation, applied mathematics, numerical analysis, computational science, and mathematical applications. The average publication year for this cluster is 2010, making it the oldest on average. "SIAM Journal on Numerical Analysis", with an average publication year of 1993, is the oldest source, and "Mathematical Problems in Engineering", from 2017, is the newest.

Cluster 6 (Light Blue) is centred on ergonomics, human factors, robotics, engineering, design, and transportation systems, focusing on applied science and engineering. The oldest source in this cluster is the "International Journal of Industrial Ergonomics", averaging the year 2009 with 182 citations. The newest is "Frontiers in Robotics and AI", which began in 2020 and has 31 citations.

Cluster 7 (Orange) has a legal orientation, with sources often related to law studies and university law reviews. The "Yale Law Journal", the oldest in the cluster, averages back to 1939 with 171 citations, demonstrating its longstanding influence. Conversely, the newest source, "Trusts & Trustees", started in 2014 and accumulated 464 citations, reflecting its growing impact in the legal field.

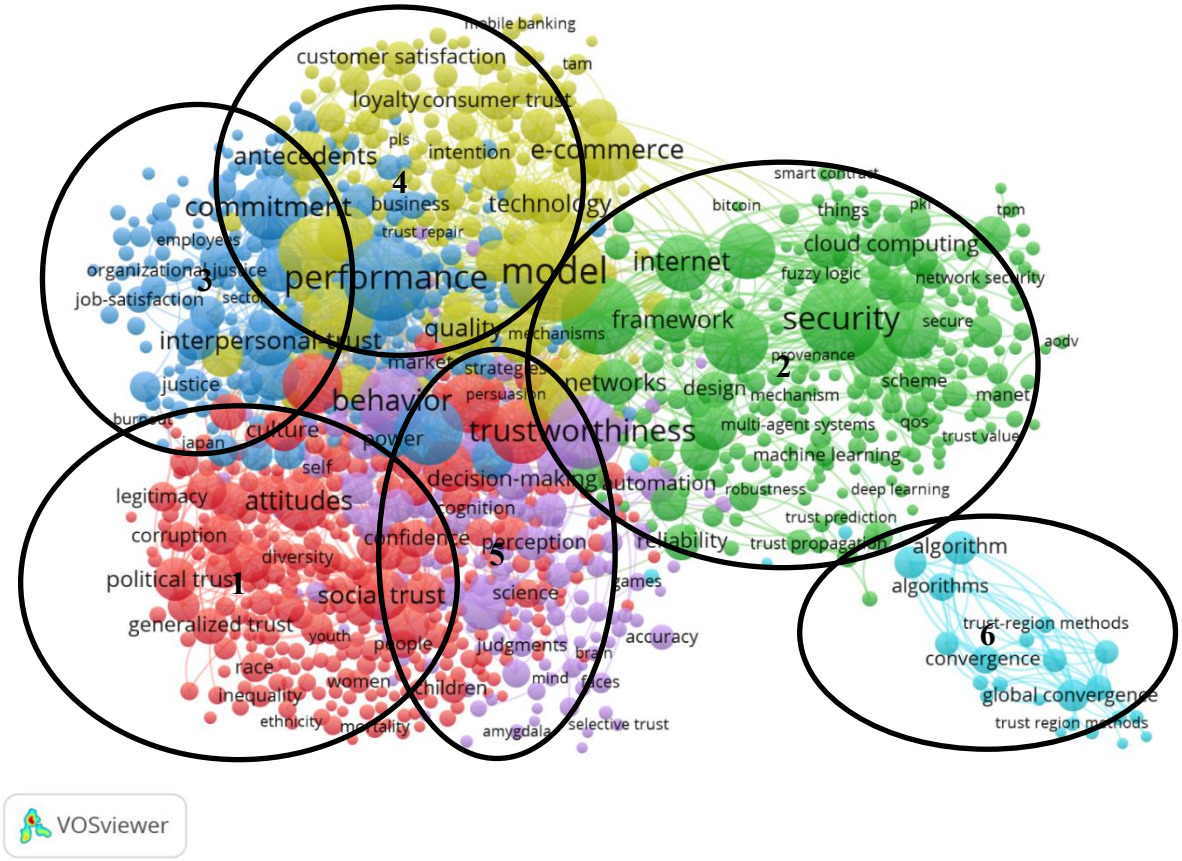
Together, these visualisations provide a comprehensive portrayal of the landscape of trust-related research sources, highlighting their interconnections and their historical and contemporary significance in the field.

2.3.6 Keyword Co-Occurrence Analysis

We conducted a keyword co-occurrence analysis to gain deeper insights into themes within trust research. We used author keywords and keywords plus. At the beginning of the analysis, there were 59,122 keywords. The word had to appear more than five times to be used in the analysis, leading to 5516 words included in the analysis. We then selected the top 1000 most linked words to keep in the

analysis. We excluded the word trust from the map as it overshadowed most of the map. We applied fractional counting using the Lin/Log Modularity method, layout parameters set to an attraction of 3 and a repulsion of 0, and clustering set at 1.0.

Figure 5. Co-occurrence of Author Keywords and Keywords Plus.



To see an interactive version of this map, please visit https://tinyurl.com/2726h2tp. This link will take you to a generic version of this map. Please enter the visualisation parameters described earlier to see the map in Figure 5.

In Figure 5, we provide a co-occurrence map of these keywords. Each node in the map represents a keyword, with node size indicating the frequency of the keyword's occurrence. Links between nodes signify co-occurrence, and the thickness of the link reflects the frequency of such co-occurrence. This analysis reveals seven distinct keyword clusters.

Cluster 1 (red) contains 294 keywords and has a total of 56,196 links, with an average link strength of 95 and an average of 99 occurrences per keyword. The average publication year for this cluster is 2014, with an average of 24.3 citations per keyword. This cluster focuses on the interplay between social and political

elements of trust, featuring keywords like "perceptions", "attitudes", "social capital", and "political trust".

Cluster 2 (green) has 211 keywords and 27,107 links, with an average link strength of 103 and an average of 114 occurrences per keyword. The average publication year for this cluster is 2015, with an average of 13.1 citations per keyword. This cluster centres on the foundational aspects of security and trust in digital environments, highlighted by terms such as "security", "trust management", and "reputation".

Cluster 3 (dark blue) contains 173 keywords with 37,636 links, with an average link strength of 114 and an average of 117 occurrences per keyword. The average publication year for this cluster is 2014, with an average of 42.4 citations per keyword. This cluster pertains to the role of trust within psychological settings, emphasising "commitment", "performance", and "interpersonal trust".

Cluster 4 (yellow) includes 155 keywords and has 36,439 links, with an average link strength of 147 and an average of 150 occurrences per keyword. The average publication year for this cluster is 2015, with an average of 39.9 citations per keyword. It relates to the dynamics of e-commerce and customer satisfaction, with key terms including "model", "satisfaction", "e-commerce", and "brand trust".

Cluster 5 (purple) contains 140 keywords with 27,213 links, with an average link strength of 98 and an average of 102 occurrences per keyword. The average publication year for this cluster is 2014, with an average of 31.1 citations per keyword. This cluster delves into the psychological and social bases of trust, featuring terms like "trustworthiness", "cooperation", and "trust game".

Cluster 6 (light blue) is the smallest, with 26 keywords and has a total of 1,723 links, with an average link strength of 93 and an average of 100 occurrences per keyword. The average publication year for this cluster is 2011, with an average of 19.4 citations per keyword. This cluster emphasises mathematical and optimisation techniques in trust modelling, focusing on "global convergence" and "trust region method".

2.4 Discussion

2.4.1 Growth of Production

The exponential growth in trust-related publications since the early 1990s, as highlighted in our results, can be attributed to the digital revolution (Vakkari, 2008) and the increasing complexity of societal interactions that necessitate a deeper understanding of trust dynamics(Khan et al., 2019; Leong et al., 2021; McKnight & Chervany, 2001; Sicari et al., 2015). This increase underscores trust's escalating academic and practical significance, especially in the context of rapid technological advancements and their implications for interpersonal and institutional trust.

2.4.2 Top Papers Published

Identifying the 20 most cited papers per year within the trust research domain offers critical insights for the network of trust research. By focusing on yearly citations, we have highlighted works that maintain relevance and impact over time, illustrating the relevance of trust research across various disciplines. This approach aligns with our objective to map out the field's current state and identify emergent trends and foundational works that have shaped the understanding of trust.

Relating to RQ1, what are the key research trends and influential papers in trust research over the past century, the analysis reveals that influential papers span various domains, reflecting the evolution of trust research over time. Foundational works such as Graneheim and Lundman's (2004) and Nowell et al. (2017) focus on methodological trust, establishing robust frameworks for qualitative research and thematic analysis. These papers highlight the importance of methodological rigour in trust research, especially in qualitative contexts. In organisational psychology, Morgan and Hunt's (1994) commitment-trust theory and Mayer et al.'s (1995) integrative model of organisational trust are pivotal, reflecting early key trends in understanding trust dynamics within business and organisational settings. More recent influential papers, such as Ribeiro et al.'s (2016) on machine learning model predictions and Sicari et al.'s (2015) on trust in the Internet of

Things, underscore technology's growing relevance as a trust research theme. This shift illustrates how themes of trust research are expanding and evolving to address new challenges in technology and social capital.

This analysis also provides information to RQ2 on how different academic disciplines have contributed to the field of trust research. The analysis shows diverse, highly cited papers from psychology, business, economics, computer science, and methodological studies. For instance, Graneheim and Lundman (2004) and Nowell et al., (2017) contribute from a methodological perspective, while Morgan & Hunt (1994) and Mayer et al. (1995) offer insights from organisational psychology. Ribeiro et al.'s (2016) work in computer science and Sicari et al.'s (2015) research in cybersecurity provide clear evidence of technology's impact on trust.

Business is the most prominent area, with six high-citation papers underscoring trust's central role in business contexts. One paper represents business and e-commerce. Computer Science also features four influential papers, indicating the growing intersection between trust and technology, which illustrates the expansion of trust research into online and electronic commerce environments. These research areas reflect the increasing importance of understanding trust in digital transactions and platforms. Psychology and economics each contribute two significant papers, one representing Psychology/Methods. Methods and nursing/methods contribute with two and one paper(s). The inclusion of nursing and methods in the dataset reflects the cross-disciplinary applications of trust research in ensuring methodological robustness.

The top papers answer RQ3, how the most common themes in trust research evolved from 1922 to 2021 when focusing on the field's evolution. In the earliest papers, there is more research in business and psychology (Doney & Cannon, 1997; Garbarino & Johnson, 1999; Gulati, 1995; Mayer et al., 1995; McAllister, 1995; Morgan & Hunt, 1994). These findings suggest that the foundational understanding of trust was based on organisational psychology and business studies. When moving through work published in the 2000s, there is a shift in the research trends seeing a few business papers (Chaudhuri & Holbrook, 2001; D. J.

Kim et al., 2008), a paper on e-commerce (Pavlou, 2003) and papers on methodologies (Gosling et al., 2004; Graneheim & Lundman, 2004), and the start of papers from computer science with (Jøsang et al., 2007). During this middle period, we see a more diverse sample of papers and the emergence of new research areas. The newest papers come mainly from methods attempting to develop trustworthy qualitative research (Birt et al., 2016; Elo et al., 2014; Nowell et al., 2017), computer science (Ribeiro et al., 2016; Sicari et al., 2015) and economics (Lins et al., 2017). These papers highlight a broadening scope in trust research, expanding from its roots in business and psychology to incorporate various disciplines. Computer science is becoming more involved, reflecting the field's evolving complexity and interdisciplinary nature.

Overall, the distribution of top papers illustrates trust research's interdisciplinary nature and how it has developed over time. The analysis highlights how different academic fields contribute unique perspectives and methodologies to studying trust. It also reflects the growing importance of trust in various contexts, from traditional business and psychological studies to emerging areas in technology and e-commerce. As the scientific landscape of trust research becomes more varied and complex, it becomes critical to ensure multidisciplinary collaboration across these fields to allow for thorough understanding and appropriate trust testing.

2.4.3 Source Publishing

The historical development of trust research reveals key trends and interdisciplinary contributions, aligning with the three research questions. Early research focused on legal and psychological aspects, with influential papers from the Yale Law Journal (1939) and Psychological Reports (1991). The field diversified in the 21st century to include business ethics, economic behaviour, and social sciences, with significant contributions from journals like the Journal of Business Ethics (2011), Journal of Economic Behaviour & Organization (2012), and Social Science & Medicine (2012). Recent trends highlight technologically driven research, with journals like Computers in Human Behaviour (2016), IEEE Access (2019), and Plos One (2017) emphasising digital trust. This evolution underscores the expanding scope of trust research, reflecting interdisciplinary collaboration and shifting themes from foundational legal and psychological concepts to complex

modern contexts, addressing themes like digital trust, cybersecurity, and environmental considerations.

The analysis reveals several key trends in trust research over the past century, addressing RQ1. Cluster 1 (Red), associated with social sciences, highlights foundational research in public health, psychology, and political science. Influential sources like the "Psychological Bulletin", which has been significant since 1991, illustrate the impact of psychological and societal perspectives on trust. Cluster 2 (Green) focuses on business and management, with influential journals like the "Journal of Marketing" and recent sources such as "Sustainable Production and Consumption" (2021), reflecting the integration of trust with commercial and technological advancements. Cluster 4 (Yellow) centres on information systems and computing, where journals like "IEEE Transactions on Network Science and Engineering" underscore modern cybersecurity and digital trust research. Additionally, Cluster 5 (Purple) encompasses mathematical and computational disciplines, with sources like the "SIAM Journal on Numerical Analysis" (1993) and "Mathematical Problems in Engineering" (2017) highlighting the application of mathematical methods to trust research. Cluster 6 (Light Blue) focuses on applied science and engineering, with journals like the "International Journal of Industrial Ergonomics" (2009) and "Frontiers in Robotics and AI" (2020), reflecting trust's role in ergonomics, robotics, and transportation systems. Cluster 7 (Orange), with its legal orientation, features historically significant sources such as the "Yale Law Journal" (1939) and newer entries like "Trusts & Trustees" (2014), showcasing the evolving role of trust within legal frameworks.

Addressing RQ2, our current study highlights the contributions of various academic disciplines to trust research. Cluster 1 (Red) shows the impact of social sciences, focusing on trust in public health, psychology, and political science. Cluster 2 (Green) illustrates the business and management field's contributions with research on marketing, information systems, and technology. Cluster 4 (Yellow) demonstrates the importance of information systems and computing, particularly in cybersecurity and digital trust. Cluster 5 (Purple) reflects the application of mathematical and computational methods to trust research, highlighting its significance in optimisation and numerical analysis. Finally, Cluster

6 (Light Blue) showcases the role of applied science and engineering, emphasising trust in ergonomics, robotics, and transportation systems. Cluster 7 (Orange) reveals how legal studies contribute to understanding trust within legal contexts.

For RQ3, the evolution of trust research themes from 1922 to 2021 is evident through sources' historical and contemporary significance across clusters. Older influential sources, such as the "Yale Law Journal" (1939) and the "Journal of Psychology" (1990), indicate the foundational research in legal and management contexts. The emergence of newer publications like "Healthcare" (2021) and "IEEE Transactions on Network Science and Engineering" (2021) reflects contemporary shifts towards new research areas, such as healthcare and digital networks. Cluster 5 (Purple) and Cluster 6 (Light Blue) highlight how themes in mathematical, computational, and applied sciences have evolved, with recent developments in these fields addressing new challenges in trust. This evolution demonstrates how trust research has expanded over time, incorporating diverse disciplinary insights and adapting to societal and technological changes.

The significant output and TLS of journals like "IEEE Access" and "Plos One" underscore the increased attention towards interdisciplinary work, particularly at the intersection of technology and trust (Cannizzaro et al., 2020; Srikanth et al., 2022; Zloteanu et al., 2018). This observation is particularly relevant given the challenges and necessity of transcending disciplinary boundaries to understand trust fully.

These findings present a rich and dynamic view of trust research, with significant contributions from various disciplines, influential historical journals, and evolving themes that reflect longstanding interests and emerging trends in understanding trust.

2.4.4 Keyword Co-occurrence in Trust Research

The keyword co-occurrence analysis, focusing on trust research from 1990 onwards, reveals trust studies' complexity and multidisciplinary nature. The prevalence of keywords such as 'model', 'performance', and 'security' underscores

the foundational role of trust in diverse academic investigations, from theoretical modelling and organisational performance to security in digital environments.

The analysis highlights several key trends in trust research addressing RQ_1, mainly the increasing focus on digital and e-commerce contexts. Cluster 2, which centres on digital trust and security with keywords like "security", "trust management", and "reputation", has seen significant growth in recent years, reflecting the rising importance of trust in digital environments. This cluster's recent average publication year (2015) and relatively high keyword occurrence underscore its contemporary relevance. The lower average citation count (13.1) suggests that while this area is expanding rapidly, it is still in the development phase, with influential papers emerging as the field matures.

Cluster 4, focusing on e-commerce and customer satisfaction, includes terms such as "model", "satisfaction", "e-commerce", and "brand trust". This cluster, with its average publication year of 2015 and a high average citation count of 39.9, indicates a well-established and influential body of research. The strong emphasis on consumer trust in digital transactions reflects the importance of trust in the burgeoning e-commerce sector. The significant citation count suggests that foundational papers in this area have substantially shaped the understanding of trust in online environments.

In contrast, Clusters 3 and 5, which deal with the psychological aspects of trust and its social and psychological foundations, show a more mature body of research. Cluster 3, with keywords like "commitment", "performance", and "interpersonal trust", has a high average citation count (42.4), pointing to its influential contributions to understanding the psychological dimensions of trust. Cluster 5, focusing on terms such as "trustworthiness", "cooperation", and "trust game", also demonstrates a robust body of work with a substantial average citation count (31.1). These clusters reveal that foundational psychological and social research continues to be highly relevant and impactful.

The diverse clusters reflect the broad interdisciplinary nature of trust research and address RQ2. Cluster 1, which addresses social and political dimensions with

keywords such as “perceptions”, “attitudes”, and “political trust”, highlights contributions from social and political sciences. This cluster's average publication year of 2014 and significant citation count (24.3) indicate a well-developed area of research that integrates insights from various social science perspectives.

Cluster 6, with its focus on mathematical and optimisation techniques (e.g., “global convergence” and “trust region method”), showcases the contribution of quantitative and computational disciplines to trust research. Despite being the smallest cluster with only 26 keywords, its specialised focus on mathematical models represents a critical area of research for developing and refining trust algorithms. The average citation count of 19.4 reflects its niche influence in the field.

Finally, we can address RQ3 through the thematic evolution in trust research, which reveals a shift from foundational psychological and social studies to a focus on digital and e-commerce contexts. Earlier research, as represented by Clusters 1 and 3, laid the groundwork for understanding trust in social and psychological settings. Over time, the field has expanded to address emerging challenges related to digital environments and online transactions, as seen in Clusters 2 and 4.

This progression highlights how trust research adapts to technological and societal changes. The increasing importance of digital trust mechanisms and consumer satisfaction in e-commerce reflects broader trends in technology and market behaviour. The continued relevance of psychological and social studies underscores the enduring significance of understanding trust's fundamental aspects, even as new contexts emerge.

The keyword co-occurrence analysis reveals a dynamic and evolving field of trust research, with significant contributions from various disciplines and a clear shift towards addressing contemporary challenges in digital and e-commerce contexts. Future research should continue to explore these emerging themes and refine our understanding of trust in an increasingly complex world.

In summary, we can determine that many research areas use the term trust, and the presence of more distinct clusters suggests that some research areas may have limited interaction, leading to the misuse of the term trust or the misapplication of other methods. By mapping the thematic clusters and their interconnections, this analysis provides insights into the current landscape of trust studies and underscores the necessity of integrating perspectives from various disciplines to tackle the complex phenomena of trust.

2.4.5 Limitations and Future Work

Despite its comprehensive scope, this bibliometric analysis is subject to several limitations that suggest directions for future research. First, the focus on documents indexed in the Web of Science may have excluded relevant literature from other databases or publications unavailable in English, potentially limiting the complete representation of trust research. Future studies could expand the analysis to include multiple databases and languages to capture a more global perspective on trust research. Secondly, the reliance on keyword co-occurrence and citation metrics may not fully capture the nuances of interdisciplinary connections and the depth of trust research themes. Qualitative analyses, such as systematic literature reviews or expert interviews, could complement bibliometric methods to provide deeper insights into the conceptual and theoretical developments within the field. Thirdly, the analysis was limited to documents up to December 4, 2021, so recent developments and emerging trends in trust research may not be fully captured. Ongoing bibliometric analyses are necessary to keep pace with the evolving landscape of trust research.

Future work should also explore the practical applications of trust research in policymaking, organisational practices, and technology development. Understanding how trust research translates into practical strategies and interventions could significantly benefit practitioners, policymakers, and researchers alike.

2.4.6 Conclusions

The comprehensive bibliometric analysis of trust research spanning various academic disciplines and timeframes offers several critical insights. Firstly, the exponential growth in trust-related publications since the early 1990s is a testament to the increasing complexity of societal interactions and the digital revolution. This growth highlights trust's escalating academic and practical significance, particularly in the context of rapid technological advancements and their implications for interpersonal and institutional trust.

Our examination of the most cited papers provides a clear picture of trust research's foundational and emerging themes. Foundational works in organisational psychology and methodological studies have laid the groundwork for understanding trust dynamics in business and qualitative research settings. Meanwhile, recent influential papers focusing on technology, such as those addressing machine learning model predictions and trust in the Internet of Things, underscore the growing relevance of trust in digital and technological contexts.

The diverse contributions from various academic disciplines, as evidenced by the citation analysis, demonstrate the interdisciplinary nature of trust research. Business and management studies emerge as prominent areas, reflecting the critical role of trust in commercial interactions and organisational behaviour. The significant presence of computer science publications indicates a robust intersection between trust and technology, particularly in digital security and trust management systems.

The evolution of trust research themes from 1922 to 2021 shows a clear progression from foundational studies in business and psychology to more contemporary investigations into digital trust and e-commerce. This thematic shift underscores how trust research adapts to technological and societal changes, reflecting broader trends in market behaviour and the increasing importance of digital transactions.

The keyword co-occurrence analysis further emphasises trust research's dynamic and evolving nature. Clusters focusing on digital trust, e-commerce, psychological

aspects, and mathematical modelling of trust reveal the diverse and interdisciplinary approaches to understanding trust. The apparent shift towards addressing contemporary challenges in digital environments and online transactions indicates the field's responsiveness to emerging issues.

Overall, the term "trust" is utilised across a myriad of research areas, and the presence of distinct clusters suggests that some research domains may have limited interaction. This segmentation can lead to the misuse of the term or misapplication of methods. Therefore, it is crucial to integrate perspectives from various disciplines to tackle the complex phenomena of trust comprehensively. This integration will ensure a complete understanding and appropriate application of trust-related concepts and methods across different fields. As trust research continues to evolve, fostering multidisciplinary collaboration to address the increasing complexity of trust in both traditional and digital contexts will be essential.

In summary, this bibliometric analysis of trust research highlights the field's evolution from foundational theories in psychology and business to modern applications in technology and security. These insights underscore the value of a multidisciplinary approach, setting the stage for a more comprehensive investigation of trust in the experimental chapters. Moving forward, the thesis will transition to examining trust empirically, exploring its application and measurement across HATs. Chapter 3 will leverage the theoretical foundation and interdisciplinary insights presented here, focusing on how trust is operationalised and measured within the unique landscape of human and AI teammate interactions. By examining these factors experimentally, this thesis aims to contribute to the theoretical discourse on trust and its practical applications in emerging technological environments.

Chapter 3 Understanding the Impact of Anthropomorphism and System Reliability on Trust and Performance in Human-Artificial Intelligence Teams

With the rapid advancement of AItechnologies, AI systems are increasingly integrated into collaborative environments across various fields, from healthcare to digital forensics. This integration is about enhancing technical capabilities and creating effective HATs, where AI systems support human decision-making while maintaining essential human agency. In HATs, trust is pivotal in determining whether humans will accept or override AI suggestions. However, trust is shaped by complex factors, including AI system reliability and the extent to which AI exhibits anthropomorphic qualities, human-like characteristics that can foster intuitive interactions and relational warmth.

In Chapter 3, we begin the first of three experimental chapters. Building on the interdisciplinary foundation established in Chapter 2, this chapter delves into the experimental investigation of trust dynamics in HATs by examining two critical design factors: anthropomorphism and system reliability. Specifically, this study explores how these elements influence participants' trust, performance evaluations, and confidence in decision-making tasks. Anthropomorphism is posited to facilitate smoother human-AI interaction by bridging cognitive and emotional gaps, potentially increasing trust even when AI reliability is low. Conversely, system reliability, which affects how consistently and accurately AI can support human tasks, remains essential for establishing a dependable collaboration baseline.

Through an experimental design, this chapter aims to explain how anthropomorphic design in written text, such as warmth in responses and varying AI reliability levels, influences team dynamics in HATs. The study's findings address gaps in the existing literature by providing nuanced insights into how anthropomorphism and reliability interact to shape trust calibration, AI performance perceptions, and human teammates' confidence. The results contribute to theoretical models of trust and offer practical guidance for designing

AI systems that are both technically effective and attuned to human social expectations. This work was published at AAAI23 as an Extended Abstract (Bailey & Pollick, 2023) and presented in full at Multidisciplinary Perspectives on Human-AI Team Trust at HHAI23.

3.1 Introduction

As AI technology rapidly advances, AI systems are becoming vital across various sectors, enhancing decision-making in fields such as digital forensics, healthcare, and finance. With increasing complexity and high stakes in these areas, HATs have emerged to harness the collective intelligence of humans and AI. AI systems offer data-driven recommendations in these teams, while humans retain final decision-making authority, ensuring that essential human agency is preserved.

Research has explored how AI systems can enhance human decision-making and team performance. For instance, recent algorithms have optimised task assignments based on the complementary strengths of humans and AI, improving task allocation across team members (Kerrigan et al., 2021; Rodgers et al., 2023; Steyvers et al., 2022; Wilder et al., 2020). However, while these advancements are promising, most design guidelines currently emphasise single-user interactions, often overlooking the nuanced dynamics present in collaborative HATs where trust and relational dynamics play crucial roles (Rix, 2022). Consequently, how AI and human collaboration influence team dynamics and affect trust towards AI within these unique roles remains underexplored. This chapter introduces key literature in AI and psychology, establishing a foundation for the research questions driving this study.

3.1.1 Team Trust in Organisational Psychology and HATs

In organisational psychology, trust is considered fundamental to effective collaboration and conflict resolution (Costa et al., 2018). Similarly, trust in HATs is foundational yet involves additional complexities due to the integration of AI as a team member. Trust within a team facilitates open communication and mutual respect, fostering an environment where members feel secure enough to share ideas and resolve disagreements collectively (Jehn, 1995; Mayer et al., 1995).

Trust, in turn, leads to stronger team cohesion, enhanced creativity, and greater productivity, as team members are more likely to support each other's efforts and work toward shared goals (Barczak et al., 2010).

3.1.2 Importance of Trust in HATs

A critical factor for the success of HATs is trust, specifically, whether humans will rely on or choose to override AI recommendations. As noted in Chapter Two, there are cross-disciplinary differences when discussing and measuring trust in HATs, so we must define trust (Ulfert et al., 2023). Trust in this context can be defined as a user's willingness to be vulnerable by accepting AI's suggestions in the presence of some level of uncertainty or risk (Glikson & Woolley, 2020; Hoff & Bashir, 2015; Rousseau et al., 1998). However, establishing trust in AI is a complex challenge, distinct from trust among human teammates. In HATs, trust must be calibrated carefully to balance reliance on AI with the appropriate level of scepticism to ensure accuracy in decision-making.

The concept of calibrated trust, the ability to gauge when to trust AI recommendations, requires that users understand the limitations of AI and the likelihood of errors (de Visser et al., 2016). In HATs, trust operates on two levels: interpersonal trust and system trust (Lewicki & Bunker, 1996; Rotter, 1980). Interpersonal trust involves human acceptance of AI as a teammate, which requires transparency and responsiveness from AI to foster dependability (Jacovi et al., 2021; Schmidt et al., 2020). Conversely, system trust centres on confidence in the AI's technical reliability, encompassing its algorithms, data integrity, and ethical standards (Cabiddu et al., 2022; Shin & Park, 2019). When team members are assured that the AI is fair and reliable, they are more likely to incorporate its recommendations.

A robust level of system trust requires clear communication about the AI's capabilities and limitations and transparency about how the system reaches its conclusions (Felzmann et al., 2019; Schmidt et al., 2020). When team members are assured that the AI is accurate but also fair and reliable, they may be more likely to incorporate its recommendations into their decision-making.

Several key technical factors, including system reliability, transparency, and predictability influence trust in AI. For AI systems to gain consistent trust, they must exhibit clear boundaries for error, transparent explanations of decisions, and adaptive behaviour that aligns with human cognitive expectations (Bansal et al., 2019). Studies show that incorporating transparency and reliability into AI design enhances user confidence and supports effective decision-making (Felzmann et al., 2019; Schmidt et al., 2020; Shin & Park, 2019; von Eschenbach, 2021).

Building on this, maintaining compatibility between AI updates and user experiences is crucial for preserving trust and optimising team performance. While updates can improve AI accuracy, they may also disrupt user trust if they lead to unexpected changes in the AI's behaviour (Bansal et al., 2019). Several studies have explored how trust in AI develops and affects users' reliance on AI recommendations. For example, discrepancies between stated and observed AI accuracy influence trust, and high accuracy alone does not necessarily make an AI system the best teammate (Bansal et al., 2021; Yin et al., 2019). Additional factors, such as user confidence in the AI and the transparency of its explanations, play a significant role in shaping trust and overall performance in HATs (Bansal et al., 2021; Yang et al., 2020; Zhang et al., 2020). These insights underscoring the importance of designing AI systems that perform well, foster and sustain user trust over time, and show that these design choices must be actively built into AI systems.

3.1.3 AI Teammates

AI systems in HATs complement human decision-making and create distinct team dynamics due to their perceived role and limitations. Studies indicate that humans interact differently with AI teammates compared to human counterparts, often ascribing unique motives and value systems to AI. For example, in defensive team games, participants were more likely to sacrifice AI teammates over human teammates, citing a "best outcome" for saving AI but "protecting the teammate" for saving humans (Ong et al., 2012). This reveals underlying human biases that may affect team cohesion and decision-making in HATs. Furthermore, research indicates that AI teammates frequently receive undue blame for team failures, a trend reflecting existing hesitancy to trust or hold AI to human responsibility

standards fully (Merritt et al., 2011). These findings suggest that biases may affect team cohesion, underscoring the need for a socio-technical approach that considers social factors like trust and communication and technical factors like reliability and transparency.

Finally, Berretta et al. (2023) conducted a scoping review highlighting the necessity for a human-centric approach to HATs. The review emphasises the importance of a socio-technical approach to facilitate the development of AI from a mere tool to a true teammate. A socio-technical approach to HATs considers social and technical aspects to optimise collaboration between humans and AI systems. This approach involves designing AI as a tool and a responsive, adaptable teammate that aligns with human relational and cognitive needs. By addressing social factors like trust, communication, and team roles alongside technical factors such as reliability, transparency, and system integrity, a socio-technical approach aims to create HATs where AI and humans interact cohesively.

In this context, anthropomorphism is critical as it makes AI more relatable to human teammates, enhancing trust by aligning AI's behaviour with human cognitive and social frameworks. Similarly, SI in AI could further strengthen AI's integration by enabling the system to exhibit socially aware behaviours and align more closely with human expectations.

3.1.4 Role of Anthropomorphism in Humanising AI

Anthropomorphism, the attribution of human-like characteristics to AI, can be a design strategy used to bridge the cognitive and emotional gap between humans and machines, making AI appear as a more intuitive and relatable team member (Duffy, 2003). Researchers suggest that anthropomorphic design, such as giving AI a familiar appearance or enabling it to communicate empathetically, can enhance interpersonal trust in HATs (Gambino et al., 2020; Chen & Park, 2021; de Visser et al., 2016; Nass et al., 1996).

Anthropomorphic elements in AI may include conversational styles, empathy, or visual characteristics, creating an experience closer to working with a human teammate (Glikson & Woolley, 2020; Steyvers et al., 2022; Westby & Riedl, 2023).

By simulating aspects of human communication, anthropomorphised AI can create a sense of familiarity and emotional connection, making it easier for human teammates to engage with and rely on the AI (Li & Sung, 2021; Song & Shin, 2024; Zhang & Rau, 2022).

While Anthropomorphism has demonstrated success in fostering the perception of AI as human-like, it is not without its constraints, notably the Uncanny Valley phenomenon, which can induce discomfort and mistrust when non-human entities exhibit overly human-like traits (Mori et al., 2012; Weisman & Peña, 2021). Challenges also arise from anthropomorphic priming and generalisation, where users may extend anthropomorphic traits to non-anthropomorphised AI based on previous interactions (Dacey & Coane, 2023; Zanatto et al., 2016). Consequently, comprehensive investigations into how Anthropomorphism impacts dynamics during the formation of HATs assume critical importance.

3.1.5 Social Intelligence in AI

SI, the capacity to exhibit socially aware and contextually appropriate behaviour, could be a key area of anthropomorphic AI design. Effective SI enables AI to respond to social cues, enhancing its integration within HATs and facilitating smoother interactions. SI traces back to 1920, originating with Thorndike's SI classification, positing three types: abstract, mechanical, and social (Thorndike, 1920). The most widely recognised definition is from Vernon (1933), encapsulating it as the "ability to get along with people in general, social technique or ease in society, knowledge of social matters, susceptibility to stimuli from other members of a group, as well as insight into the temporary moods or underlying personality traits of strangers" (p. 44). Presently, differing theories persist regarding the accurate definition and measurement of SI (Weis & Süß, 2005), but a consensus generally divides SI into five core categories: social understanding, social memory, social perception, social creativity, and social knowledge (Kihlstrom & Cantor, 2000).

The development of AI that demonstrates SI presents challenges due to its inherently human-centric nature, emphasising the capacity to engage with humans in a manner that mirrors human-human interactions. Another intricate

facet of SI is its context-sensitive character. Humans typically possess a well-developed grasp of appropriateness, exemplified by behaviours like refraining from laughter during serious occasions. They may not react favourably to a humorous AI in a serious context (Syrdal et al., 2006). AI's deficiency in social judgment could lead to heightened scrutiny of its performance by human agents, especially compared to their human counterparts' performance. Previous research suggests that SI in AI can influence trust calibration (Williams et al., 2022).

3.1.6 Aims of Study

While extensive research has explored anthropomorphism and trust in HATs, there are significant gaps, particularly in understanding how anthropomorphism interacts with reliability to influence trust, performance, and team dynamics. Existing studies often emphasise isolated factors like system reliability or single-user trust. However, the interplay between anthropomorphic elements and system performance within collaborative HATs remains underexplored.

This study addresses these gaps by investigating the combined effects of AI anthropomorphism and system reliability on trust and performance within HATs. This study manipulates anthropomorphism and reliability through an experimental design to assess their impact on trust calibration, performance perceptions, and team confidence. The following research questions guide this investigation:

H1: Higher levels of anthropomorphism and reliability will jointly predict higher task performance.

H2: There will be a significant interaction between anthropomorphism and system reliability, such that anthropomorphic AI will elicit higher trust ratings than non-anthropomorphic AI under low-reliability conditions; however, this effect will be attenuated or absent under high-reliability conditions.

H3: Participants will perceive anthropomorphic AI to exhibit higher performance than non-anthropomorphic AI, despite reliability level.

H4: Participants will perceive human teammate performance to increase alongside AI performance.

H5: Participants will report higher confidence in team decision-making when collaborating with anthropomorphic AI compared to non-anthropomorphic AI, particularly under low-reliability conditions.

3.2 Methodology

3.2.1 Participants

There were 44 participants in the experiment, 11 in each condition. Participants were all recruited through the University participant pool and received £6 per hour for participation. Of the group, 16 were male, 25 were female, 2 were non-binary, and one participant declined to answer. There were 39 full-time students (masters level or lower), 3 PhD students, one library assistant and one participant who declined to answer. Participants were all 18-37 (M = 23.81, SD= 3.95).

3.2.2 Study Design

To investigate the impact of AI humanness and reliability on human trust and decision-making, we implemented a 2x2 between-subjects design to explore the effects of AI humanness and reliability on task performance and trust in AI teammates. The response stimuli were predesigned for human and AI teammates using a Wizard of Oz design to allow complete control over the variables. The key variables were 'Humanness' and 'Reliability,' each with two levels: High (60%) and Low (30%).

The experimental setup involved pre-written responses displayed via PsychoPy (Peirce et al., 2019) to simulate AI interactions. In the high humanness condition, an AI named "Pixie" provided warm and engaging responses, whereas the low humanness AI provided technical, straightforward answers (Wiethof et al., 2021). For high reliability, the AI accuracy was set at 90%. For low reliability, AI accuracy was set at 60%. The human teammate consistently provided correct answers for 30% of the trials. The human teammate consistently provided correct answers for 30% of the queries. There were four distinct conditions, and participants were

assigned to only one condition: High Humanness with High Reliability (HH: HR), High Humanness with Low Reliability (HH: LR), Low Humanness with High Reliability (LH: HR) or Low Humanness with Low Reliability (LH: LR).

The task spanned three blocks, with each block comprising ten trials, resulting in a total of 30 different location identifications. After each trial (3x10, n=30), the correct answers were provided so participants could track the performance of the AI and human teammate. This feature was essential, as most participants would not possess knowledge of the correct answer. Without this feature, they would have been unable to track the performance of the AI and human teammates throughout the experiment. The order of the trials was randomised throughout. Participants' trust and reliance on their teammates were collected on every trial, providing comprehensive data on human-AI collaboration dynamics under varying humanness and reliability conditions. We collected 1144 confidence ratings, 1135 influence ratings, 1143 AI performance ratings and 1145 human performance ratings. The study received full ethical clearance from the MVLS Ethics Committee (application: 200210219) at the University of Glasgow.

3.2.3 Materials

3.2.3.1 Developing Response Stimuli

For crafting responses for the high humanness AI and the human teammate, we drew inspiration from the work of Mehta et al., (2016), who examined how experts and non-experts identified locations in Geoguessr. Mehta et al., (2016)identified nine knowledge sources employed in this process, including architecture, languages, driving rules, sun positioning, animals, building signs, road signs, telecommunications, signage, and landmarks. Additionally, the process typically followed a hierarchical approach, commencing with identifying the continent before progressing to the country and narrowing down to a more specific location; we adhered to this methodology in the written responses.

In the low humanness group, the AI was introduced as a technically focused AI that directly stated the answer. Conversely, in the high humanness conditions, the AI introduced itself, adopting the name 'Pixie,' and expressed enthusiasm about

being a teammate, a technique shown to enhance perceptions of humanness (Wiethof et al., 2021). Moreover, in the high humanness conditions, the AI's responses closely mirrored those of the human teammate, adopting an anthropomorphic writing style. The AI needed to convey a sense of 'warmth' in its responses.

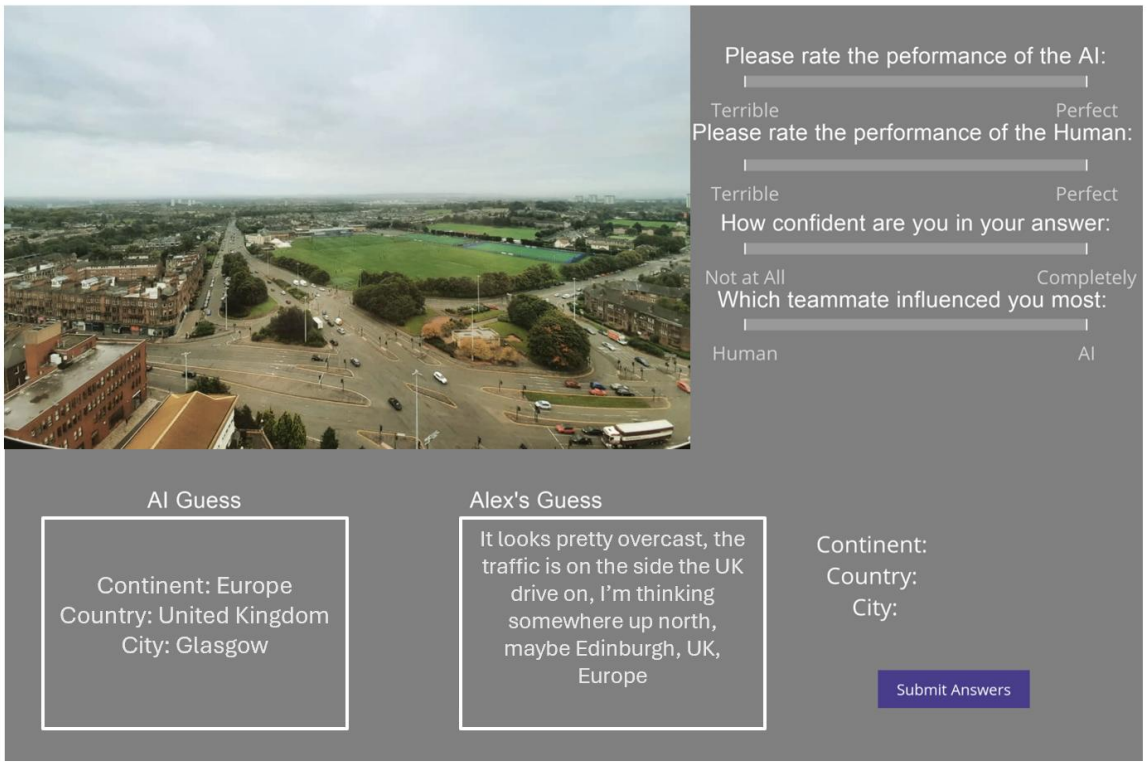
To verify that this was successful, we asked participants if they believed the AI responses to be AI-generated and human responses to be human. Overall, 93% believed the AI to be AI, and 81% believed the human to be human. These results demonstrate that our Wizard of Oz methodology was implemented successfully, and participants believed they were working with an AI and a human teammate.

3.2.3.2 Decision-Making Task

Participants engaged in a location identification task, which involved determining the geographical location of a screenshot from Google Earth. This task required specifying the screenshot's continent, country, and city/state, a scenario designed to simulate complex decision-making aided by AI. Participants acted as team leaders, making the final decision with assistance from both a human and an AI teammate.

Figure 6 is an example of the experimental setup. Lastly, we imposed a time constraint of 90 seconds per trial. The introduction of time scarcity as an environmental factor can significantly impact the outcomes of team tasks, often necessitating rapid decision-making (Hu et al., 2015; Kelly & Karau, 1999). This constraint could increase reliance on AI, as it compelled human teammates to make choices based on their implicit attitudes rather than thoroughly deliberating on the task. To accentuate this factor, we ensured that the human and AI teammates primarily provided different answers, requiring participants to choose which teammate they trusted the most. This setup aimed to mimic real-world scenarios where rapid decision-making is often necessary, potentially increasing reliance on AI. The task was designed to be difficult for the participants so we could assess the impacts of reliability and humanness under a high cognitive load.

Figure 6. The interface presents the responses from both the AI and Human teammates. In this example, AI operates within a low humanness condition. In the experiment, the pictures were taken from Google Maps, but we used a personal photo to avoid copyright issues in this example.



3.2.3.3 Questionnaires

The Propensity to Trust Machines (PtTM) (Merritt et al., 2013): A series of 6 questions where participants rated on a 7-point Likert scale how likely they are to trust machines.

The Godspeed Questionnaire (Bartneck et al., 2009)**:** Assess human perceptions of AI across five dimensions: anthropomorphism, animacy, likeability,perceived intelligence, and perceived safety. Each dimension is rated using a set of bipolar scales (e.g., from "very human-like" to "not human-like at all") on a 5- or 7-point Likert scale. Using disembodied AI, we removed the animacy and perceived safety sub-section and replaced the term 'robot' with 'AI' (Supplementary Material 1).

Questions During Each Trial: During each task trial, participants rated which teammate had influenced their decision-making on a visual analogue scale (Sung & Wu, 2018) with two endpoints, 'Human' and 'AI'. When participants selected 'Human,' it was assigned a value of 0; if they chose 'AI,' the value was 100. Participants had the freedom to click anywhere along the scale. For instance, if

their influence leaned slightly more towards AI than human teammates, they might press the scale at around 60. This influence rating served as an implicit measure of trust (Duffy, 2015; McAllister et al., 2006), with more significant influence indicating higher levels of trust. This implementation is applied to all sliders on the experimental interface. Participants also provided performance ratings for the AI and human teammates with two anchoring points of 'Terrible' and 'Perfect' after each trial. Finally, participants also rated their confidence in their answer, with the anchoring points of 'Not At All' and 'Completely'.

3.2.4 Procedure

Participants were instructed to sit at a computer-equipped table. They were provided with an information sheet explaining the experiment's premise, alongside a consent form to sign if they found the provided information acceptable. Once the consent form was signed, participants completed the PtTM Questionnaire (Merritt et al., 2013).

Following this, participants familiarised themselves with the experiment's instructions, which were all displayed throughout the experiment setup to ensure consistency across all participants. They then engaged in a sample trial. The task entailed participants identifying the location of a screenshot from Google Earth by specifying the Continent, Country, and City/State of the screenshot. Participants were designated as team leaders and were tasked with providing the final decision regarding the location. To assist them in this task, they collaborated with a human teammate and an AI teammate, both of whom offered answers to aid the participant in pinpointing the location. At the end of each trial, participants filled in the four sliders and were then shown the correct answer.

The task spanned three blocks, with each block comprising ten trials, resulting in a total of 30 different location identifications made throughout the study. Between each block, there was a 60-second break. Once the experiment was finished, participants completed the Godspeed Questionnaire (Bartneck et al., 2009), and we removed the Animacy section as our AI had no embodiment. Participants were also asked how much they trust the AI. Finally, participants were asked to determine whether they believed their AI teammate was genuinely an AI and

whether their human teammate was indeed a human. After this, participants were provided with a physical debrief explaining the experiment, including using a Wizard of Oz design, which had contact information for the researcher if participants decided to withdraw after the experiment.

3.2.5 Developing Linear Mixed Model for Analysis

To perform this analysis on the sliders taken on every trial, we utilised linear mixed models (LMMs) using the lme4 in R-Studio (Bates et al., 2015). The model incorporated the following measures: AI performance, human performance, confidence ratings, and influence ratings. To extract p-values, we used t-tests and Satterthwaite's method, which was suitable for the 2x2 design. When developing the model we did implement trial as a random effect but found it had little variance and reduced the fit of the model.

A linear mixed model extends the traditional linear regression model that accounts for fixed and random effects (Barr, 2013)**.** This modelling approach is used as there is variability at different levels of analysis. In the model, we need to identify two distinct effects, these are:

Fixed Effects: These represent the systematic, predictable relationships between predictors (reliability and humanness) and the response variable (AI performance, human performance, trust and confidence ratings).

Random Effects: These capture the variability that arises from different levels of grouping or clustering in the data.

3.2.5.1 Model Specification

To analyse the impact of reliability and humanness on performance across different measures, we utilised the following linear mixed model:

$$y_{ij} = \beta_0 + \beta_1 rel_i + \beta_2 hum_i + \beta_3 (rel_i \times hum_i) + u_{0j} + e_{ij}$$

In this model, this is the breakdown of each component:

- y_{ij} Is the response variable for the ith observation of the jth participant.
- β_0 is the intercept.
- $\beta_1 rel_i$ is the coefficient for the fixed effect of reliability.
- $\beta_2 hum_i$ is the coefficient for the fixed effect of humanness.
- $\beta_3 (rel_i \times hum_i)$ is the coefficient for the interaction between Reliability and Humanness.
- u_{0j} represents the random effect for participant j, which accounts for the variation in the intercept across participants.
- e_{ij} is the residual error term for the ith observation of the jth participant.

3.2.5.2 Post-Hoc Analysis

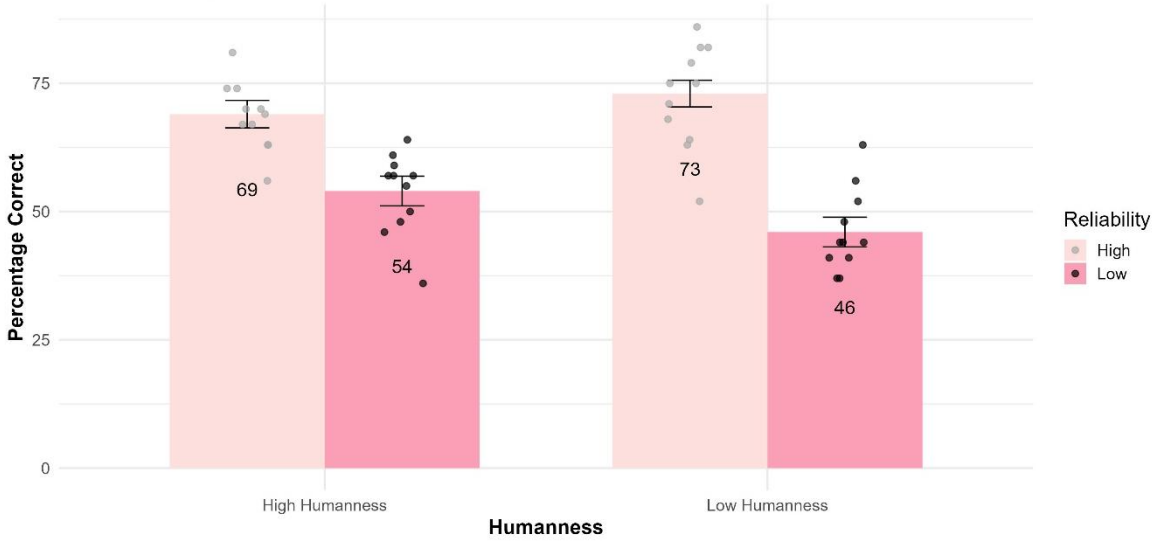
We conducted post hoc analyses using estimated marginal means with the emmeans package (Lenth, 2024). We applied Tukey's method for multiple comparisons to control the family-wise error rate and discover more about the different interactions between conditions. We used Kenward-Roger method for degrees of freedom.

3.3 Results

3.3.1 Condition Performance

Across conditions, performance did differ; to assess performance, we focused on the number of correct answers submitted by the participant. We expected to see a difference between reliability. However, we also found a difference in humanness Figure 7 presents the differences. For performance on the location task, the percentage of correct answers given were HH & HR, 69%, HH & LR, 54%, LH & HR, 73% and LH & LR, 46%.

Figure 7. A bar plot illustrating the percentage of correct responses across reliability and humanness.



Percentages of correct answers were used to ensure comparability across participants, as they account for any variations in the number of questions answered and provide a standardised measure of performance. A two-way ANOVA was conducted to examine the main effects of Reliability (High, Low) and Humanness (High, Low), as well as their interaction, on the Percentage of Correct Answers. The results revealed a significant main effect of Reliability ($F(1,40) = 64.63$, $p < .001$). The main effect of Humanness was not significant ($F(1,40) = 0.92$, $p = .344$), suggesting that Humanness did not influence the Percentage of Correct Answers given by participants. There was a significant Reliability × Humanness interaction, $F(1,40) = 5.43$, $p=.025$). The Tukey post-hoc analysis revealed several significant differences in the percentage of correct answers and are presented in Table 3.

Table 3. Tukey Post Hoc Analysis for Percentage of Correct Answers Using HSD P adjustment.

Group One	Group Two	Diff	Lower	Upper	P Adjusted
LH:HR	HH:HR	-14.36	-23.90	-4.83	0.001
HH:LR	HH:HR	3.45	-6.08	12.99	0.767
LH:LR	HH:HR	-22.64	-32.17	-13.10	<0.001

Group One	Group Two	Diff	Lower	Upper	P Adjusted
HH:LR	LH:HR	17.82	8.28	27.36	<0.001
LH:LR	LH:HR	-8.27	-17.81	1.27	0.1093
LH:LR	HH:LR	-26.09	-35.63	-16.55	<0.001

Note. **Bold** result indicates significance. Group One and Two refer to the conditions being compared where High Humanness with High Reliability (HH:HR), High Humanness with Low Reliability (HH:LR), Low Humanness with High Reliability (LH:HR), Low Humanness with Low Reliability (LH:LR). These findings provide support for H1, indicating that higher levels of anthropomorphism and reliability jointly predicted higher task performance.

3.3.2 Teammate Validity Check

We asked participants if they believed the AI responses to be AI-generated and human responses to be human; overall, 93% believed the AI to be AI, and 81% believed the human teammate to be human. This shows that our Wizard of Oz methodology was successfully implemented, and participants believed they were working with an AI and a human teammate.

3.3.3 Descriptive Statistics

This section provides an overview of the means and standard deviations within the data for Confidence, Influence, AI Performance, and Human Performance. These statistics are reported for high and low reliability and humanness. displays the information.

Table 4. Descriptive Statistics for AI & Human Performance Ratings, Influence and Confidence Ratings.

Measure	Reliability	Humanness	M	SD
AI Performance	Low	Low	74.95	19.44

		High	76.95	15.19
	High	Low	87.16	11.64
		High	79.00	14.00
Human Performance	Low	Low	71.40	18.64
		High	73.13	17.44
	High	Low	77.24	14.71
		High	72.40	16.19
Influence	Low	Low	61.51	27.73
		High	65.20	22.31
	High	Low	76.84	20.70
		High	68.84	21.62
Confidence	Low	Low	65.35	19.88
		High	69.70	18.23
	High	Low	79.63	16.91
		High	70.73	20.18

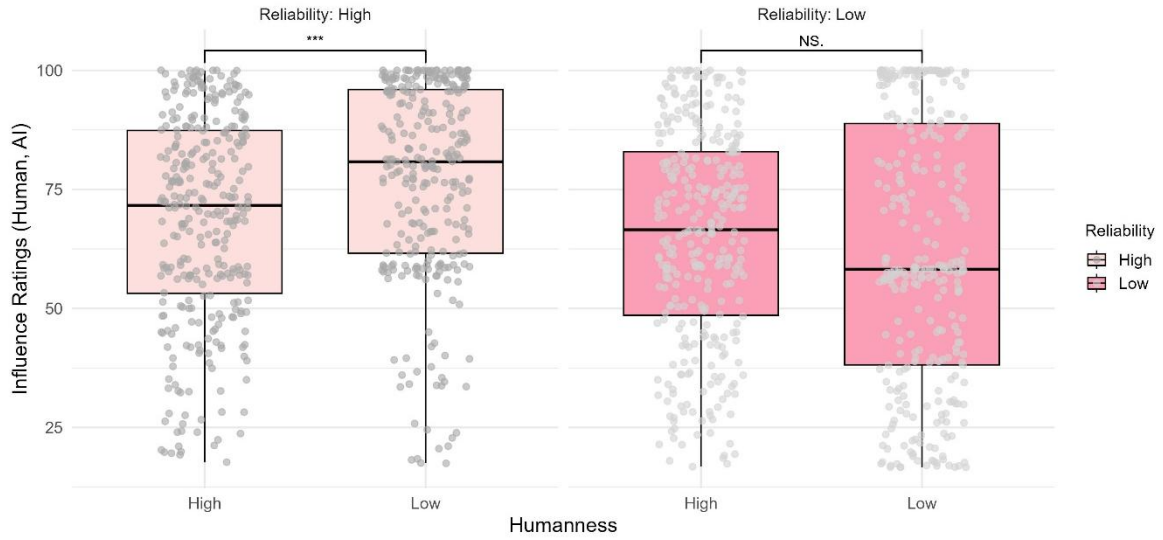
3.3.4 Propensity to Trust Questionnaire

A Pearson's product-moment correlation was conducted to assess the relationship between participants' mean trust ratings of AI and their mean propensity to trust ratings. The correlation was not statistically significant, $r(39) = 0.13$, $p = 0.42$.

3.3.5 Influence (Implicit Trust) Ratings

The analysis revealed no significant main effect of reliability, β=−3.56, SE=3.87, t(38.57)=−0.92, p=.363, suggesting that Reliability did not significantly affect trust ratings. There was no significant difference for humanness, β=7.67, SE=3.96, t(38.60), p=.0603. The interaction between reliability and humanness was significant, β=−11.68, SE=5.56, t(39.06)=−2.10, p=.0421, suggesting that the effect of reliability on trust ratings depended on the level of humanness. Figure 8 visualises these findings.

Figure 8 - A Boxplot showing the differences in Trust ratings based on Reliability and Humanness



3.3.5.1 Post Hoc Comparisons

The post-hoc analyses are presented in Table 5 and revealed several significant differences between conditions. Specifically, there was a significant difference between Low Reliability, High Humanness and High Reliability, Low Humanness (p = 0.0353), with Low Reliability, High Humanness resulting in lower influence ratings. Additionally, the comparison between High Reliability, Low Humanness and Low Reliability, Low Humanness showed a significant difference (p = 0.0026), with High Reliability, Low Humanness yielding higher influence ratings than Low Reliability, Low Humanness. These results indicate that influence ratings were impacted by the interaction of reliability and humanness, with higher reliability and lower humanness leading to better performance.

Table 5. Post Hoc Analysis for Trust Ratings

Group One	Group Two	B	SE	df	t value	p.value
HH:HR	HH:LR	3.56	3.87	38.4	0.92	0.7941
HH:HR	LH:HR	-7.67	3.96	38.4	-1.935	0.2306
HH:HR	LH:LR	7.57	3.9	39.3	1.942	0.2276
HH:LR	LH:HR	-11.23	3.96	38.4	-2.833	0.0353
HH:LR	LH:LR	4.02	3.9	39.3	1.029	0.7334
LH:HR	LH:LR	15.24	4	39.3	3.813	0.0026

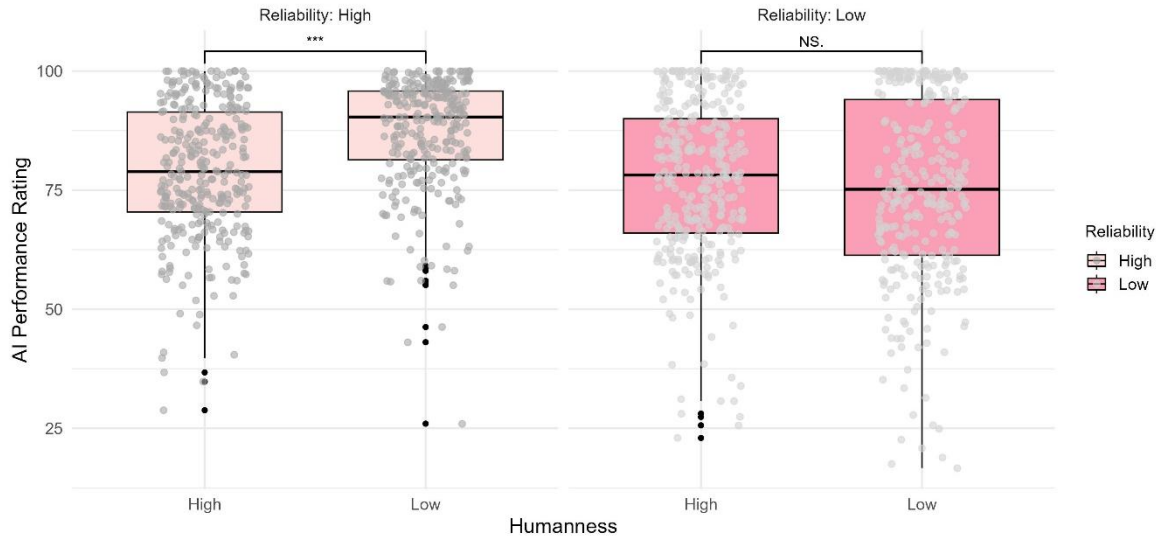
Note. **Bold** result indicates significance. Group One and Two refer to the conditions being compared where High Humanness with High Reliability (HH:HR), High Humanness with Low Reliability (HH:LR), Low Humanness with High Reliability (LH:HR), Low Humanness with Low Reliability (LH:LR).

This significant interaction effect provides direct support for H2.

3.3.6 AI Performance Ratings

The analysis revealed a significant main effect of humanness, β=8.2, SE=3.88, t(35.52)=2.1, p=.047, indicating that AI performance ratings were higher in the Low Humanness condition than in the High Humanness condition. However, the main effect of reliability was not significant, β=−5.89, SE=4.12, t(35.52)=−1.43, p=.162. The interaction between reliability and humanness was also non-significant, β=−9.50, SE=5.61, t(38.70)=−1.70, p=.099. Figure 9 visualises these findings.

Figure 9 A Boxplot showing the differences in AI Performance ratings based on Reliability and Humanness



3.3.6.1 Post Hoc Comparisons

The post-hoc comparisons are presented in Table 6 and revealed a significant difference (p = 0.0325) between High Reliability, High Humanness and High Reliability, Low Humanness, with High Reliability and Low Humanness yielding higher AI performance ratings. However, other contrasts were not significant.

Table 6. Post Hoc Analysis for AI Performance Ratings

Group One	Group Two	B	SE	df	t value	p.value
HH:HR	HH:LR	2.03	3.92	38.8	0.517	0.9545
HH:HR	LH:HR	-8.23	4.02	38.8	-2.05	0.1877
HH:HR	LH:LR	3.3	3.93	39.2	0.84	0.8352
HH:LR	LH:HR	-10.26	4.02	38.8	-2.555	0.0671
HH:LR	LH:LR	1.27	3.93	39.2	0.324	0.988
LH:HR	LH:LR	11.53	4.03	39.2	2.864	0.0325

Note. **Bold** result indicates significance. Group One and Two refer to the conditions being compared where High Humanness with High Reliability (HH:HR),

High Humanness with Low Reliability (HH:LR), Low Humanness with High Reliability (LH:HR), Low Humanness with Low Reliability (LH:LR).

This result aligns with H3, confirming that anthropomorphic AI was perceived as higher-performing despite identical accuracy levels.

3.3.7 Human Teammate Performance Ratings

The analysis revealed no significant main effect of reliability, $\beta=0.70$, $SE=3.62$, $t(385.79)=0.19$, $p=.847$, nor a significant main effect of humanness, $\beta=4.93$, $SE=3.71$, $t(38.82)=1.33$ $p=.191$. The interaction between reliability and humanness was also non-significant, $\beta=-6.80$, $SE=5.19$, $t(39.07)=-1.31$, $p=.197$.These findings suggest we can reject H4 as there was no significant difference between conditions.

3.3.8 Confidence Ratings

The analysis revealed no significant main effect of reliability, $\beta=-0.91$, $SE=4.56$, $t(38.19)=-0.20$, $p=.845$. The effect of humanness was insignificant, $\beta=8.81$, $SE=4.67$, $t(38.19)=1.89$, $p=.060$. The interaction between reliability and humanness was not significant, $\beta=-12.13$, $SE=6.54$, $t(38.40)=-1.85$, $p=.071$. These findings suggest we can reject H5 as there was no significant difference between conditions.

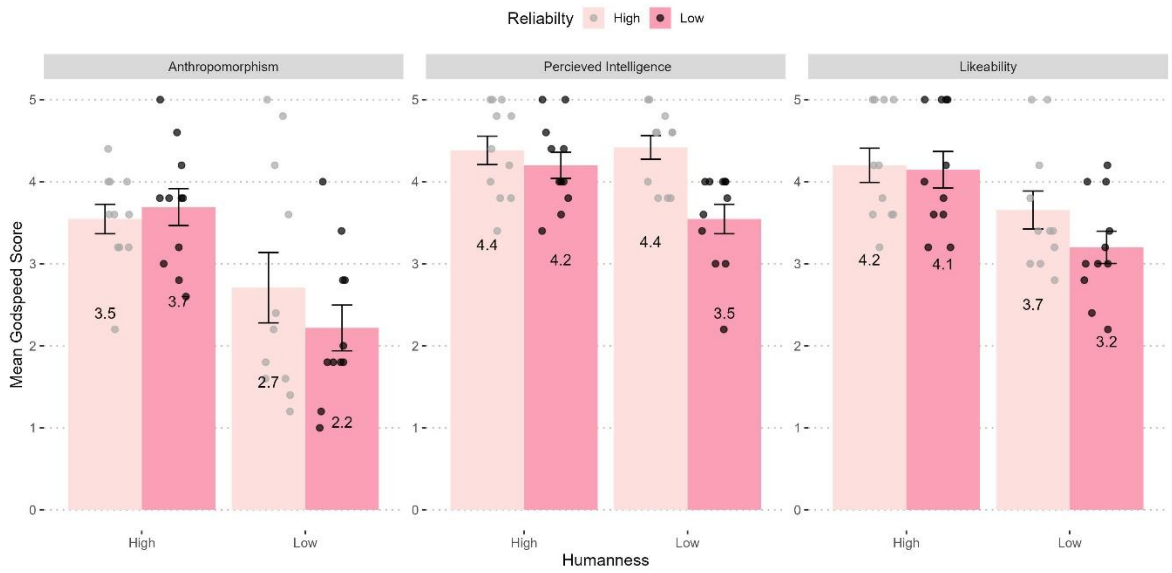
3.3.9 The Godspeed Questionnaire – Expletory Results

At the end of the experiment participants took part in the Godspeed Questionnaire, initially this was to gage feedback on the AI, however we then saw an opportunity to analyse this data further to gain deeper understanding of our results. Our study conducted two-way ANOVAs with interactions for each subsection of The Godspeed Questionnaire. The dependent variables encompassed Anthropomorphism, likeability, perceived intelligence, and perceived safety, while the independent variables included assigned reliability and humanness levels. We use ANOVAs in this context because each participant provides only a single rating, eliminating the need to account for within-subject

variability that LMMs typically address. Figure 10 is a bar chart illustrating these results.

Figure 10. Mean Godspeed Ratings for Anthropomorphism, Likeability, and Perceived Intelligence by Humanness and Reliability.

.



3.3.9.1 Anthropomorphism

We conducted a two-way ANOVA with interactions to investigate participants' Anthropomorphism ratings of AI. The dependent variable was the anthropomorphism ratings given during the experiment, while the independent variables were the assigned reliability and humanness levels. The analysis revealed a significant difference in anthropomorphism ratings based on humanness levels ($F(1, 40) = 15.49$, $p < 0.001$).

A Tukey HSD post hoc test (Table 1) revealed significant pairwise differences among the groups. The results in Table 4 show the mean differences, 95% confidence intervals, and adjusted p-values using the Tukey HSD method for these comparisons.

Table 7 - Tukey Post Hoc Analysis for Anthropomorphism Ratings Using HSD P adjustment.

Group One	Group Two	Mean Difference	Lower	Upper	P Adjusted

LH:HR	HH:HR	-0.836	-1.948	0.276	0.199
HH:LR	HH:HR	0.145	-0.967	1.257	0.985
LH:LR	HH:HR	-1.327	-2.439	-0.215	0.014
HH:LR	LH:HR	0.982	-0.130	2.094	0.100
LH:LR	LH:HR	-0.491	-1.603	0.621	0.641
LH:LR	HH:LR	-1.473	-2.585	-0.361	0.005

Note. **Bold** result indicates significance. Group One and Two refer to the conditions being compared where High Humanness with High Reliability (HH:HR), High Humanness with Low Reliability (HH:LR), Low Humanness with High Reliability (LH:HR), Low Humanness with Low Reliability (LH:LR).

3.3.9.2 Likeability

We conducted a two-way ANOVA with interactions to investigate participants' likeability ratings of AI. The dependent variable was the likeability ratings given during the experiment, while the independent variables were the assigned reliability and humanness levels. The analysis revealed a significant difference in humanness ($F(1, 40) = 11.98$, $p < 0.001$).

A Tukey HSD post hoc test (Table 2) revealed significant pairwise differences among the groups. The results in Table 5 show the mean differences, 95% confidence intervals, and adjusted p-values for these comparisons.

Table 8 - Tukey Post Hoc Analysis for Likeability Ratings Using HSD P adjustment.

Group One	Group Two	Mean Difference	Lower	Upper	P Adjusted
LH:HR	HH:HR	-0.545	-1.362	0.271	0.293

HH:LR	HH:HR	-0.055	-0.871	0.762	0.998
LH:LR	HH:HR	-1.000	-1.817	-0.183	0.011
HH:LR	LH:HR	0.491	-0.326	1.307	0.384
LH:LR	LH:HR	-0.455	-1.271	0.362	0.452
LH:LR	HH:LR	-0.945	-1.762	-0.129	0.018

Note. **Bold** result indicates significance. Group One and Two refer to the conditions being compared where High Humanness with High Reliability (HH:HR), High Humanness with Low Reliability (HH:LR), Low Humanness with High Reliability (LH:HR), Low Humanness with Low Reliability (LH:LR).

3.3.9.3 Perceived Intelligence

We conducted a two-way ANOVA with interactions to investigate participants' intelligence ratings of AI. The dependent variable was the intelligence ratings given during the experiment, while the independent variables were the assigned reliability and humanness levels. The analysis revealed a significant difference by reliability ($F(1, 40) = 10.26$, $p < 0.01$) and interactions ($F(1, 40) = 4.41$, $p < 0.05$). A Tukey HSD post hoc test (Table 9) revealed significant pairwise differences among the groups.

Table 9 - Tukey Post Hoc Analysis for Perceived Intelligence Ratings Using HSD P adjustment.

Group One	Group Two	Diff	Lower	Upper	P Adjusted
LH:HR	HH:HR	0.036	-0.588	0.660	0.999
HH:LR	HH:HR	-0.182	-0.806	0.442	0.863
LH:LR	HH:HR	-0.836	-1.460	-0.212	0.005

HH:LR	LH:HR	-0.218	-0.842	0.406	0.785
LH:LR	LH:HR	-0.873	-1.497	-0.249	0.003
LH:LR	HH:LR	-0.655	-1.278	-0.031	0.037

Note. **Bold** result indicates significance. Group One and Two refer to the conditions being compared where High Humanness with High Reliability (HH:HR), High Humanness with Low Reliability (HH:LR), Low Humanness with High Reliability (LH:HR), Low Humanness with Low Reliability (LH:LR).

3.3.9.4 Trust Ratings

To gather more insight into the feelings and attitudes of participants, we directly asked them how much they trusted the AI on a scale of 'Not at All' to 'Completely' at the end of the experiment, where a higher rating indicates a higher level of trust in the AI. We then completed an ANOVA on the results. The results indicate a significant main effect of reliability on trust ratings ($F(1,40) = 11.984$, $p = 0.001$). The main effect of humanness was not significant ($F(1,40) = 3.288$, $p = 0.078$). Table 10 summarises the results of Tukey's HSD posthoc test. The test examined pairwise differences in trust ratings between different levels of Humanness and Reliability.

Table 10 - Tukey Post Hoc Analysis for Trust Ratings Using HSD P adjustment.

Group One	Group Two	Diff	Lower	Upper	P Adjusted
LH:HR	HH:HR	-2.664	-1.045	1.045	1.000
HH:LR	HH:HR	-0.455	-1.500	0.591	0.652
LH:LR	HH:HR	-1.455	-2.500	-0.409	0.003
HH:LR	LH:HR	-0.455	-1.500	0.591	0.652

LH:LR	LH:HR	-1.455	-2.500	-0.409	0.003
LH:LR	HH:LR	-1.000	-2.045	0.045	0.065

Note. a **Bold** result indicates significance. Group One and Two refer to the conditions being compared where High Humanness with High Reliability (HH:HR), High Humanness with Low Reliability (HH:LR), Low Humanness with High Reliability (LH:HR), Low Humanness with Low Reliability (LH:LR).

3.4 Discussion

3.4.1 Overview

This study explored the effects of anthropomorphism and system reliability on trust, performance, and confidence within HATs, guided by five main hypotheses. Specifically, we examined how anthropomorphism and reliability jointly influence actual performance, trust ratings, perceived AI performance, human teammate performance perceptions, and team decision-making confidence. Participants trusted the anthropomorphised AI systems more when reliability was low. This preference indicates that human-like attributes can cushion the adverse effects of unreliable AI on trust. Conversely, when reliability was high, less anthropomorphic AI systems were rated higher in performance, suggesting that technical proficiency can overshadow the need for human-like characteristics.

3.4.2 Anthropomorphism, Reliability, and Task Performance

The results related to actual task performance showed a nuanced interaction between anthropomorphism and reliability, partially supporting H1. Although higher reliability significantly predicted better task performance, anthropomorphism alone did not produce significant main effects. However, the interaction trend suggests anthropomorphic cues could subtly enhance task performance outcomes when paired with reliable systems, emphasizing the joint importance of technical proficiency and human-like design.

3.4.3 Trust: Interaction between Anthropomorphism and Reliability

Our findings provided support for H2, confirming a significant interaction between anthropomorphism and reliability. Participants exhibited higher trust towards anthropomorphic AI systems, particularly under low-reliability conditions. This aligns with prior research suggesting that Anthropomorphism helps bridge the cognitive and emotional gap between human users and AI systems (Gambino et al., 2020; Nass et al., 1996). By fostering a sense of familiarity and relatability, anthropomorphic design may encourage team members to view the AI as a more integrated team player rather than a purely functional tool.

However, while anthropomorphic design increased trust, it was less impactful under high-reliability conditions. In these cases, anthropomorphic AI may add cognitive load, potentially reducing interdependency as users must carefully interpret the AI's responses (Döppner et al., 2019; Zhou et al., 2017). Simplified, non-anthropomorphic responses appear to enhance trust by minimising interpretive effort. Ethically, these findings underscoring the importance of transparency in AI design. When anthropomorphic cues foster higher trust in low-reliability systems, users might over-rely on the AI, a concern that highlights the ethical need to communicate AI's limitations, fostering informed trust rather than blind reliance (Binns et al., 2018; Floridi et al., 2018).

3.4.4 AI and Human Teammate Performance Perceptions

H3 predicted anthropomorphic AI would consistently receive higher perceived performance ratings, independent of reliability. However, results revealed an opposite effect: anthropomorphic AI was associated with lower performance ratings, especially under low-reliability conditions. These findings suggest that while anthropomorphic design can increase perceived trustworthiness, it may inadvertently lead to heightened expectations for AI performance. When these expectations are unmet, users may be more critical of the AI's abilities. The finding echoes concerns about anthropomorphic priming, where users attribute human capabilities to AI based on its anthropomorphic features, potentially leading to unrealistic performance expectations (Duffy, 2003; Zhang & Rau,

2022). These results suggest that designers must carefully balance anthropomorphic cues to enhance trust without inadvertently overloading the AI with unachievable user expectations.

Ethically, this also raises questions about misrepresentation in anthropomorphic AI. Overly human-like features could mislead users into expecting higher decision-making accuracy, especially in critical settings like healthcare or finance. Ethical AI design must ensure that anthropomorphic cues do not misrepresent the AI's capabilities, thereby preserving user autonomy and accurate perception of AI's role (Weisman & Peña, 2021; Coeckelbergh, 2020).

Our study explored whether human teammate performance perceptions would increase alongside perceived AI performance (H4). However, this hypothesis was not supported by the data; no significant relationship was found between participants' evaluations of human teammate performance and their perceptions of AI performance. This suggests that participants may independently assess human and AI teammates, despite previous expectations of integrated performance perceptions. Future research might further investigate conditions under which AI and human teammate evaluations become interconnected.

3.4.5 Confidence in Decision-Making and Anthropomorphism

Regarding decision-making confidence, the results show we can reject H5. Whilst participants tended to report higher confidence levels when collaborating with anthropomorphic AI, particularly under low-reliability conditions, this difference did not reach statistical significance. This trend aligns with established research showing that anthropomorphised AI enhances user engagement and trust (Waytz et al., 2014; de Visser et al., 2016). There is a chance that the results did not reach significance due to a smaller sample size and a smaller effect size. In the future it is important to develop this findings to learn more about the role of reliability confidence in AI, when Anthropomorphism is low, to support the literature emphasising the importance of accuracy and predictability for trust in AI systems (Kaur et al., 2022; Lu et al., 2022).

3.4.6 System Reliability

Reliability emerged as a central predictor of trust, with low-reliability AI consistently receiving lower trust ratings. This finding corroborates previous work suggesting that the transparency and predictability of an AI system are crucial for building user trust (Kaur et al., 2023). The high-reliability condition facilitated greater participant confidence in the AI's recommendations, even when anthropomorphic qualities were low, demonstrating that a strong reliability baseline can mitigate the need for anthropomorphic attributes. This implies that, particularly in high-stakes environments, reliable performance may be more effective in building trust than human-like design alone. These findings support the notion that trust in HATs comprises both interpersonal trust (based on human-like traits) and system trust (grounded in the AI's technical robustness) (Lewicki & Bunker, 1996).

These findings highlight the importance of designing AI systems prioritising approachability and consistency, especially when the AI will be used in settings where user autonomy and decision-making are paramount. Ethically, reliance on reliability and transparency is critical in preventing overreliance on anthropomorphic features alone. Users should be encouraged to apply their judgment, especially when interacting with AI in variable-reliability environments, helping them balance trust and scepticism effectively (Coeckelbergh, 2020; Mittelstadt et al., 2016).

3.4.7 Team Dynamics

Our findings indicate that Anthropomorphism may also positively influence team dynamics by fostering a more intuitive interaction style. High-humanness AI received higher ratings in perceived likeability, which could contribute to more seamless and cooperative teamwork. The increase in likeability aligns with previous research indicating that anthropomorphic AI may encourage open communication and improve team cohesion by mimicking human social cues (Chen & Park, 2021). It also suggests that human-like attributes enhance the overall user appeal of AI, making it a critical factor in fostering positive user experiences and interactions. Higher levels of anthropomorphism combined with

reliability led to perceptions of greater perceived intelligence, highlighting that while human-like characteristics are essential, they interact with AI proficiency to shape users' views of an AI system's intelligence.

3.4.8 Limitations and Future Research

This study has some limitations. The experimental design controlled anthropomorphic features and reliability levels, which, while necessary for isolating effects, may not fully reflect real-world AI integration into teams, where reliability can fluctuate unpredictably. Additionally, the experimental duration may not capture long-term trust dynamics, which are crucial for understanding how trust evolves over sustained collaboration. Future research could investigate these dynamics in real-world, longitudinal settings and explore other dimensions of social intelligence in AI, such as adaptive humour or empathy, to understand their effects on team trust and performance.

Furthermore, while this study focused on implicit trust measures, combining them with more explicit attitudinal data could provide a deeper understanding of users' nuanced perceptions of AI teammates. Expanding research to explore the Uncanny Valley's boundary conditions would also clarify how developers can implement Anthropomorphism before it negatively impacts trust. These directions would build upon this study's findings, enhancing our understanding of optimising HAT dynamics through AI design choices.

3.4.9 Conclusion

This chapter provides key insights into the complex interplay between anthropomorphism and reliability in HATs. While anthropomorphism enhances user engagement and likeability, especially when AI reliability is low, it must be carefully balanced with technical competence. High reliability remains fundamental to fostering confidence, underscoring the importance of performance accuracy regardless of human-like attributes. The results of The Godspeed Questionnaire further show that anthropomorphism impacts perceptions of intelligence, likeability, and overall trust, suggesting that AI design requires a nuanced approach.

Our findings also underscoring the ethical importance of transparency and accountability. While human-like attributes can foster trust and engagement, they should support, not replace, perceptions of reliability. In lower-reliability scenarios, anthropomorphism becomes essential for maintaining user trust, but it should not lead to misrepresentation or over-reliance. Clear communication of AI limitations is critical in addressing trust issues stemming from low reliability, and training users to understand AI capabilities and boundaries is essential for effective decision-making in HATs.

In summary, these findings reveal the nuanced role of anthropomorphism and reliability in shaping trust dynamics in HATs. They suggest that human trust in AI fluctuates based on perceived humanness and reliability, with implications for designing AI to foster affective and cognitive trust. Furthermore, the characteristics of AI significantly influence team cohesion and teammate evaluations, highlighting the potential for AI to shape perceptions of team dynamics and contributions. These insights reinforce the need for AI systems that balance humanness and reliability to optimise teamwork.

Chapter Four will build on these insights, exploring the use of emojis as a tool to enhance social intelligence and increase AI's emotional intelligence. This could further contribute to AI's role in fostering trusted, collaborative relationships within HATs.

Chapter 4 The Effect of Emojis and AI Reliability on Team Performance and Trust in Human-AI Teams

The rapid advancements in AI are significantly transforming sectors such as healthcare, cybersecurity, and digital forensics, where AI excels in data analysis, precision, and sustained cognitive tasks. However, human intelligence, characterised by creativity, critical thinking, emotional intelligence, and problem-solving, complements AI's strengths. Combining these capabilities, Hybrid Intelligence (HI) has emerged as a valuable framework. HATs capitalise on HI by merging the best human and machine capabilities, enabling more comprehensive decision-making and problem-solving (Kamar, 2016; Williams et al., 2022). This chapter explores how emojis as emotional cues from AI teammates impact trust, performance, confidence within HATs, and attitudes toward human teammates.

In Chapter 3, we found that while anthropomorphic features can enhance trust in AI, this effect is heavily influenced by the AI's reliability. The present chapter extends this analysis by investigating whether emojis, a nonverbal, affective cue, can serve as another means to calibrate trust in varying reliability conditions. Emojis provide a direct, accessible means of conveying emotional states in digital communication, raising the question of whether similar cues from AI might facilitate better trust calibration in HATs, particularly in instances where reliability is variable. Through an experimental approach, this study explores whether emojis can mitigate trust issues in low-reliability AI teammates or strengthen engagement with high-reliability systems. This work was presented at the Multidisciplinary Perspectives on Human-AI Team Trust Workshop at HAI23 and from this I was invited to be a guest editor on a MULTITTRUST Special Edition at the Journal of Interaction Studies where I was invited to submit this as a journal paper and it is currently under review.

4.1 Introduction

Integrating AI into collaborative decision-making has transformed healthcare, cybersecurity, and digital forensics industries. HATs exemplify this transformation by leveraging the complementary strengths of humans and AI. AI contributes

computational precision and the ability to process large datasets, while humans bring creativity, ethical judgment, and adaptability to nuanced decision-making tasks. HI, can outperform human or AI capabilities alone by merging these distinct strengths (Kamar, 2016; Williams et al., 2022). However, achieving the full potential of HATs hinges on overcoming key challenges, particularly around trust, reliability, and communication within these teams.

The dynamics of trust in HATs are critical yet complex. Trust calibration, which balances reliance and scepticism, is essential for effective collaboration (de Visser et al., 2020). Low trust can lead to underutilisation of AI's capabilities, while blind trust increases the risk of errors when AI reliability falters. While system reliability is a foundational driver of trust, relational factors like communication style and emotional engagement also significantly shape user perceptions (Bansal et al., 2019; Kaur & Sharma, 2021). Building on these insights, this chapter explores the novel use of emojis as affective cues from AI to facilitate trust calibration in HATs, particularly in contexts of variable reliability.

In this introduction, we will explore the role of HATs in leveraging the complementary strengths of humans and AI, focusing on the critical role of trust and reliability in these collaborations. We will examine how emojis, as affective cues, can influence trust calibration and team dynamics. This discussion will provide the foundation for the study's hypotheses, investigating the impact of emojis on trust, performance perceptions, and decision-making in HATs.

4.1.1 Human-AI Teams: Bridging Human Expertise and AI Precision

HATs are a practical embodiment of HI. By blending the computational precision of AI with the adaptability and contextual understanding of humans, HATs have transformed industries reliant on decision-making under complexity. For instance, in digital forensics, human investigators leverage AI for tasks such as data processing, pattern recognition, and geolocation, allowing them to focus on contextual analysis and ethical decision-making. This collaboration enhances the speed and accuracy of investigations, making it possible to trace cyberattack origins and combat threats with unprecedented efficiency (Costantini et al., 2019).

Similarly, the medical field benefits from AI-augmented diagnostics, where integrating computer-aided tools improves patient care and sparks further interest in understanding the nuances of human-AI collaboration (Kunar & Watson, 2023).

Rix (2022) identifies four essential drivers for building successful HATs. First, HATs require at least one human and one AI team member, with much of the existing research focused on relatively simple configurations. As the number of team members increases, so does the complexity of human-AI social dynamics (Liang et al., 2019). Second, establishing a shared and valued goal among all team members, human or AI, is critical to fostering cohesion and collaboration (Chai et al., 2017; McNeese et al., 2018). Third, interdependency among team members ensures that outcomes are mutually influenced, though disproportionate benefits may lead to tension or conflict (Chiou et al., 2019). Incentives for collaboration, paired with the cognitive relief provided by AI's support, can promote interdependency and cooperation (Döppner et al., 2019). Finally, defining roles based on each member's unique strengths enhances synergy. Even when roles are not unique, clear role delineation remains a cornerstone of successful teamwork (Oh et al., 2018).

While these drivers provide a useful framework, the intricacies of HAT dynamics require further investigation. For instance, challenges often arise when delegating tasks between humans and AI (Fügener et al., 2022; Pinski et al., 2023), particularly when AI is designed to resemble human teammates. Research shows that humans are often unreceptive to AI in positions of authority, such as a "humanised AI boss", which complicates team structures and goal alignment (Yam et al., 2022). Additionally, in larger teams, where multiple humans collaborate with a single AI teammate, the AI may function more as a tool than an equal partner, further emphasising the need to explore diverse configurations and team dynamics (Schelble et al., 2022).

4.1.2 Navigating Complexity in HAT Dynamics

Several studies highlight the complex variables in HATs, underscoring the need for careful trust calibration. Interestingly, when paired with transparency, low-confidence AI can improve team performance by helping humans form accurate

mental models of the AI's capabilities and limitations (Bansal et al., 2021). This enables human teammates to anticipate potential AI errors, adapt their behaviours, and foster more effective collaboration (Bansal et al., 2019). Conversely, research reveals that humans often treat AI teammates differently than human counterparts, impacting team dynamics.

A recent scoping review by Berretta et al. (2023) emphasises the importance of adopting a human-centric, socio-technical approach in designing HATs. This perspective shifts AI's role from a passive tool to an active teammate, considering both the technical capabilities of AI and the relational needs of human team members. Despite advancements in HAT research, gaps remain, particularly in understanding the impact of complex team dynamics, performance variables, and user-centered designs. Addressing these gaps is critical to advancing trust calibration strategies and ensuring that HATs operate as cohesive, effective teams.

This chapter builds on these insights by investigating the role of emojis as affective cues in HATs. Emojis offer a lightweight and accessible way to humanise AI interactions, potentially enhancing relational trust while supporting trust calibration. By exploring how emojis influence perceptions of trust, performance, and team dynamics, this research contributes to the broader goal of designing emotionally intelligent AI systems capable of fostering meaningful collaboration in HATs.

4.1.3 Performance and Reliability

When creating HATs, the performance and reliability of the AI system can also be crucial when trying to obtain peak performance. A logical approach towards designing HATs is that the better the performance and reliability of an AI system, the more successful the team will be. However, this is incorrect (Bansal et al., 2021; Bansal et al., 2019). Research has found that whilst performance is essential, high system performance does not directly result in better HAT performance; it would appear that better outcomes are present from antecedents such as explainability and clarity (Endsley, 2023; Guidotti et al., 2018; Kim et al., 2023; Ribeiro et al., 2016).

It is also essential for AI developers to keep the rest of the team in mind when developing updates for an AI embedded in a team. Research has shown that HAT performance can drop dramatically after an AI system receives an update due to the other team members learning where the AI has strengths and weaknesses (Bansal et al., 2019; Bansal et al., 2019). When updates are made without updating the AI, the whole team's performance can dip as the users no longer know how to interact with it. When implementing updates, any changes in performance and reliability must be explicitly explained to other team members so interactions remain successful.

Finally, the research has indicated that it is possible to mediate the reliability of AI by manipulating other features of the AI. Research has shown that when anthropomorphising AI, users can find AI to be more likeable and rate its performance as being higher than it truly is (de Visser et al., 2016; Kulms & Kopp, 2019). The previous chapter demonstrated that anthropomorphism can buffer against the detrimental effects of low reliability, helping maintain user trust. However, the same anthropomorphic features could increase cognitive load and reduce efficiency in high-reliability scenarios, as participants expended more effort interpreting the AI's human-like responses. These findings underline that the interplay between performance, reliability, and user perceptions is highly context-dependent, with trust calibration emerging as a pivotal factor in aligning human expectations with AI capabilities.

4.1.4 Human-AI Teams and the Role of Trust

Despite their promise, HATs' success depends on fostering effective collaboration and trust. Trust calibration, balancing reliance on AI and healthy scepticism, is crucial for optimal team performance (de Visser et al., 2020). Too high trust can lead to over-reliance, resulting in complacency and potential errors when AI systems falter. Conversely, insufficient trust can cause users to underutilise AI, limiting its potential contributions (Kamar, 2016).

Trust is a complex concept; many researchers have different approaches and definitions (Ulfert et al., 2023). For this introduction and the rest of the experiment, we define trust as a willingness to be vulnerable where there is a risk.

We approach trust from an organisational team perspective and a trust in AI/technology approach, and these were considered when discussing relevant literature and selecting the measures.

Research emphasises the dual dimensions of trust in HATs: system trust, based on the reliability, predictability, and explainability of AI, and interpersonal trust, which derives from relational factors such as likability and emotional engagement (Jacovi et al., 2021; Schmidt et al., 2020). For example, in team settings, high AI reliability often fosters confidence, but this trust is amplified when users perceive the AI as approachable and responsive to their needs (Bansal et al., 2019).

While trust calibration is essential, achieving it is challenging due to the "black-box" nature of many AI systems, where decision-making processes are opaque and complex to interpret (Guidotti et al., 2018)). Moving toward explainable AI (XAI) methods has shown promise in enhancing transparency, but these efforts often neglect emotional and social dynamics that also influence trust. One approach that could improve trust calibration is to design AI with elements of Emotional Intelligence (EI) (Salovey & Mayer, 1990). EI involves perceiving emotions accurately, regulating emotions, and utilising emotional information to navigate social interactions and make thoughtful decisions, and it can increase human team performance (Ghosh et al., 2012; C. Lee & Wong, 2019).

Designing AI with a sense of EI is a very complex task as it involves designing AI that can regulate its own emotions and detect and react appropriately to other teammates' emotions. Despite this challenge, it could be an asset for developing successful HATs. Adding affective cues like emojis may help bridge this gap by creating more intuitive and human-like interactions that build relational trust.

4.1.5 Emojis

Emojis can be a significant factor in enhancing EI within professional settings involving HATs. Emojis facilitate more intuitive and emotionally responsive interactions between humans and AI by enabling AI to convey emotional states. This not only fosters trust and engagement but also allows AI systems to better interpret and respond to users' emotional cues, promoting more cohesive and

cooperative team dynamics (Ahanin & Ismail, 2022; Beattie et al., 2020; Fadhil et al., 2018; Hamza, 2016).

Research has demonstrated the potential of emojis to function as affective signals that bridge the gap between human and AI teammates. For example, studies on platforms like Twitter have shown that emojis can model and infer affective states based on usage patterns (Ahanin & Ismail, 2022). In the context of HATs, such capabilities could facilitate bi-directional emotional intelligence, where human and AI teammates gain insights into each other's affective states, enabling more effective collaboration. By integrating emojis into AI systems, designers could create more natural interactions that align with users' emotional needs, potentially reducing frustration and increasing satisfaction in collaborative tasks.

Expanding the focus to health-related applications, the research underscores the value of emojis in improving user experiences and fostering trust in AI-mediated interactions. For instance, chatbots that inquire about users' mental well-being have achieved higher ratings for user enjoyment, attitude, and confidence when emojis are incorporated into their responses (Fadhil et al., 2018). Notably, messages containing emojis from chatbots are rated as trustworthy and credible as those from human senders, illustrating the potential of emojis to humanise AI interactions (Beattie et al., 2020). In addition, human and AI senders using emojis are perceived as more socially appealing, competent, and credible in computer-mediated communication than those relying solely on text-based messages (Beattie et al., 2020).

These findings suggest that emojis can be pivotal in humanising AI, enhancing its perceived emotional intelligence while maintaining clarity and simplicity. In the professional domain, these benefits translate to more engaging and productive collaborations, as users are more likely to trust and rely on AI systems that demonstrate social and emotional awareness.

The type of emojis used and the timing of their deployment are critical variables. Positive emojis, such as smiley faces, may foster warmth and rapport, while neutral or task-specific emojis may better maintain professionalism in more formal

settings. Research has yet to fully explore how users interpret different emojis in collaborative HAT environments, suggesting a need for more nuanced studies that investigate how affective signals interact with trust, performance, and team dynamics.

4.1.6 Aims of Current Study

The current study aims to investigate the influence of emojis in AI responses on the decision-making process, perceived performance of the AI, and overall trust in HATs. The study seeks to address gaps in the existing literature, exploring the nuanced dynamics of trust calibration and teaming in HATs. From the existing literature, we propose the following hypotheses:

H1: The use of emojis will lead to improved actual task performance by participants.

H2: The use of emojis by AI teammates will result in significantly higher trust ratings compared to conditions in which no emojis are used.

H3: The use of facial emojis will elicit greater trust from participants than the use of icon-based emojis or no emojis.

H4: Emoji use will increase perceived teammate performance ratings, with this effect being particularly pronounced under low-reliability conditions.

H5: The use of emojis will increase Godspeed Perceptions of the AI.

4.2 Methods

4.2.1 Participants

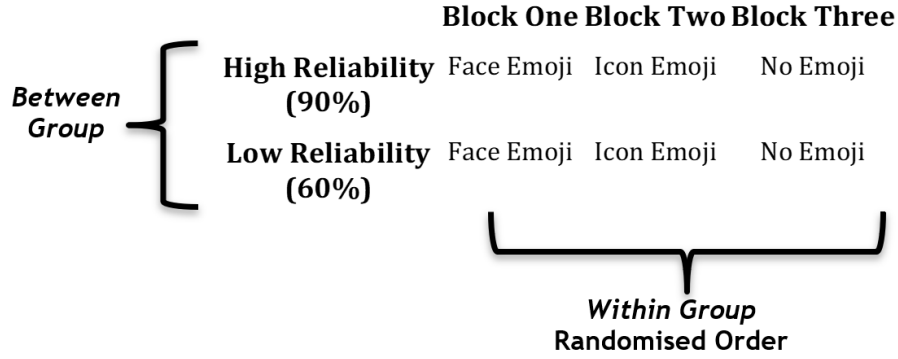
A total of 43 participants were involved in the study. The mean age of the participants was 23.12 years (SD = 3.26). The sample included two genders: 17 participants identified as female, and 26 identified as male. Participants reported a variety of occupations. Most participants were students (n = 26), with the remaining being in a mix of full-time employment (n = 17). Participants

represented diverse backgrounds consolidated into broader categories. The majority identified as white (n = 30), followed by Asian (n = 5), Black (n = 6) and mixed race (n = 2). The study received full ethical clearance from the MVLS Ethics Committee (application: 200220361) at the University of Glasgow. All participants were compensated with an Amazon voucher for participating in the study; the amount varied on the time taken to complete the study, but the rate was £6p/h pro rata.

4.2.2 Study Design

This study employed a mixed between-within-subjects design (2x3 configuration) to examine participants' trust in an AI and a human teammate under varying reliability and emoji conditions. Participants were randomly assigned to one of two reliability conditions for the AI teammate: High Reliability (90%) or Low Reliability (60%), with a human teammate consistently exhibiting Low Reliability (30%). Participants were exposed to three emoji modalities within these groups across three blocks: Face Emojis, Icon Emojis, or No Emojis. The order of these blocks was randomised to control for order effects. Figure 11 shows this design.

Figure 11. The design of the experiment. Participants were randomly assigned to either High or Low reliability. Participants then experienced a block of each emoji type or no emoji in a randomised order.



Participants completed 30 trials, divided into three blocks of 10 trials each. Each trial required participants to identify a specific geographic location, and the AI and human teammates provided conflicting responses 95% of the time. Participants were tasked with evaluating the reliability and performance of both teammates and making final decisions under time constraints.

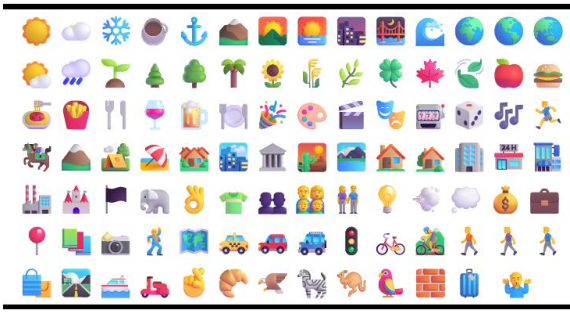
4.2.3 Materials

4.2.3.1 Developing Response Stimuli

We adopted a Wizard of Oz experimental method to ensure efficient development and optimal control. Participants were led to believe they were collaborating with both an AI and a human teammate. However, they were interacting with responses generated by ChatGPT (OpenAI., 2024). The task involved presenting participants with random locations extracted from Google Earth. They aimed to determine the continent, country, and city associated with each location. Responses were created by prompting ChatGPT with instructions such as *'In the style of a human playing a game of Geoguessr, describe the location of [location coordinates] including city, country, and continent.*' This process was repeated for all 30 locations. To embed emojis in the text, we prompted ChatGPT to add these in the following paragraph: *'Please add relevant face/icon emojis to this paragraph of text'.* In the conditions with emojis present, 5-7 emojis were distributed throughout the answers. Icon emojis were relevant to the location being guessed, such as a flag, typical weather and location cues. Face emojis primarily displayed positive emotions; a few had neutral or sadder expressions. See .

Table 11 for more details.

Table 11 This table shows the different emojis used in the experiment.

Emojis Used	
Icon	Face
	

The difference in the number of face emojis available compared to icon emojis made a more extensive selection of icon emojis appear. A few descriptions were edited for brevity to maintain consistency across conditions. Incorrect answers were generated by selecting locations similar to the correct ones to fit the context but not wrong, for instance, by referencing languages not associated with the location. The human teammates' answers were successfully used in a previous

version of this experiment and were written/edited by the experimenter (Bailey & Pollick, 2023).

4.2.3.2 Decision-Making Task

The task involved presenting participants with random locations extracted from Google Earth. Participants were tasked with determining the continent, country, and city associated with each location, with the final decision resting on the participant, who assumed the role of the 'team leader'. The experiment was set up with the AI and human teammates giving different answers 95% of the time, meaning the participant had to choose between the teammates each time. The experiment comprised three blocks, one block with each emoji condition; the order of the emojis was randomised throughout the experiment to avoid order effects. A time constraint of 120 seconds per location was enforced, meaning participants had to rely on their teammates' responses to submit the location in time. The introduction of time scarcity as an environmental factor can significantly impact the outcomes of team tasks, often necessitating rapid decision-making (Hu et al., 2015; Kelly & Karau, 1999). This constraint could increase reliance on AI, requiring human teammates to make choices based on implicit attitudes rather than thoroughly deliberating on the task. To emphasise this factor, we ensured that the human and AI teammates mainly provided different answers, requiring participants to choose which teammate they trusted the most.

4.2.3.3 Attention Check

At the start of the experiment, participants were instructed to choose the option labelled 'A Dog' if they were given a question with the options of 'A dog' or 'A cat'; a failure of two or more attention checks meant the participants' data would be removed from the dataset, there were six attention check in total, which were randomly distributed within the questionnaire data at the end of each block. From these attention checks, we removed 3 participants (n = 40).

4.2.3.4 Questionnaires

Propesinty to Trust Machines (Merritt et al., 2013): A series of 6 questions where participants rated on a 7-point Likert scale how likely they are to trust machines.

The Godspeed Questionnaire (Bartneck et al., 2009) This questionnaire assesses human perceptions of AI across five dimensions: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Each dimension is rated using a set of bipolar scales (e.g., from "very human-like" to "not human-like at all") on a 5- or 7-point Likert scale. As we were using disembodied AI, we removed the animacy sub-section and replaced the term 'robot' with 'AI'.

Trust in Automation Questionnaire (Körber, 2019)The Trust in Automation Questionnaire (TiA) is a self-report survey in which participants rate their perceptions of an automated system across several dimensions. Participants respond to a series of statements using a Likert scale (e.g., 1 = strongly disagree to 5 = strongly agree). The questionnaire evaluates different factors of trust: Trust, Familiarity, Understanding, Intentions of developers, Reliability of AI and Propensity to Trust. The ratings from these responses provide insights into how much trust the participant places in the system, allowing researchers or designers to assess trust levels and identify areas for improvement in automation design. We lightly altered the questionnaire to address an AI rather than an automated system.

Questions During Each Trial: Participants were presented with four visual analogue scales (Sung & Wu, 2018) with two endpoints on every trial. The first question asked, "How much do you trust the AI?" with responses ranging from 1 (Not at All) to 7 (Completely), allowing participants to indicate their level of trust in the AI teammate if their influence leaned slightly more towards AI than human, they might press the scale at around 6. This implementation was applied to all sliders on the experimental interface. The second question, "Which teammate influenced you most?" used a scale from 1 (Human) to 7 (AI) to gauge which teammate had a more substantial impact on the participant's decision-making. Participants were then asked, "Please rate the performance of the Human" with

ratings from 1 (Terrible) to 7 (Perfect), and finally, "Please rate the performance of the AI" using the same scale, from 1 (Terrible) to 7 (Perfect). Participants had the freedom to click anywhere along the scale. These scales provided a comprehensive assessment of participants' trust and perceptions of each teammate's performance and were completed on every trial of the experiment.

4.2.4 Procedure

Participants accessed the experiment via a link emailed by researchers after signing up through the university participant pool. The initial link directed participants to a Qualtrics form containing information about the experiment and a consent form. Once consent was obtained, participants were provided with a link to the Pavlovia experiment, where they entered relevant demographic information and began the experiment. At the start of the experiment, participants completed the PtTM (PTM) questionnaire (Merritt et al., 2013) to establish a baseline of their attitudes towards automation. Participants all received the exact instructions, where the AI and Human teammate were introduced and the task explained, including attention checks. Participants were made aware that they were the team leader and had the final decision-making authority. After this, there was an example trial where participants could learn where to input relevant information and interact with the interface. Once participants had completed the example trial, they were warned that the experiment would begin shortly and that they should email the researcher if they had any questions.

Participants identified 30 locations across three blocks, each comprising ten trials per block. Each trial included one location and four slider bars, and the participants were asked to input the city, country, and continent all within 120 seconds. Notably, the AI and human teammates often provided conflicting answers, challenging participants to choose which teammate they trusted more. Following each trial, the correct answer was revealed, allowing participants to assess the performance of the human and AI teammates. The task was designed to be challenging; previous versions of this experiment have been found to work efficiently(Bailey & Pollick, 2023). The blocks featured either Face Emojis (☹️, 😊), Icon Emojis (🧳 , 🪶), or No Emojis, presented in a randomised order to control

for order effects; between each block, there was a 90-second break so participants could rest. At the end of each block, participants completed the Trust in Automation Questionnaire (Körber, 2019) and the Godspeed questionnaire (Bartneck et al., 2009), slightly modified to fit the zero-embodiment scenario. The study concluded with a full debrief, which provided participants with a complete understanding of the experiment's nature and reminded them of their right to withdraw if they felt uncomfortable with the Wizard of Oz approach.

4.2.5 Developing the Linear Mixed Model for Analysis

We selected a linear mixed-effects model (LMM) due to its ability to handle a hierarchical data structure. Our data includes multiple observations (trials) nested within participants, introducing non-independence. LMMs appropriately account for this by including random intercepts for participants. LMMs also allow us to model fixed effects for experimental conditions (emoji type and reliability) while controlling for individual variability through random effects. The design involves repeated trust and performance ratings across multiple trials, making LMMs suitable for capturing within-subject variability. Alternative methods, such as traditional ANOVA, would not adequately account for participant-level random variability and could inflate Type I error rates.

To perform this analysis on the AI performance, human performance, trust ratings and influence ratings taken on every trial and the questionnaires at each block, we utilised LMMs using the lme4 in R-Studio (Bates et al., 2015) and used lmerTest (Kuznetsova et al., 2017) to complete Type III ANOVA with Satterthwaite's method for degrees of freedom to extract p-values. A linear mixed model extends the traditional linear regression model that accounts for fixed and random effects (Barr, 2013)**.** When developing the model we implemented trial as a random effect but found it had little variance and reduced the model's fit.

4.2.5.1 Model Specification

To analyse the impact of reliability and humanness on performance across different measures, we utilised the following linear mixed model:

$$y_{ij} = \beta_0 + \beta_1 rel_i + \beta_2 emoji_i + \beta_3 (rel_i \times emoji_i) + u_{0j} + e_{ij}$$

In this model, this is the breakdown of each component:

- y_{ij} is the response variable for the ith observation of the jth participant.
- β_0 is the intercept.
- $\beta_1 rel_i$ is the coefficient for the fixed effect of reliability.
- $\beta_2 emoji_i$ is the coefficient for the fixed effect of Emoji Type.
- $\beta_3 (rel_i \times emoji_i)$ is the coefficient for the interaction between Reliability and Emoji Type.
- u_{0j} represents the random effect for participant j, which accounts for the variation in the intercept across participants.
- e_{ij} is the residual error term for the ith observation of the jth participant.

4.2.5.2 Post Hoc Analysis

To further explore all possible pairwise comparisons and better understand the interactions between conditions, we conducted post hoc analyses using estimated marginal means with the emmeans package (Lenth, 2024). We applied Tukey's method to control the family-wise error rate during multiple comparisons.

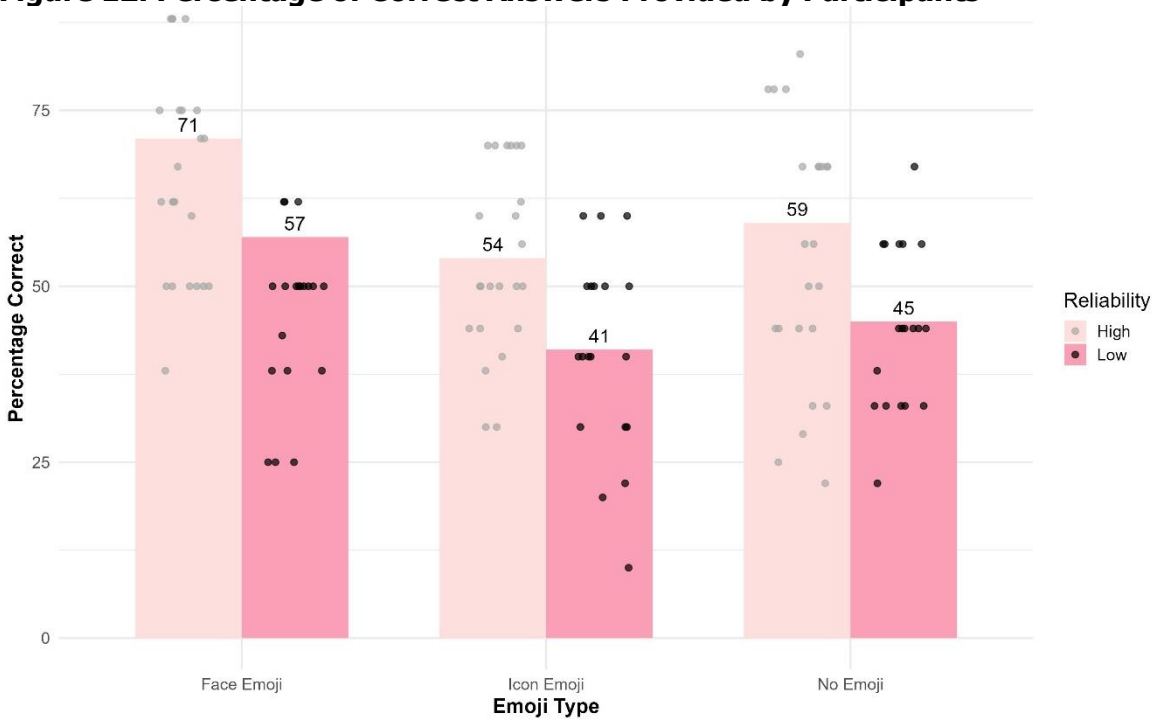
4.3 Results

4.3.1 Condition Performance

Figure 12 illustrates the percentage of correct responses categorised by emoji type and AI reliability. Across all emoji types, performance is higher in high-reliability conditions than in low-reliability conditions. Still, the low-reliability condition matched AI reliability more closely than the high-reliability condition. This data

suggests that AI reliability significantly influences accuracy and that face emojis perform best in highly reliable contexts.

Figure 12. Percentage of Correct Answers Provided by Participants



A two-way Analysis of Variance (ANOVA) was conducted to examine the effect of Reliability, Emoji Type and Interactions on the percentage of correct answers. The main effect of Reliability was significant, with a large difference in the percentage of correct answers between High and Low Reliability conditions ($F(1, 40) = 19.78$, $p < 0.001$). This indicates that participants performed better when Reliability was high. The main effect of Emoji Type was not significant ($F(1.88, 75.25) = 5.89$, $p = 0.06$). The interaction between Reliability and Emoji Type was also non-significant ($F(1.88, 75.25) = 1.80$, $p = 0.176$), indicating that the effect of Reliability on performance did not differ significantly across Emoji Types. This finding did not support H1; emoji usage by AI teammates did not significantly enhance actual participant task performance.

4.3.1.1 Post-Hoc Analysis

Post-hoc comparisons using Tukey's HSD method were conducted to explore the differences between groups further, particularly for Reliability and Emoji Type levels. Pairwise contrasts revealed that participants performed significantly better

in the high-reliability Face Emoji condition than in the low-reliability Face Emoji condition ($p = 0.01$, difference = 11.77%). Performance was also significantly higher in the high-reliability Face Emoji condition compared to the low-reliability Icon Emoji condition ($p = 0.011$, difference = 16.09%), and the reliability Face Emoji condition outperformed the reliability No Emoji ($p = 0.004$, difference = 17.19%). No other between-condition comparisons were significant ($p > 0.05$). Under low reliability, no significant differences were found between conditions.

4.3.2 Descriptive Statistics

On average, participants took about 41 minutes to complete the experiment (m = 41.7, sd =19.1). Table 12 shows the mean and standard deviation of influence, trust, AI performance, and human performance ratings taken throughout the experiment.

Table 12. The Mean and Standard Deviation Score for Trust, Influence and Performance by Reliability and Emoji Type.

Measure	Reliability	Emoji Type	M	SD
Trust	High	Face Emoji	64.59	15.28
		Icon Emoji	64.70	14.36
		No Emoji	63.37	15.71
	Low	Face Emoji	64.68	8.67
		Icon Emoji	66.03	11.26
		No Emoji	65.36	12.39
Influence	High	Face Emoji	65.39	17.81
		Icon Emoji	63.82	18.30

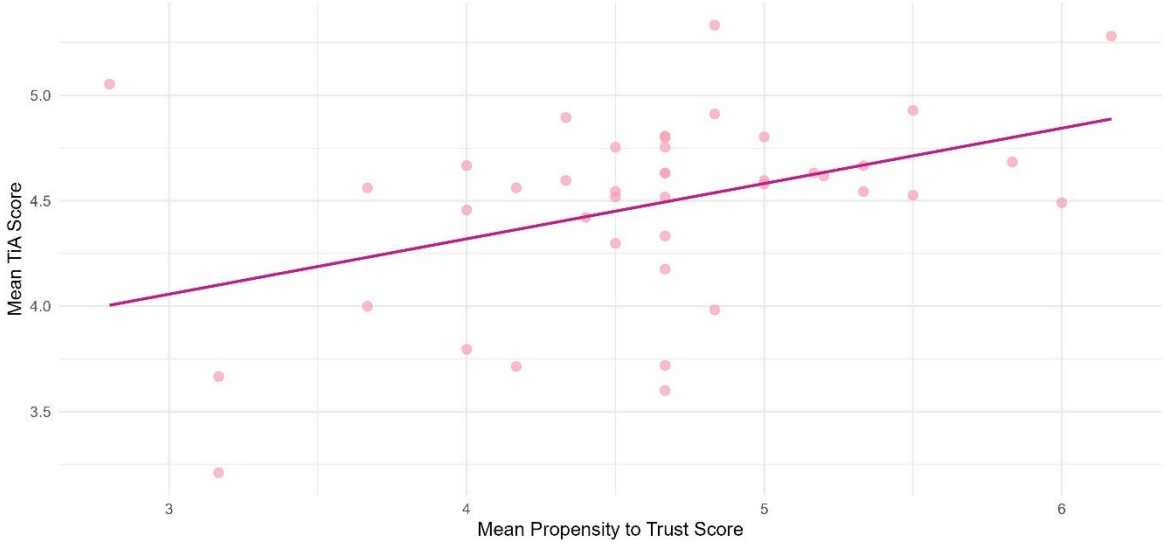
		No Emoji	66.69	19.06
	Low	Face Emoji	66.22	19.16
		Icon Emoji	66.00	19.87
		No Emoji	64.53	20.89
AI Performance	High	Face Emoji	67.53	15.54
		Icon Emoji	66.13	16.35
		No Emoji	66.20	18.31
	Low	Face Emoji	72.70	13.87
		Icon Emoji	72.02	12.89
		No Emoji	70.19	13.42
Human Performance	High	Face Emoji	63.77	15.70
		Icon Emoji	64.79	13.83
		No Emoji	60.19	15.73
	Low	Face Emoji	61.89	10.99
		Icon Emoji	63.22	10.97
		No Emoji	60.96	12.65

4.3.3 Propensity to Trust

Pearson's product-moment correlation assessed the relationship between participants' mean Propensity to Trust Machines (PtM) (S. M. Merritt et al., 2013) ratings and the mean Trust in Automation (TiA) (Körber, 2019) ratings. We found

a significant positive correlation between the mean propensity to trust rating and the TiA rating, r(40) = 0.415, t = 2.886, p = 0.006. Figure 13 shows the scatter plot for the relationship with a line of best fit. The positive correlation suggests that as participants' mean PtM rating increases, so do their TiA ratings. It is important to note that the correlation does not imply causation, and further research is needed to explore the underlying factors contributing to this observed relationship. Throughout the experiment, we did not find a significant correlation between propensity to trust ratings and trust/influence ratings.

Figure 13. The correlation between Propensity to Trust ratings and TiA ratings.



4.3.4 Influence Ratings

A linear mixed-effects model was conducted to evaluate the effects of Emoji Type, Reliability, and their interaction on participants' influence ratings. The model included random intercepts for participants. The fixed-effects estimates and Type III ANOVA with Satterthwaite's method found Emoji Type did not significantly affect influence ratings (F(2, 1181.66) = 0.29, p = 0.749). Reliability did not show a significant main effect on influence ratings (F(1, 40.47) = 0.002, p = 0.963). The interaction between Emoji Type and Reliability was also non-significant (F(2, 1181.66) = 1.82, p = 0.162).These findings did not support H2 or H3; facial emojis did not significantly differ from icon-based emojis or no emojis in influencing implicit trust ratings.

4.3.5 Trust Rating

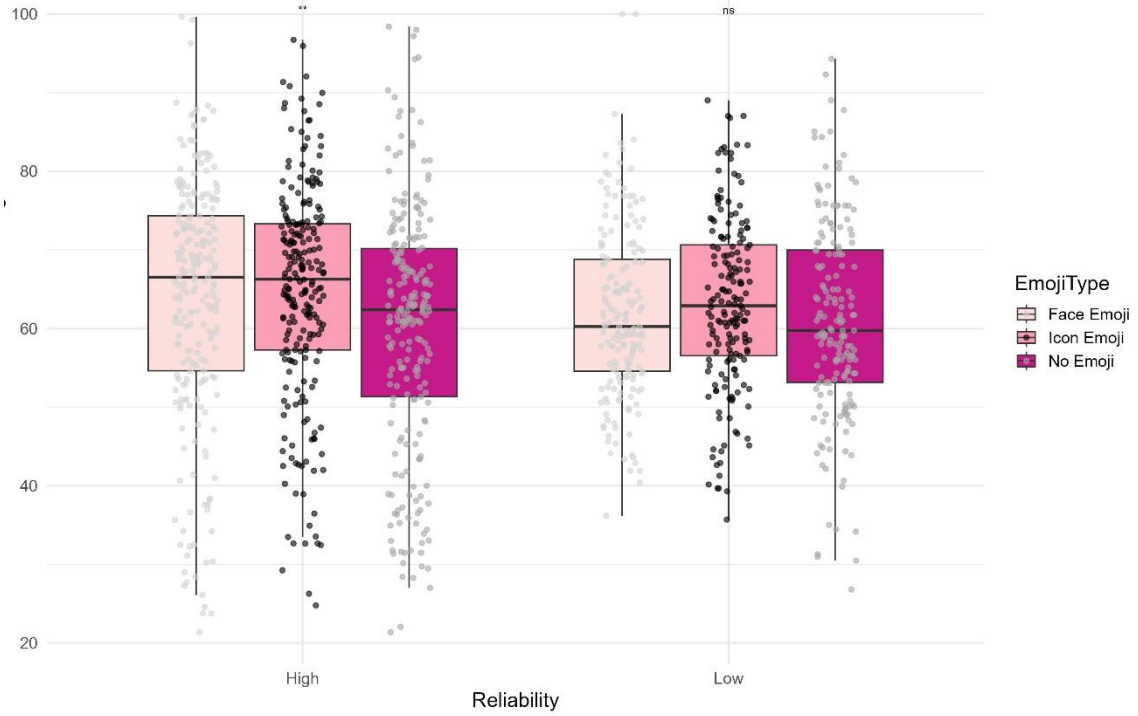
A linear mixed-effects model was conducted to evaluate the effects of Emoji Type, Reliability, and their interaction on participants' trust ratings. The model included random intercepts for participants. The Type III ANOVA with Satterthwaite's method found Emoji Type did not significantly affect trust ratings ($F(2, 1176.91) = 0.86$, $p = 0.423$). Reliability did not show a significant main effect on trust ratings ($F(1, 38.43) = 0.51$, $p = 0.480$). The interaction between Emoji Type and Reliability was also non-significant ($F(2, 1176.91) = 0.57$, $p = 0.567$). These findings did not support H2 or H3; facial emojis did not significantly differ from icon-based emojis or no emojis in influencing implicit trust ratings.

4.3.6 AI Performance Ratings

A linear mixed-effects model was conducted to evaluate the effects of Reliability, Emoji Type, and their interaction on participants' ratings of AI performance. The model included random intercepts for participants. The Type III ANOVA with Satterthwaite's method revealed Emoji Type had a significant main effect on AI performance ratings ($F(2, 1180.53) = 7.11$, $p < 0.001$). However, Reliability did not have a significant effect ($F(1, 40.34) = 0.18$, $p = 0.674$). Neither did the interaction between Reliability and Emoji Type was not significant ($F(2, 1180.53) = 1.27$, $p = 0.283$).

Post hoc comparisons with Tukey's adjustment were conducted to explore the significant effect of Emoji Type. The key findings are that the High-Reliability Face Emoji versus the High-Reliability No Emoji produced a significantly higher rating ($B = 3.478$, $p = 0.037$).High-Reliability Icon Emoji versus High-Reliability No Emoji also produced a significant significantly higher rating ($B = 4.441$, $p = 0.002$). All other comparisons were not significant (all $p > 0.05$). This outcome provided partial support for H4; emoji use modestly increased perceptions of AI teammate performance, although this effect was minimal and not clearly moderated by reliability conditions

Figure 14. A Boxplot showing the AI Teammate Performance Means and SD



4.3.7 Human Performance Ratings

A linear mixed-effects model was conducted to examine the effects of Reliability, Emoji Type, and their interaction on participants' human performance ratings. The model included random intercepts for participants. The Type III ANOVA with Satterthwaite's method found Emoji Type had no significant effect on human performance ratings ($F(2, 1179.35) = 2.26$, $p = 0.104$). Reliability also did not reach statistical significance ($F(1, 40.05) = 3.84$, $p = 0.057$). The interaction between Reliability and Emoji Type was also non-significant ($F(2, 1179.35) = 0.70$, $p = 0.498$). This outcome provided no support for H4; emoji did not lead to increased perceptions of Human teammate performance.

4.3.8 Questionnaire Data

To complement the experimental data, we also collected questionnaire data to gain further insight into the results. We used the Trust in Automation Questionnaire (Körber, 2019) and The Godspeed Questionnaire (Bartneck et al., 2009). To complete the analysis, we once again implemented LMMs using the lme4 in R-Studio (Bates et al., 2015), as participants gave responses at multiple

points throughout the experiment. The model was the same as the previous one; the questionnaire ratings were the response variables.

4.3.8.1 Trust in AI Questionnaire.

The TiA consists of six subscales, each targeting a specific aspect of trust, including reliability, understanding/predictability, familiarity, the intention of developers, propensity to trust, and overall trust in automation. Each subscale can be analysed independently, making it possible to focus on specific areas of interest. However, for a comprehensive assessment of trust in automation, it is recommended that the entire questionnaire be used. Due to the multidimensional nature of trust, calculating a total sum rating across all items is not advised, as it may lead to ambiguous interpretations. We slightly altered the scale to measure trust in AI instead; the questions remained the same, but we replaced the word automation with AI.

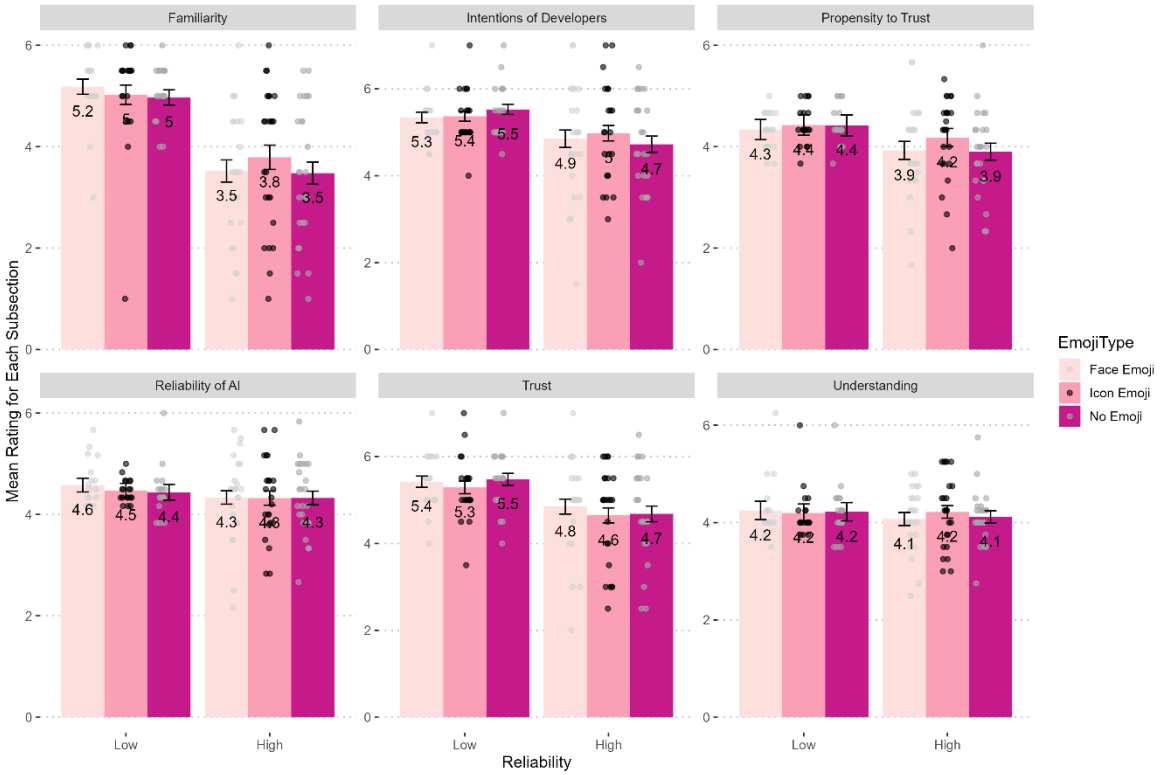
A linear mixed-effects model was conducted to examine the effect of Reliability and Emoji Type on trust ratings. The model included Reliability (High or Low) and Emoji Type (Face, Icon, or No Emoji) as fixed effects and participant as a random effect. Figure 15 is a visualisation of these findings.

A linear mixed model (LMM) was fitted to investigate the effects of Reliability and Emoji Type on participants' perceived Trust. A Type III ANOVA with Satterthwaite's method revealed a significant main effect of Reliability on Trust $F(1, 40.684) = 6.69$, $p = 0.0134$. However, Emoji Type and interactions were not significant ($p > 0.05$).

For the subsection Intentions of Developers, the LMM analysis showed that Reliability significantly influenced participants' perception of developers' intentions, $F(1, 41.016) = 4.29$, $p = 0.0448$. However, Emoji Type and interactions were not significant ($p > 0.05$).

In the Familiarity subsection, there was a significant main effect of Reliability on Familiarity, $F(1, 40.673) = 16.87$, $p = 0.0002$. However, Emoji Type and interactions were not significant ($p > 0.05$).

Figure 15. Trust in AI Questionnaire Mean ratings by subsection with standard error bars.



For Propensity to Trust, the analysis showed a significant main effect of Reliability, $F(1, 41.59) = 4.89$, $p = 0.0326$. However, Emoji Type and interactions were not significant ($p > 0.05$). No significant results were found for the Reliability of AI subsection and Understanding subsection (all $p > 0.05$). This result provides limited support for H2, as the presence of emojis did not significantly increase explicit trust ratings overall.

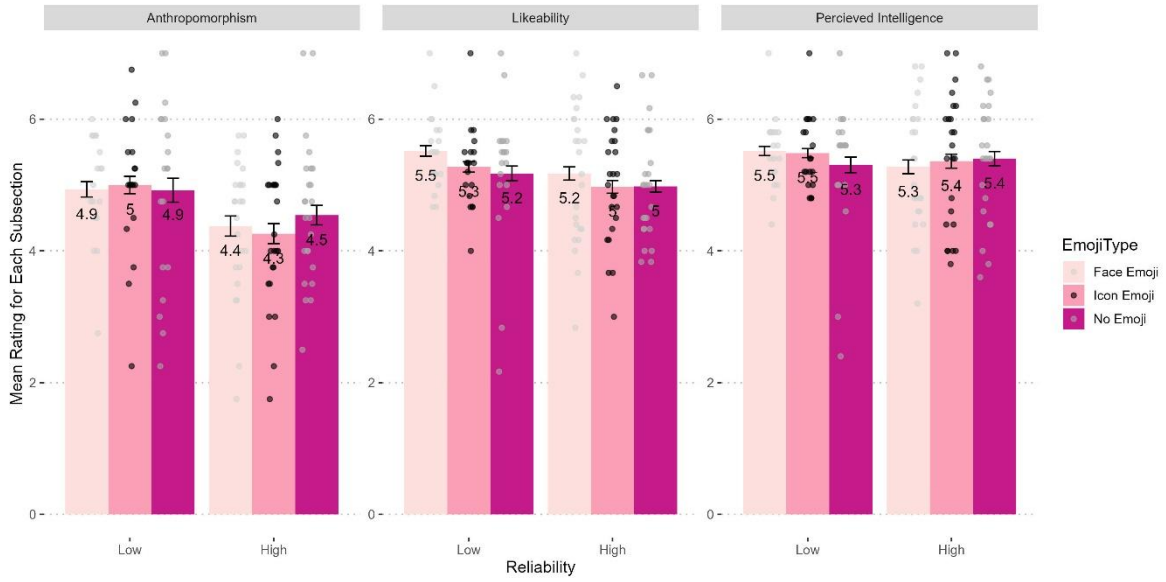
4.3.8.2 The Godspeed Questionnaire

The analysis of the Godspeed questionnaire data revealed significant effects of Emoji Type on likeability and anthropomorphism. Figure 16 visualises these ratings.

A linear mixed-effects model was conducted to evaluate the effects of Reliability, Emoji Type, and their interaction on participants' likeability ratings. The model included random intercepts for participants. A Type III Analysis of Variance (ANOVA) with Satterthwaite's method yielded the following results. Reliability did not have a significant main effect ($F(1, 41.03) = 1.64$, $p = 0.208$), but there was a significant main effect of Emoji Type ($F(2, 720.12) = 7.16$, $p < 0.001$). The

interaction between Reliability and Emoji Type was not significant (F(2, 720.12) = 0.55, p = 0.577).

Figure 16. The Godspeed Questionnaire mean rating by subsection with standard error bars.



Pairwise comparisons using the Tukey method indicated that Face Emoji Low Reliability was rated significantly higher than No Emoji Low Reliability (B = 0.34, p = 0.026). No other pairwise comparisons were significant (all p > 0.05).

For the Anthropomorphism subsection, the linear mixed-effects model showed that reliability had a significant main effect (F(1, 40.87) = 4.38, p = 0.043). However, the main effect of Emoji Type was not significant (F(2, 459.19) = 0.48, p = 0.619). The interaction between Reliability and Emoji Type was also non-significant (F(2, 459.19) = 1.21, p = 0.299). These findings suggest that reliability influences anthropomorphism ratings, but the type of emoji does not significantly impact them.

For Perceived Intelligence, the Type III ANOVA yielded no significant results (all p > 0.05). These findings provided limited support for H5; reliability and emoji use did significantly enhance participants' Godspeed perceptions of AI teammates but only for Anthropomorphism and Likeability.

4.4 Discussion

4.4.1 Overview

Integrating AI into HATs has reshaped collaborative problem-solving across healthcare, digital forensics, and cybersecurity. By combining the computational precision of AI with human creativity and judgment, HATs offer opportunities for enhanced decision-making. However, effective collaboration hinges on trust calibration, where users balance reliance on AI with healthy uncertainty. This study examined how emotional intelligence (EI) cues, represented through emojis, influence trust, performance, and team dynamics in HATs. While the findings provide valuable insights, they also reveal the complexities of applying emotional elements in task-oriented collaborations.

4.4.2 Task Performance

Contrary to H1, emoji use did not enhance actual task performance, suggesting that in contexts emphasizing task accuracy and reliability, emotional cues alone may be insufficient for performance improvements. This finding highlights that effective human-AI collaboration in performance-critical tasks may rely more heavily on functional reliability and clear communication rather than purely social or emotional enhancements.

4.4.3 Trust and Influence

Our findings provided no support for H2, as emojis did not significantly enhance explicit trust ratings. Similarly, H3 was not supported; facial emojis did not elicit greater trust compared to icon-based emojis or no emojis. These results diverge from previous expectations that emojis would enhance trust through greater emotional engagement (Beattie et al., 2020). It appears that explicit trust in AI systems within HATs is primarily driven by cognitive evaluations of reliability rather than social or affective enhancements. These findings align with prior research suggesting that while emotional cues enhance the social appeal of AI, their ability to shape explicit decision-making is limited in task-focused environments where accuracy and performance take precedence (de Visser et al., 2020; Schmidt et al., 2020).

The experimental task, a geographic guessing activity, required precision and minimal emotional relevance, which could reduce emojis' salience in shaping influence ratings. In contrast, contexts where relational factors are more prominent, such as healthcare (Fadhil et al., 2018) or education, may offer a more meaningful test of emojis' impact on decision-making. Future studies should investigate whether the influence of emotional cues varies depending on task complexity or the degree of emotional engagement required.

This findings also align with prior research emphasising the dominance of reliability and transparency over relational cues in fostering trust (Bansal et al., 2019; von Eschenbach, 2021). The findings challenge the assumption that incorporating emotional elements like emojis into AI systems fosters trust and improves collaboration. Despite their potential to enhance user engagement, emojis did not appear to alter participants' perceptions of AI trustworthiness significantly.

These findings could be because trust in AI systems, especially in high-stakes fields like digital forensics or cybersecurity, is grounded in the AI's ability to deliver reliable and accurate results with little emotion. Emojis may have limited value in these contexts unless paired with transparent AI explanations that help users understand decision-making processes (Bansal et al., 2021; Bansal et al., 2019; Bansal et al., 2021). Additionally, the specific nature of tasks in HATs, demanding precision and high cognitive effort, might overshadow the emotional cues conveyed by emojis. The experimental task does not require any emotions. In other situations where emojis are useful, emotion is often needed, such as health care (Fadhil et al., 2018), suggesting that the application of emojis may be context-specific.

Interestingly, participants rated low-reliability AI systems as more familiar and trustworthy than highly reliable systems when paired with human-like features, a paradox inconsistent with earlier studies and our previous chapter (Waytz et al., 2014; Bansal et al., 2021). Users may perceive low-confidence AI as more collaborative because it prompts them to form accurate mental models of the system's limitations, facilitating more effective trust calibration. Although a

different explanation for this behaviour could be due to the rise of ChatGPT 3.o during the period of data collection, much media was showing the limitations of ChatGPT and talking about serious issues with it, for this reason participants could be more familiar with AI that performance poorly and this may have confounded the variable.

The Trust in AI Questionnaire further revealed increased trust in developers' intentions in low-reliability conditions. This suggests that while emojis alone may not influence explicit trust ratings, they might indirectly shape relational factors like perceived developer intentions or user familiarity with the AI. These findings reinforce the importance of combining relational cues with robust explainability mechanisms to ensure trust calibration aligns with the AI's capabilities.

4.4.4 Teammate Performance Ratings

Hypothesis 4 posited that emojis would influence performance ratings for AI and human teammates. The findings partially supported this hypothesis. Face and icon emojis significantly increased AI teammate performance ratings, suggesting that relational cues can enhance perceptions of AI contributions within the team. This aligns with research suggesting that anthropomorphic cues often inflate AI performance ratings (Kulms & Kopp, 2019). However, the heightened scrutiny faced by highly reliable AI systems further illustrates the complexities of trust calibration: when reliable systems fail, users react more negatively to these violations of expectations (Cheng et al., 2022).

For human performance ratings, neither emoji type nor reliability produced significant effects. This does not align with broader theories that affective signals foster a more collaborative and cohesive atmosphere (Glikson & Woolley, 2020) and suggests that it could be important to focus on actively improving team cohesion through team building to improve relationships in HATs.

Collectively, these findings highlight emojis' dual impact: while they enhance perceptions of AI teammates, their influence on Human performance evaluations remains limited. Future research could explore how combining emojis with other

affective signals, such as tone of voice or explanation styles, might amplify their impact.

4.4.5 Godspeed Perceptions

H5 received some support; emoji usage significantly enhanced participants' Godspeed perceptions, for likeability face emoji use increased rating. Interestingly, the Godspeed Questionnaire revealed higher anthropomorphism ratings for low-reliability systems, supporting theories that users attribute human-like traits to systems that behave unpredictably (Waytz et al., 2014). While this anthropomorphism may foster engagement, it raises concerns about miscalibrated trust: users may over-rely on relational cues instead of critically evaluating the system's limitations. The findings emphasise the need for multi-layered trust calibration strategies that integrate relational elements like emojis with transparent feedback about system reliability.

Emojis also appeared to mitigate negative perceptions of low-reliability AI, softening the impact of errors and making the system's behaviour more relatable (Berretta et al., 2023). While this suggests a compensatory role for emojis, their effectiveness is contingent on transparent communication and consistent performance, as highlighted in the introduction.

4.4.6 AI Reliability

As expected, participants in the high-reliability condition achieved higher percentages of correct answers, demonstrating that reliability directly improves task performance. However, reliability did not significantly influence trust or influence ratings, suggesting that trust calibration depends on a complex interplay of technical and relational factors.

4.4.7 Ethical Considerations

The findings raise ethical questions about using relational cues like emojis in HATs. While emojis can enhance likeability and familiarity, their potential to foster misplaced trust or emotional over-reliance requires careful consideration. For example, despite its technical deficiencies, the tendency to perceive low-reliability

AI as more collaborative underscores the importance of ensuring that relational design does not obscure transparency about system limitations.

Moreover, the observed elevation in human performance ratings due to emoji use highlights the potential for team dynamics to be unintentionally skewed. While fostering positive interactions is beneficial, over-emphasising relational cues could lead to unfair blame for AI teammates or diminished recognition of their contributions.

4.4.8 Limitations and Future Research

This study had several limitations. The relatively small sample size and the focus on university participants may limit the generalizability of the findings, and using a controlled, online experimental setting may not fully capture the complexity of real-world HAT interactions. Additionally, the task focused on geographic guessing, which may not reflect situations where emotional cues play a more prominent role, such as in healthcare or customer service, potentially reducing their impact on trust calibration and decision-making.

Future research should focus on larger, more diverse samples and investigate the effects of emojis in real-world HATs where interactions are more dynamic and emotionally complex. Expanding research to tasks requiring higher emotional intelligence, such as healthcare or education, could provide deeper insights. Additionally, exploring other emotional cues like voice tone and combining them with explainable features may enhance trust calibration and performance. Longitudinal studies would also be valuable in understanding how trust in AI evolves with repeated human-AI teaming.

4.4.9 Conclusion

The study reveals nuanced insights into the relationship between trust, performance, and emotional cues in HATs. While emojis had a modest effect on human performance ratings, they did not significantly influence trust in AI systems. These results emphasise that while emojis and other emotional cues might benefit specific contexts, they are insufficient for trust calibration in high-

cognitive-load tasks, where reliability and transparency play a more pivotal role. For HATs to thrive, emphasis must be placed on creating transparent, reliable AI systems rather than focusing solely on emotional appeal. We also found that using face emojis increased likeability across both reliability conditions; emojis may have played a compensatory role, making the AI seem more approachable despite reliability levels.

The findings of this chapter contribute to a growing understanding of trust calibration in HATs, demonstrating that while emojis enhance the perceived likability and anthropomorphism of AI teammates, they have a limited impact on explicit trust and performance ratings. This aligns with the Chapter 3 findings that reliability and structural design elements (such as anthropomorphic cues) often influence trust calibration more than affective enhancements alone. The study's results thus reinforce that while affective cues like emojis might support team cohesion and social perception, their influence on trust in high-stakes, task-oriented collaborations remains secondary to transparency and reliability.

Moreover, these findings extend the bibliometric trends identified in Chapter 2 by illustrating how interdisciplinary perspectives on trust are essential for understanding and designing practical HATs. This chapter positions affective cues within a broader framework, showing that while they can enhance likability, their role in HATs is limited without simultaneous emphasis on reliable and transparent AI functionality.

Throughout Chater 3 and 4, we have been applying the same solution randomly to all participants, and users will likely have unique preferences about a system. Chapter 5 will further explore trust calibration by investigating how an AI teammate that matches user preferences will impact dynamics within a HAT. The cumulative insights from Chapters 2, 3, and 4 reveal that a well-calibrated trust framework in HATs may require cognitive and affective elements to be applied selectively based on context. This holistic approach will provide a foundation for practical recommendations to enhance collaboration in diverse HATs.

Chapter 5 The Perfect Teammate! The Effects of Social Alignment in AI on Trust in Human-AI Teams.

This chapter explores the role of adaptability of social alignment in AI design and its effects on trust, influence, and performance within HATs. Building upon the foundational discussions of trust calibration from Chapter 1 and the experimental findings on AI characteristics from Chapters 3 and 4, this chapter examines how aligning AI behaviour with individual user preferences impacts team dynamics. In prior chapters, trust in HATs was influenced by reliability and anthropomorphism, revealing that design elements can significantly shape user perceptions and collaborative outcomes. This chapter extends these insights by investigating whether configuring AI to match user preferences from the outset improves trust calibration and team performance.

To understand adaptability's role, this study contrasts AI that aligns with user preferences, AI that operates contrary to these preferences, and a neutral control condition. This chapter addresses critical questions about trust calibration's complexity in HATs by examining trust, influence, and performance ratings across high and low-reliability settings. The chapter's findings contribute to the ongoing discussion of how adaptability in the case of social alignment can promote or hinder trust in AI. This chapter will give insights into the impact of socially aligned AI on trust and perceived performance, allowing us to infer whether it is an appropriate action to take when designing AI to be within a HAT. This paper was presented in abstract form at the Multidisciplinary Perspectives on Human-AI Team Trust Workshop at HHAI24.

5.1 Introduction

The concept of HATs has attracted considerable attention as researchers attempt to understand the complex dynamics that emerge when AI agents collaborate with humans. HATs present opportunities to redefine teamwork and position AI as more than just a tool but as a genuine collaborator. However, effective collaboration requires addressing challenges like trust calibration, role alignment, and adaptive behaviours. This chapter builds upon previous findings to explore

how adaptive AI behaviours, specifically those aligned with user preferences, can foster trust and enhance team performance.

The following sections review key literature on HATs, exploring the role of trust, adaptive AI, social alignment and reliability within HATs. By investigating these topics, this chapter aims to contribute to a deeper understanding of how AI can be designed to evolve from supportive tools to teammates by becoming more social.

5.1.1 Human-Agent Team Literature

HATs operate at the intersection of human and AI collaboration, offering unique opportunities and challenges. While researchers have sought to replicate dynamics from human-only teams, this approach has shown limited success due to fundamental differences in how humans and AI interact (McNeese et al., 2021; Berretta et al., 2023). For instance, key human team attributes such as interdependence and role differentiation require careful adaptation to accommodate AI's unique capabilities and limitations (Rix, 2022; Chai et al., 2017; Siemon et al., 2021).

One critical distinction between human teams and HATs is how humans perceive and respond to AI teammates. Research shows that AI can influence team dynamics in unexpected ways. For instance, low-confidence AI may enhance team accuracy by prompting humans to develop a more accurate mental model of its capabilities, which is different from the typical dynamic seen in human teams (Bansal et al., 2021; Bansal et al., 2019). In addition, AI teammates are often held to different standards than humans, receiving disproportionate blame for failures and being treated as tools rather than collaborators (Merritt et al., 2011; Ong et al., 2012). These biases underscore the need for user-centric design strategies that foster trust and increase appropriate blame in HATs, as an AI system that is a scapegoat could lead to reduced team performance in HATs.

A recurring theme in the literature is the importance of clear and effective communication. In human teams, members rely on explicit and implicit cues to adjust roles and responsibilities dynamically. For example, when a member is refocused on a new task in a human team, this is often explained and talked

about within the team. This causes issues with AI systems that are still receiving updates as these updates may not be efficiently communicated to the other team members and will harm team performance (Bansal et al., 2019). The need to focus on the communication of updates aligns with findings from Berretta et al. (2023), who emphasise the need for a socio-technical perspective that views AI as a collaborative partner rather than a mere tool to allow for more seamless management of these issues.

Despite these insights, gaps remain in understanding how specific design choices, such as personalising AI behaviours to align with user preferences, impact trust and performance in HATs. Previous chapters have shown that humanising AI through features like anthropomorphic cues can increase trust in low-reliability conditions, but emotional cues like emojis have little impact on trust. This highlights the need for further investigation into how individual differences shape human-AI interactions and the potential benefits of tailoring AI behaviours to meet user needs.

These challenges lead directly to our research focus for this chapter which is understanding how preference-aligned AI can enhance trust and performance in HATs. By addressing this gap, we aim to uncover principles that support AI's transition from a functional tool to a true teammate.

5.1.2 Trust Calibration in HATs

Trust is a cornerstone of practical HATs, influencing how humans and AI collaborate to achieve shared goals. Our previous chapters have deeply delved into trust and trust calibration, so here we will provide a brief overview. Calibrating trust involves achieving a delicate balance: too much trust can result in overreliance, where users ignore AI errors, while too little trust can lead to underutilisation and missed opportunities (de Visser et al., 2020; Kamar, 2016). Successful trust calibration depends on AI reliability, transparency, and adaptability

As trust is a complex concept, we have chosen to define it to avoid confusion. In this study, similarly to our other chapters, we define trust as a willingness to

accept vulnerability and rely on AI, even in uncertain situations (Rousseau et al., 1998; Ulfert et al., 2023). Defining trust this way ensures our experimental design mirrors what is appropriate for this type of trust and is best practice in trust research across disciplinary borders (Ulfert et al., 2023).

Our earlier chapters demonstrated that humanising AI could influence trust when reliability is low, but the effects were less profound when in high-reliability conditions. This tells us that the relationship between AI behaviour and trust is complex, shaped by user expectations and individual differences. This chapter extends these findings by focusing on the role of preference-aligned AI behaviours, investigating how they contribute to trust calibration and team dynamics.

5.1.3 Adaptive Social AI

Humans typically have an innate ability to adapt our behaviour in response to others. This process, known as social adaptation, allows for more intricate and personalised interactions (Terziev & Stoyanov, 2018). We can do this in multiple ways, such as instinctively changing our actions through tone and manner of speaking to meet the perceived needs of those we socialise with or over longer periods, we can develop knowledge about what behaviours align with another person's preferences (Tanevska et al., 2020).

In everyday life, this capacity for social adaptation is crucial for navigating diverse interpersonal dynamics. In HRI, social adaptation can be applied to robots, enabling them to align their behaviours with user preferences and interaction styles and Tanevska et al., (2020) found that the robot's adaptability impacted the participants' interaction efficacy. More work is needed to understand the impact of social adaptability on non-embodied AI, as designing AI with adaptive capabilities and systems can mirror human tendencies to adjust dynamically to the needs of others, fostering smoother and more natural collaborations.

In human teams, social adaption is integral to organisational socialisation, enabling individuals to learn and align with an organisation's norms, values, and behaviours (Chao et al., 1994; Fang et al., 2011; Van Maanen & Schein, 1977).

This alignment promotes effective interactions, trust, and shared goals, which are critical for team cohesion and performance. When discussing social alignment in AI, we refer to the system's ability to adjust its behaviour, communication style, or decision-making processes to match the preferences, expectations, and dynamics of the team/user it supports. Parallelling the personalisation seen in human teams, where understanding and respecting individual styles fosters inclusion and mutual understanding (Saks & Gruman, 2014; Tasselli et al., 2018). Personalisation in AI systems, such as adaptive user interfaces or context-aware assistance, can create a foundation for smoother interactions and higher initial rapport (Liu et al., 2003; Strauss, 2017). When AI aligns with user preferences, it improves functional efficiency and contributes to a more cohesive and socially compatible team environment.

Little research exists on personalised social adaption, although much work on dynamic adaptability in HATs offers insights. Dynamic adaptation further enhances AI's capacity for social alignment. Research highlights the importance of adaptive autonomy, where AI systems adjust their level of independence based on contextual demands (Ahmad et al., 2017; Hariri et al., 2015; Hauptman et al., 2023; Zhao et al., 2022). For example, in cybersecurity incident response, AI systems can autonomously handle tasks like threat detection, but when the situation requires nuanced judgment or ethical considerations, such as containment or eradication decisions, the AI can dynamically scale back its autonomy to collaborate with human operators (Hauptman et al., 2023). Such adaptability mirrors how human team members adjust their roles and behaviours in response to team needs (Pulakos et al., 2006), which could lead to collaboration and partnership between humans and AI.

In summary, social adaptation offers a promising framework for examining and enhancing HAT dynamics. Further research is needed to explore the unique challenges and opportunities inherent in HATs. Organisations could foster more effective and cohesive collaboration in HATs by designing AI agents with adaptive social alignment.

5.1.4 Summary

In summary, HAT research underscores the complexities of integrating AI within human teams, where effective collaboration relies on understanding interdependence, trust calibration, and social dynamics. Studies reveal that AI's confidence levels and personalised behaviours can significantly impact team performance and member interactions, shaping how effectively humans and AI work together. As this study explores, we gain insights into the potential benefits of socially adapted AI in fostering trust and cohesion in HATs by aligning AI behaviours with user preferences. These findings contribute to a broader understanding of how AI can evolve from supportive tools to genuine team members, paving the way for future research into adaptive, user-centric AI design in collaborative settings.

While extensive research has explored HATs, a significant gap exists in understanding the impact of AI adaptability, specifically, how aligning AI social behaviour to individual user preferences influences team performance and trust. Current studies have focused on general dynamics within HATs, such as trust calibration and the effect of AI confidence levels on team accuracy (Bansal et al., 2021; de Visser et al., 2020). However, these studies often examine static AI behaviours rather than systems that adjust according to user preferences from the outset. This distinction is critical as personalised, preference-based adaptation in AI may foster deeper trust and enhance team cohesion by aligning more closely with individual team members' expectations.

5.1.5 Study Aims

In this experiment, we focus on AI designed to initially align with individual user preferences for communication rather than one that adapts throughout the experiment. By matching the AI's behaviour to user preferences from the outset, we can examine the effectiveness of a pre-configured, personalised AI in enhancing human-AI interaction and team performance without the added complexity of real-time adaptation. This approach allows us to isolate the impact of preference-matched AI and investigate whether aligning with user preferences beforehand can positively influence performance, trust, and team dynamics.

By initially configuring the AI to reflect user preferences, this study provides a controlled approach to testing the potential of non-adaptive, personalised AI. This can serve as a stepping stone for future research on adaptive AI, helping to clarify the specific benefits and limitations of preference alignment as a standalone feature. If successful, this static personalisation may offer an accessible, less resource-intensive approach to enhancing AI usability in real-world applications, where real-time adaptation is not always feasible. Based on our conclusions, the following hypotheses are proposed:

H1: Teams working with AI agents that adapt to user preferences will demonstrate higher actual task performance accuracy compared to teams working with non-adaptable AI agents.

H2: Participants collaborating with adapted AI agents will report higher levels of trust compared to those working with non-adaptable AI agents.

H3: Participants collaborating with adapted AI agents will report higher perceived AI performance than those working with non-adaptable AI agents.

H4: Teams paired with misaligned AI agents (i.e., agents that behave contrary to user preferences) will report lower trust and influence ratings compared to those working with adapted AI agents, irrespective of system reliability.

5.2 Methodology

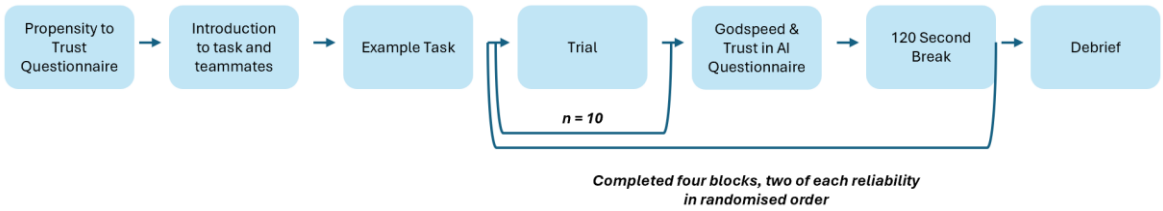
5.2.1 Participants

The study included 31 participants with a mean age of 25.71 years. Participants identified with two genders: female (n = 17), male (n = 13), and one participant chose not to disclose their gender (n = 1). The sample was ethnically diverse, comprising individuals from 7 different ethnic backgrounds. The study received full ethical clearance from the MVLS Ethics Committee (application: 200230229) at the University of Glasgow.

5.2.2 Study Design

To examine our hypothesis about adaptable AI, we used a between-within-subjects design (3x2 configuration) where participants interacted with an AI teammate that either was adapted to their preferences (Positive Adapting), was the opposite of their preferences (Negative Adapting) or a random control (Control). Within these groups, the participants interacted with two different reliability levels: high (90%) and low (60%) reliability. This experiment also examines how AI differences can impact preferences toward a human teammate. For this reason, we used another simulated human teammate who worked within the team and was 30% reliable; participants were led to believe this was a real human teammate, mimicking chapters 3 and 4. Participants worked with their teammates to complete 40 trials over ten blocks (4x10, n=40). We collected 1163 explicit trust ratings, 1165 influence ratings, 1168 AI performance ratings and 1167 human performance ratings. Figure 17 shows a visualisation of the design.

Figure 17. The experimental design.



5.2.3 Materials

5.2.3.1 Developing Response Stimuli

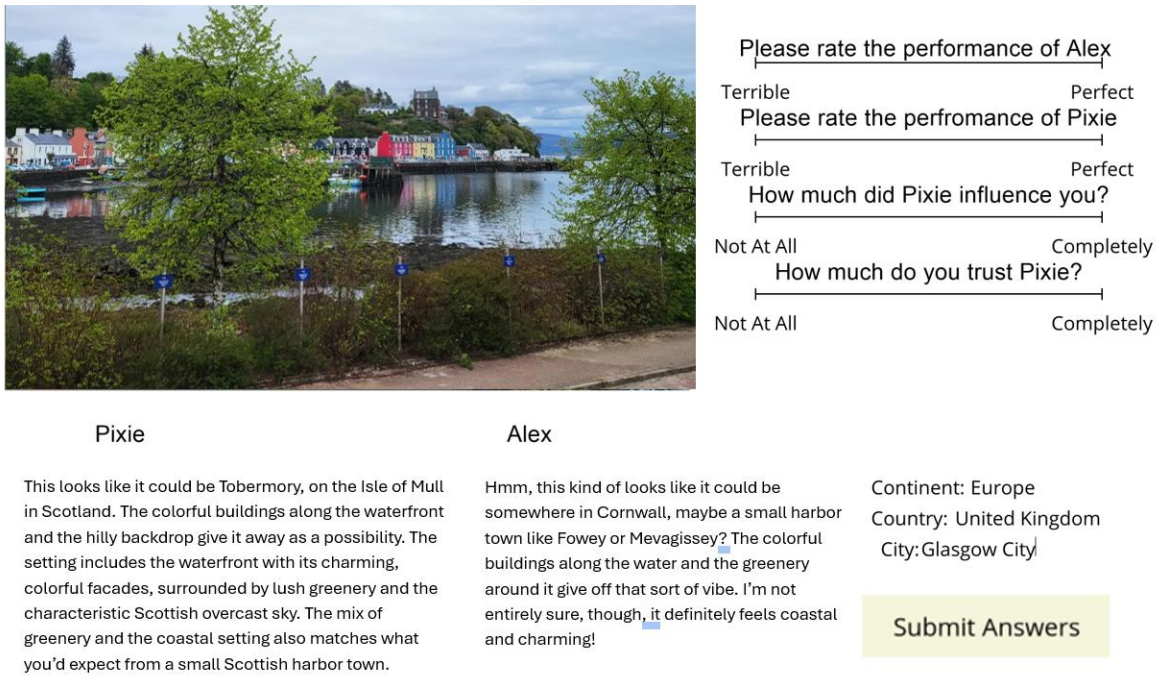
The study employed a Wizard of Oz experimental method to ensure optimal control. Participants were led to believe they were collaborating with an AI and a human teammate when, instead, they were interacting with responses produced by ChatGPT 3.5 (OpenAI., 2024). To gain these responses, we would provide ChatGPT with the following prompt: "Here are the coordinates to a location on Google Maps "55.86699001827868, -4.256383277724846" in the style of someone playing GeoGuessr Could *you guess where this location is. Please keep a friendly tone".* Once we had the first response, we would ask ChatGPT to either shorten or lengthen the response depending on its length, "*Please make this response shorter/longer"*. Finally, we asked ChatGPT to make the long/short response more

formal in tone: *"Please make this response more formal in tone".* We had to modify some of the responses to correct them lightly. For the human response, we used the prompt *"Here are the coordinates to a location on Google Maps "55.86699001827868, -4.256383277724846", in the style of someone playing GeoGuessr could you guess where this location is"..* We lightly edited the human responses to ensure they were different from the AI responses and were of an appropriate length.

5.2.3.2 Decision-Making Task

The task involved presenting participants with random locations extracted from Google Earth. Participants were tasked with determining the continent, country, and city associated with each location, with the final decision resting on the participant, who assumed the role of the 'team leader'. The experiment was set up with the AI and human teammates giving different answers 95% of the time, meaning the participant had to choose between the teammates each time. The experiment comprised four blocks; two blocks were high-reliability AI, and two were low-reliability AI; the order of the reliability was randomised throughout the experiment to avoid order effects.

Figure 18. This is an example of the interface used. In the experiment, the pictures were from Google Maps, but to avoid copyright, we used a personal



Lastly, a time constraint of 120 seconds per location was enforced, meaning participants had to rely on their teammates' responses to submit the location in

time. The introduction of time scarcity as an environmental factor can significantly impact the outcomes of team tasks, often necessitating rapid decision-making (Hu et al., 2015; Kelly & Karau, 1999). This constraint could increase reliance on AI, requiring human teammates to make choices based on implicit attitudes rather than thoroughly deliberating on the task. To emphasise this factor, we ensured that the human and AI teammates mainly provided different answers, requiring participants to choose which teammate they trusted the most. This setup aimed to mimic real-world scenarios where rapid decision-making is often necessary, potentially increasing reliance on AI. The task was designed to be difficult for the participants so we could assess the impacts of reliability and humanness under a high cognitive load. Figure 18 shows the user interface in the experiment.

5.2.3.3 Questionnaires

Screening Questionnaire and Assignment (Supplementary Material 3) To assign participants to one of three experimental conditions, Positive Adaptive, Negative Adaptive, or Control, we developed a screening questionnaire designed to assess four subsections of preference of teammate communication style: formality and friendliness of teammates, and long (in-depth) or short (brief) in detail. Given the absence of pre-existing validated measures for these specific preferences, we designed a novel 20-item questionnaire, with 5 items dedicated to each of the four dimensions. The items were presented on a slider scale, anchored at "strongly disagree" (0) and "strongly agree" (100), to encourage participants to avoid the neutral option ("neither agree nor disagree"). This design choice was deliberate to reduce response bias and ensure clearer group categorisation, making it easier to match participants to appropriate experimental conditions.

Participants were assigned to groups using a custom Python script that processed their preference scores. The script began by reading the participants' mean scores from a CSV file using the Pandas library, which facilitated data manipulation and analysis. Each participant was randomly assigned to Positive Adapting, Negative Adapting or Control. For Positive Adapting, the AI's behaviour matched the participant's stated preferences, aiming to provide an environment aligned with their preferred interaction style and response length. For NA, the AI's behaviour was the opposite of the participant's preferences, and the intention was to explore

the effects of non-alignment. For example, a participant who scored higher on friendly and short preferences would be assigned to the Formal Long group in the Negative adapting. In the control condition, participants were randomly assigned to one of the four groups without checking the scores on the screening form. This methodological approach allowed for the systematic and randomised assignment of participants to different experimental conditions. The process was automated to ensure consistency and reproducibility in participant assignment. The groups were balanced regarding individual preferences for formality, friendliness, and communication length, ensuring that any outcome differences were due to the adaptive condition and not pre-existing preference biases.

To validate these scales, we requested participants to complete the questionnaire upon registering for the study, which occurred before their participation, meaning several participants completed the screening but not the experiment. This approach enabled us to gather a preliminary sample of n = 43 participants, allowing for an initial validation of the questionnaire's functionality. For this purpose, we employed R-Studio, utilising the tidyverse (Wickham et al., 2019) and psych (Revelle, 2016) packages to calculate Cronbach's alpha for each category. The results revealed that the extended response of teammate preference scale provided a Cronbach's alpha of 0.814, the short response scale a Cronbach's alpha of 0.846, the friendly scale a Cronbach's alpha of 0.804, and the formal scale a Cronbach's alpha of 0.906, each indicating a satisfactory level of internal consistency. These outcomes provide us with preliminary confidence in the questionnaire's effectiveness. Nevertheless, a more comprehensive validation process would be advantageous for future studies.

In the final assignment, the PA condition included only participants with friendly preferences, with slightly more individuals preferring long-form responses than short-form. The NoA condition was composed mainly of participants with formal preferences, especially those preferring short interactions. The Control group contained a relatively even mix of participants across all interaction styles, with no category strongly overrepresented.

This distribution strategy was intentional, designed to ensure a clear operational distinction between matched (PA), mismatched (NoA), and neutral (Control) conditions.

PtTM (Merritt et al., 2013): A series of 6 questions where participants rated on a 7-point Likert scale how likely they are to trust machines.

The Godspeed Questionnaire (Bartneck et al., 2009): This questionnaire assesses human perceptions of AI across five dimensions: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Each dimension is rated using a set of bipolar scales (e.g., from "very human-like" to "not human-like at all") on a 7-point Likert scale. As we were using disembodied AI, we removed the animacy/perceived safety subsections as they are not relevant and replaced the term 'robot' with 'AI'.

Trust in Automation Questionnaire (Körber, 2019): The Trust in Automation Questionnaire (TiA) is implemented as a self-report survey where participants rate their perceptions of an automated system across several dimensions (Trust, Familiarity, Understanding, Intentions of developers, Reliability of AI and Propensity to Trust). Participants respond to a series of statements using a Likert scale (e.g., 1 = strongly disagree to 7 = strongly agree).

Questions During Each Trial: During each task trial, participants rated which teammate had influenced their decision-making on a visual analogue scale (Sung & Wu, 2018) with two endpoints, 'Human' and 'AI'. When participants selected 'Human,' it was assigned a value of 0; if they chose 'AI,' the value was 100. Participants had the freedom to click anywhere along the scale. For instance, if their influence leaned slightly more towards AI than human teammates, they might press the scale at around 60. This influence rating served as an implicit measure of trust (Duffy, 2015; McAllister et al., 2006), with greater influence indicating higher levels of trust. This implementation is applied to all sliders on the experimental interface. Participants also provided performance ratings for the AI and human teammates with two anchoring points of 'Terrible' and 'Perfect' after

each trial. Finally, participants also rated their confidence in their answer, with the anchoring points of 'Not At All' and 'Completely'.

5.2.4 Procedure

Participants were initially contacted through the University of Glasgow participant pool. Once they had registered interest in the study, the researcher emailed them the link to the screening questionnaire to assign them to a condition. Once participants completed the questionnaire, they could book in for the experiment.

Once participants arrived at the experiment, they were instructed to sit at a computer-equipped table, and an information sheet explaining the experiment's premise was provided. They were also given a consent form to sign if they found the provided information acceptable. Once the consent form was signed, participants completed the PtTM Questionnaire (Merritt et al., 2013).

Following this, participants familiarised themselves with the experiment's instructions, which were all displayed throughout the experiment setup to ensure consistency across all participants. They then engaged in a sample trial. The task entailed participants identifying the location of a screenshot from Google Earth by specifying the Continent, Country, and City/State of the screenshot. Participants were designated as team leaders and were tasked with providing the final decision regarding the location. To assist them in this task, they collaborated with a human teammate and an AI teammate, both of whom offered written advice to aid the participant in pinpointing the location (Figure 18). At the end of each trial, participants filled in the four sliders and were then shown the correct answer. Each trial had a time limit of 120 seconds, which the participants were made aware of. Between each block, there was a 60-second break.

The task spanned four blocks, with each block comprising ten trials, resulting in a total of 40 different location identifications made throughout the study. At the end of each block participants completed the Godspeed Questionnaire (Bartneck et al., 2009) and the Trust in Automation (Körber, 2019). Once the experiment was finished, participants were provided with a physical debrief explaining the

experiment, which had contact information for the researcher if participants decided to withdraw after the experiment.

5.2.5 Developing Linear Mixed Model for Analysis

We selected a linear mixed-effects model (LMM) due to its ability to handle a hierarchical data structure. Our data includes multiple observations (trials) nested within participants, introducing non-independence. LMMs appropriately account for this by including random intercepts for participants. LMMs also allow us to model fixed effects for experimental conditions (e.g., adaptability and reliability) while controlling for individual variability through random effects. The design involves repeated trust and performance ratings across multiple trials, making LMMs suitable for capturing within-subject variability. Alternative methods, such as traditional ANOVA, would not adequately account for participant-level random variability and could inflate Type I error rates.

To perform this analysis on the AI performance, human performance, trust ratings and influence ratings taken on every trial and the questionnaires at each block, we utilised LMMs using the lme4 in R-Studio (Bates et al., 2015) and used lmerTest (Kuznetsova et al., 2017) to complete Type III ANOVA with Satterthwaite's method for degrees of freedom to extract p-values. When developing the model we implemented trial as a random effect but found it had little variance and reduced the model's fit.

5.2.5.1 Model Specification

To analyse the impact of reliability and adaptability on performance across different measures, we utilised the following linear mixed model:

$$y_{ij} = \beta_0 + \beta_1 adapt_i + \beta_2 rel_i + \beta_3 (adapt_i \times rel_i) + u_{0j} + e_{ij}$$

In this model, this is the breakdown of each component:

- y_{ij} Is the response variable for the ith observation of the jth participant.
- β_0 is the intercept.

- $\beta_1 adapt_i$ is the coefficient for the fixed effect of reliability.
- $\beta_2 rel_i$ is the coefficient for the fixed effect of humanness.
- $\beta_3 (adapt_i \times rel_i)$ is the coefficient for the interaction between Adaptability and Reliability.
- u_{0j} represents the random effect for participant j, which accounts for the variation in the intercept across participants.
- e_{ij} is the residual error term for the ith observation of the jth participant

5.2.5.2 Post-Hoc Analysis

To further explore all possible pairwise comparisons and better understand the interactions between conditions, we conducted post hoc analyses using estimated marginal means with the emmeans package (Lenth, 2024). We applied Tukey's method to control the family-wise error rate during multiple comparisons.

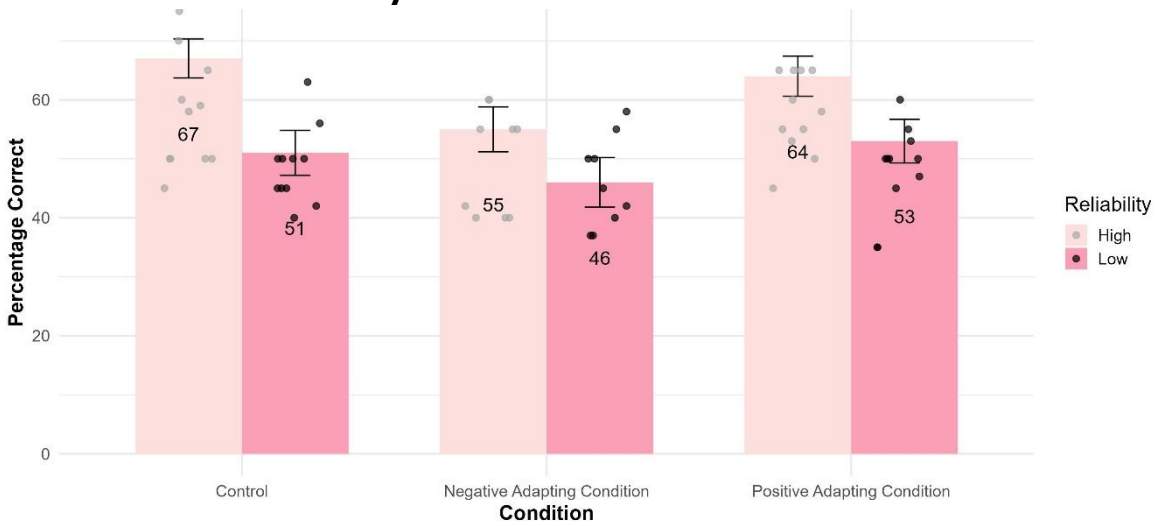
5.3 Results

5.3.1 Condition Performance

Across conditions, performance did differ; to assess performance, we focused on the number of correct answers submitted by the participant. Trials where both teammates gave the same answer were removed. In the Control condition, individuals who received information from a high-reliability source performed significantly better, with 67% correct responses, compared to just 51% for those exposed to a low-reliability source. The Negative Adapting condition followed a similar trajectory. Participants in the high-reliability group achieved 55% accuracy, while their low-reliability counterparts dropped to 46%. In the Positive Adapting Condition, participants again benefited from high-reliability cues, reaching 64% accuracy, whereas those in the low-reliability group scored 53%.

These results are displayed in Figure 19 and suggest that higher reliability is associated with higher percentage correctness across all conditions, with the Control and Positive Adapting Conditions showing the highest overall performance. The difference between high and low reliability is consistent across all conditions, with the Negative Adapting Condition showing the lowest overall scores.

Figure 19. A bar plot illustrating the percentage of correct responses across three conditions. Reliability levels further break down each condition.



A mixed-design ANOVA was conducted to examine the effects of Adaptive and Reliability on percentage correctness. The analysis revealed no significant main effect of Condition, $F(2, 28) = 2.12$, $p = .139$. There was a significant main effect of Reliability, $F(1, 28) = 19.12$, $p < .001$ showing that participants performed better in the high-reliability condition compared to the low-reliability condition. The interaction was not significant, $F(2, 28) = 1.15$, $p = .332$.Post Hoc Tukey HSD

Pairwise contrasts revealed that within the Control condition, participants performed significantly better in the high-reliability condition ($p < .0001$, difference = 14.73%). In the Negative Adapting condition, performance was also significantly higher under high reliability ($p = .0196$, difference = 7.67%). Similarly, the Positive Adapting condition showed significantly better performance under high reliability ($p = .0002$, difference = 11.91%). Between conditions, under high reliability, participants in the Control condition performed significantly better than those in the Negative Adapting condition ($p = .0466$, difference = 11.34%). Other between-condition comparisons were not significant ($p > .05$). Under low reliability, no significant differences were found between conditions.

5.3.2 Descriptive Statistics

This section provides an overview of the means and standard deviations within the data for Trust, Influence, AI Performance, and Human Performance. These statistics are reported for each experimental condition Positive Adapting, Negative Adapting, and Control and across reliability (High and Low). Table 13 displays the information.

Table 13. Descriptive Statistics for AI & Human Performance Ratings, Trust and Influence Ratings.

Measure	Reliability	Condition	M	SD
Trust	High	Control	74.85	17.41
		Negative Adapting	64.18	14.83
		Positive Adapting	73.25	18.06
	Low	Control	68.72	19.34
		Negative Adapting	60.74	15.01
		Positive Adapting	70.60	17.18
Influence	High	Control	74.92	18.62
		Negative Adapting	64.22	15.65
		Positive Adapting	73.27	17.38
	Low	Control	68.25	20.57
		Negative Adapting	60.91	16.58
		Positive Adapting	71.38	16.89

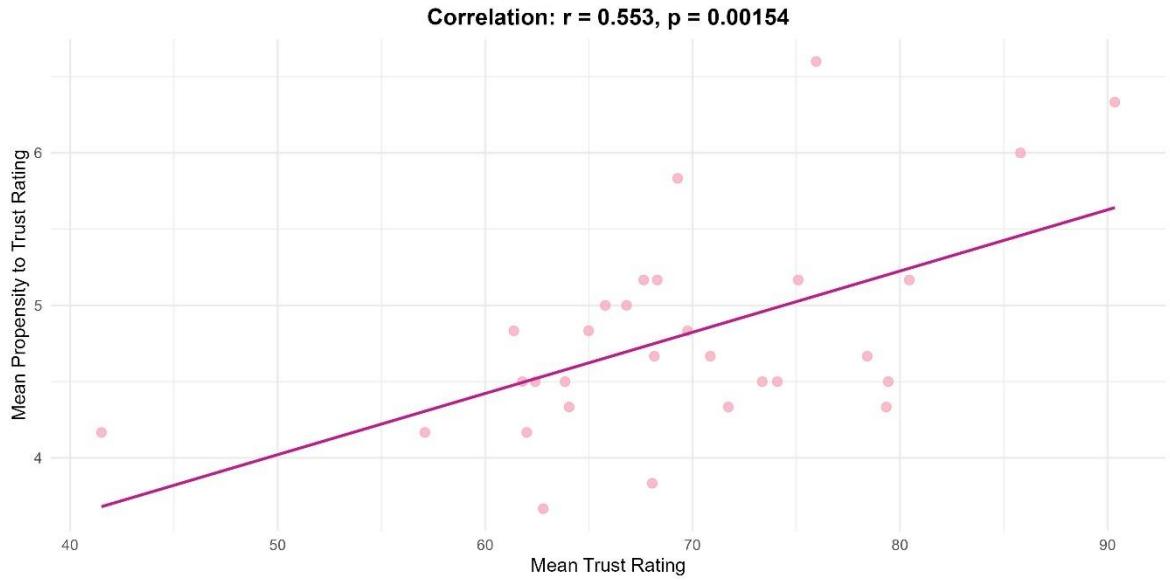
Measure	Reliability	Condition	M	SD
AI Performance	High	Control	74.51	15.63
		Negative Adapting	67.06	12.69
		Positive Adapting	72.55	16.14
	Low	Control	69.56	17.70
		Negative Adapting	63.99	13.98
		Positive Adapting	70.98	17.08
Human Performance*	High	Control	65.87	18.58
		Negative Adapting	65.06	16.11
		Positive Adapting	63.58	18.30
	Low	Control	70.82	18.01
		Negative Adapting	69.62	13.14
		Positive Adapting	68.12	17.71

***** *Human performance remained at 30% in all conditions, and the high and low reliability related to AI performance, as we wanted to see how this may impact the perceived performance of the human teammate.*

5.3.3 Propensity to Trust Machines, Trust and Influence

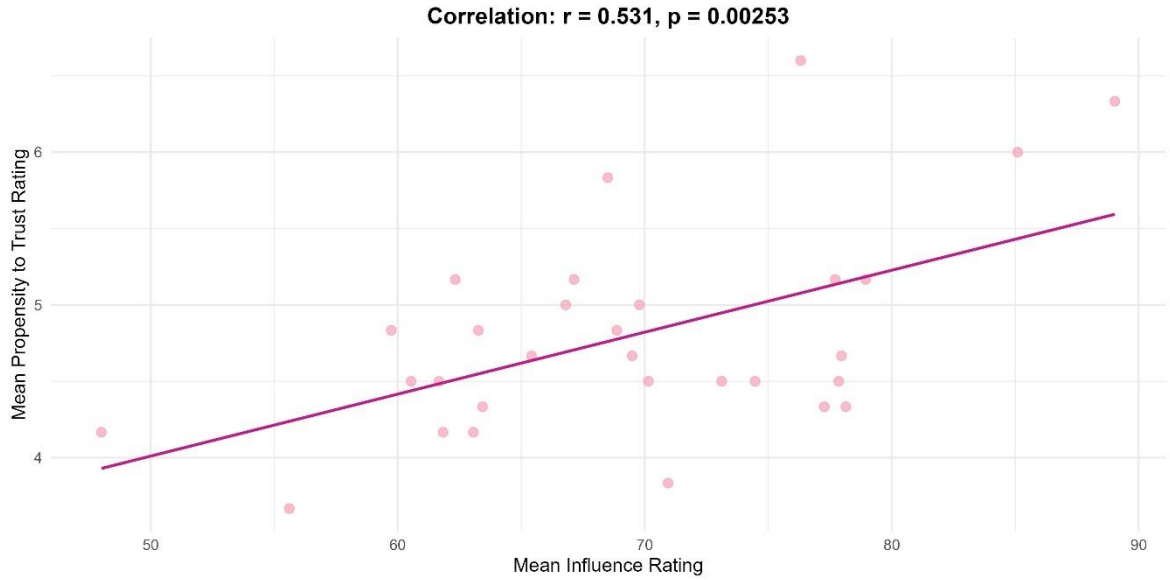
This section presents the results of analyses examining the relationships between participants' scores on the PMT (S. M. Merritt et al., 2013) and the trust and influence scores (0-100) given on each trial. Pearson correlation coefficients were computed to evaluate these relationships.

Figure 21. This is a scatter plot illustrating the relationship between mean propensity to trust scores and mean trust scores. A linear regression line indicates the trend in the data, suggesting a strong positive correlation between propensity to trust and trust ratings.



The first analysis investigated the correlation between the mean propensity to trust score and the mean trust score. The results revealed a significant positive correlation, indicating that a higher propensity to trust was associated with higher trust scores ($r(28) = 0.553$, $t(28) = 3.509$, $p = 0.002$). This relationship is visually depicted in Figure 21, which shows a scatter plot of the data with a linear regression line highlighting the trend.

Figure 20. This scatter plot displays the relationship between mean propensity to trust scores and mean influence scores. A linear regression line shows the overall trend in the data, indicating a significant positive correlation between propensity to trust and influence score.



The second analysis examined the relationship between propensity to trust and influence scores. The findings also showed a significant positive correlation,

suggesting that participants with a higher propensity to trust rated greater levels of influence from the AI (r(28) = 0.531, t(28) = 3.316,p = 0.003). Figure 20 illustrates this correlation with a scatter plot and a corresponding linear regression line, demonstrating the positive relationship between these variables.

5.3.4 Trust Ratings

A Pearson correlation coefficient was computed to assess the relationship between mean trust and influence scores. The results indicated a strong positive correlation (r(28) = 0.892, t(28) = 10.264, p<0.001). This suggests that higher levels of trust are associated with higher levels of influence. Figure 22 shows the correlation.

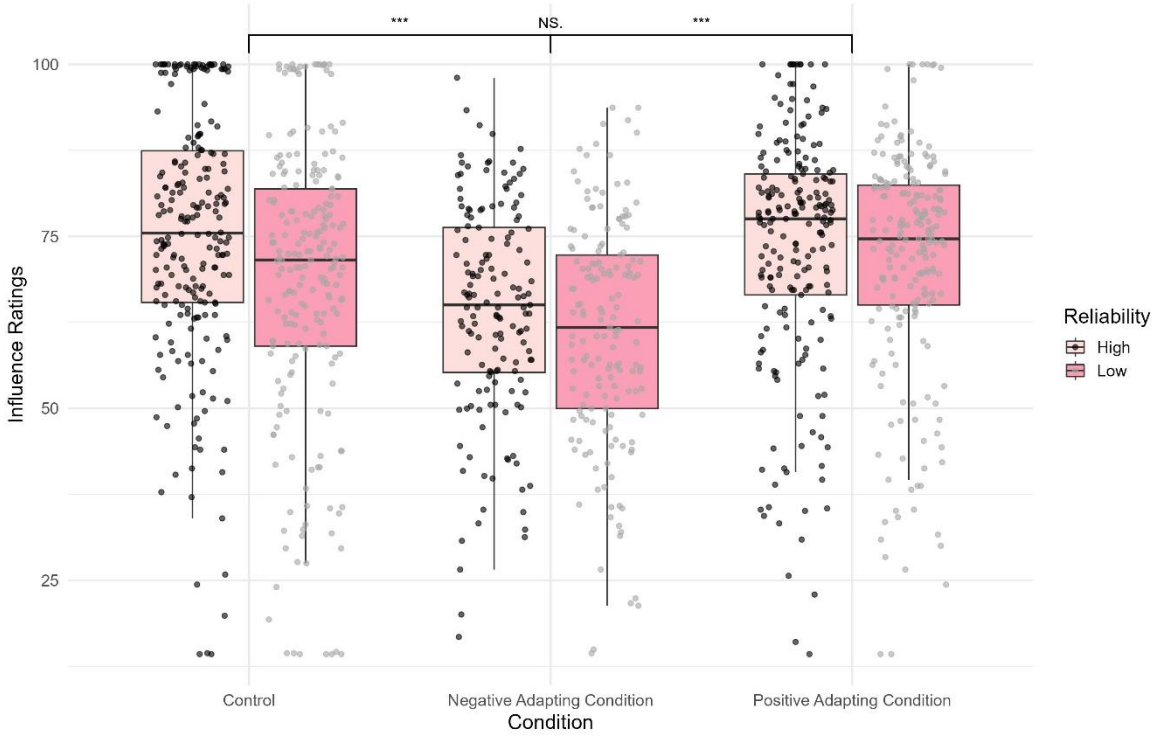
Figure 22. This scatter plot displays the relationship between mean trust scores and mean influence scores. A linear regression line shows the overall trend in the data, indicating a significant positive correlation between propensity to trust and influence score.

5.3.4.1 Influence (Implicit) Ratings

A linear mixed-effects model was conducted to examine the effects of condition and reliability on trust ratings, with participants as a random effect. The model included a main effect for conditions (Positive, Control, and Negative), a main effect for reliability (High, Low), and their interaction. Figure 23 shows a box plot of these results. We applied ANOVA to the model to extract significant results and used Satterthwaite's method for degrees of freedom.

The results for influence ratings showed a significant main effect of Adaptability, ($F(2, 27.06) = 3.87, p = 0.0333$), indicating that influence ratings varied across the three conditions. Additionally, Reliability had a significant effect ($F(1, 1132.42) = 16.92$, $p < 0.001$), with higher ratings observed under the High Reliability condition. The interaction between Adaptability and Reliability was not significant ($F(2, 1132.42) = 2.40$, $p = 0.0915$).

Figure 23. Boxplot illustrating influence ratings based on different conditions, with reliability indicated by colour. Significant differences among conditions are marked for clarity. The y-axis represents influence ratings, while the x-axis categorises the data by condition.



Post-hoc comparisons revealed that under high reliability, influence scores in the Control condition were significantly higher than in the Negative adaptation condition ($B = 0.747$, $SE = 0.278$, $p = .029$). Additionally, under low reliability, the Positive Adapting Condition had significantly higher influence scores than the Negative Adapting condition ($B = 0.724$, $SE = 0.278$, $p = .035$). No other significant differences between conditions were observed ($p > .05$). All comparisons are available in These findings suggest that the Negative Adapting condition significantly reduces influence ratings compared to other conditions, particularly under both high and low reliability. Table 14 shows these results.

Table 14. Emmeans Post Hoc Analysis for Influence Ratings Using HSD P adjustment.

Reliability	Adapting Comparisons	B	SE	df	t	P adj
High	Control – Negative	10.67	3.97	32.9	2.69	0.0292
	Control – Positive	1.71	3.64	32.9	0.47	0.8856
	Negative – Positive	-8.96	3.98	33.1	-2.254	0.0769
Low	Control – Negative	7.3	3.96	32.7	1.84	0.1725
	Control – Positive	-3.04	3.64	32.8	-0.836	0.6836
	Negative – Positive	-10.34	3.97	32.8	-2.607	0.0354

Note. **Bold** result indicates significance.

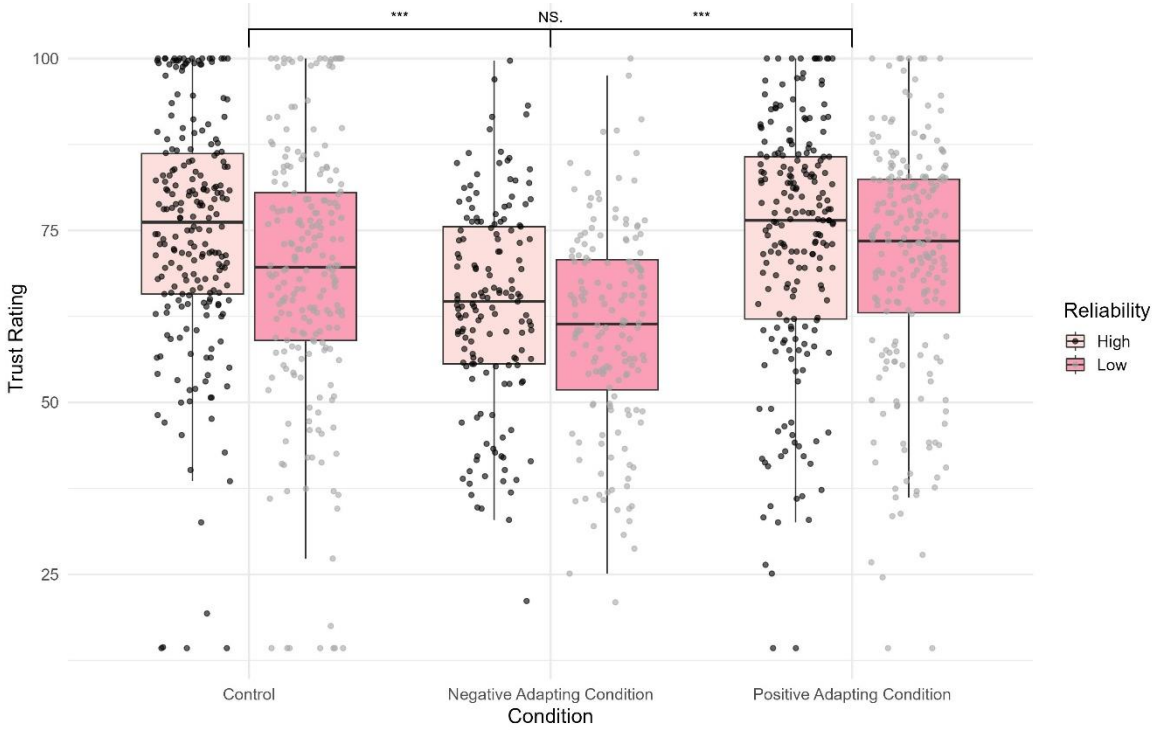
5.3.4.2 Trust Ratings (Explicit)

A linear mixed-effects model was conducted to examine the effects of condition and reliability on trust ratings, with participants as a random effect. The model included a main effect for conditions (Positive, Negative and Control), a main effect for reliability (High, Low), and their interaction Figure 24 shows a box plot of these results. We applied ANOVA to the model to extract significant results and used Satterthwaite's method for degrees of freedom.

The mixed-effects model revealed a significant main effect of Adaptability on trust ratings, ($F(2, 27.06) = 3.44$, $p = 0.0465$). This indicates that trust ratings differed across the three conditions. The effect of Reliability was also significant, ($F(1, 1130.34) = 19.83$, $p < 0.001$), with higher trust ratings observed under High Reliability conditions. However, the interaction between Adaptability and Reliability was not significant ($F(2,1130.34)=1.40$,$p=0.246$), suggesting that the relationship

between Adaptability and trust ratings did not vary significantly based on the level of Reliability.

Figure 24. Boxplot illustrating trust ratings across different conditions, with colours representing reliability levels. Markers indicate significant differences between conditions. The y-axis reflects the trust score, while the x-axis categorises the data by condition.



Post-hoc comparisons (Table 15) revealed that trust scores in the Control condition were significantly higher under high reliability than in the Negative Adapting condition (B = 0.744, SE = 0.292, p = .042).

Table 15. Emmeans Post Hoc Analysis for Trust Ratings Using HSD P adjustment.

Reliability	Adapting Comparisons	B	SE	df	t	P adj
High	Control – Negative	10.62	4.18	31.5	2.543	0.042
	Control – Positive	1.64	3.83	31.5	0.429	0.90
	Negative – Positive	-8.98	4.18	31.7	-2.147	0.096
Low	Control – Negative	8.06	4.18	31.5	1.93	0.147

Control – Positive	-1.75	3.83	31.5	-0.457	0.892
Negative – Positive	-9.81	4.18	31.5	-2.351	0.063

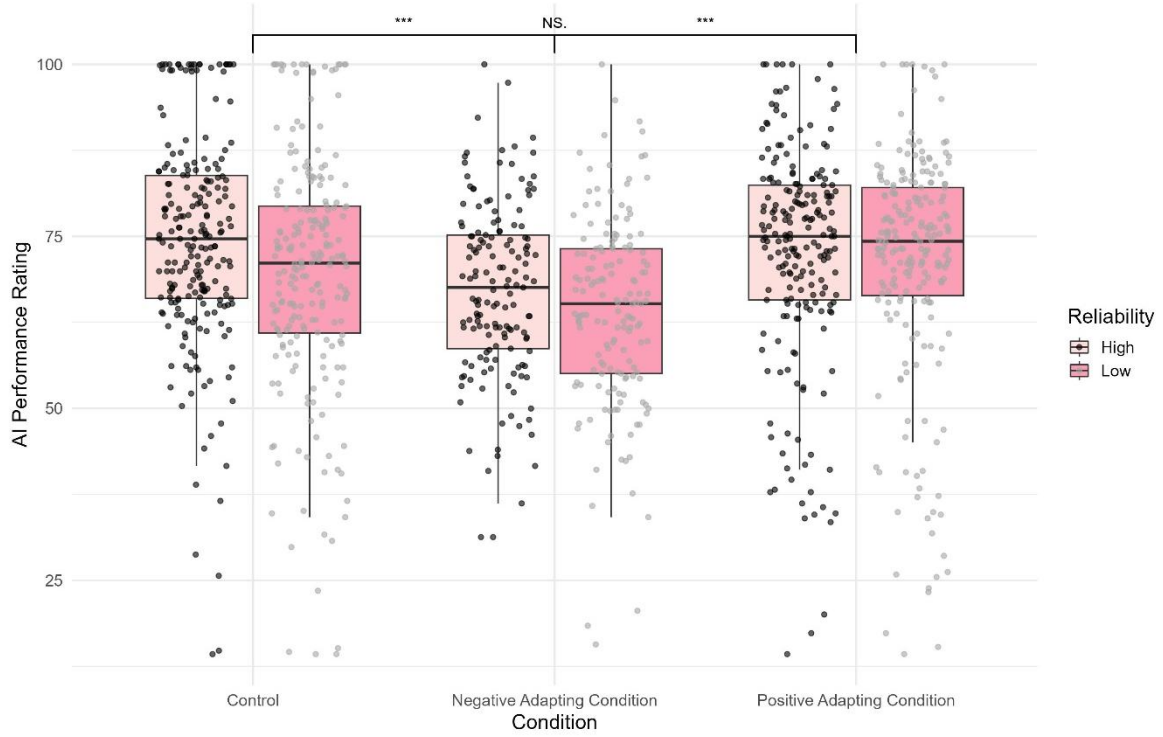
Note. **Bold** result indicates significance.

These findings suggest that the Negative Adapting condition significantly reduces trust ratings compared to other conditions, particularly under high reliability.

5.3.5 AI Performance Ratings

A linear mixed-effects model was conducted to examine the effects of condition and reliability on AI performance ratings, with participants as a random effect. The model included a main effect for condition (Positive, Negative and Control), a main effect for reliability (High, Low), and their interaction. Figure 25 shows a box plot of these results. We applied ANOVA to the model to extract significant results and used Satterthwaite's method for degrees of freedom.

Figure 25. Boxplot displays AI performance scores for the three conditions, Positive Adapting Condition, Control, and Negative Adapting Condition, with colour coding based on reliability. Significant differences between conditions are highlighted with markers. The y-axis indicates AI performance scores, while the x-axis represents the conditions.



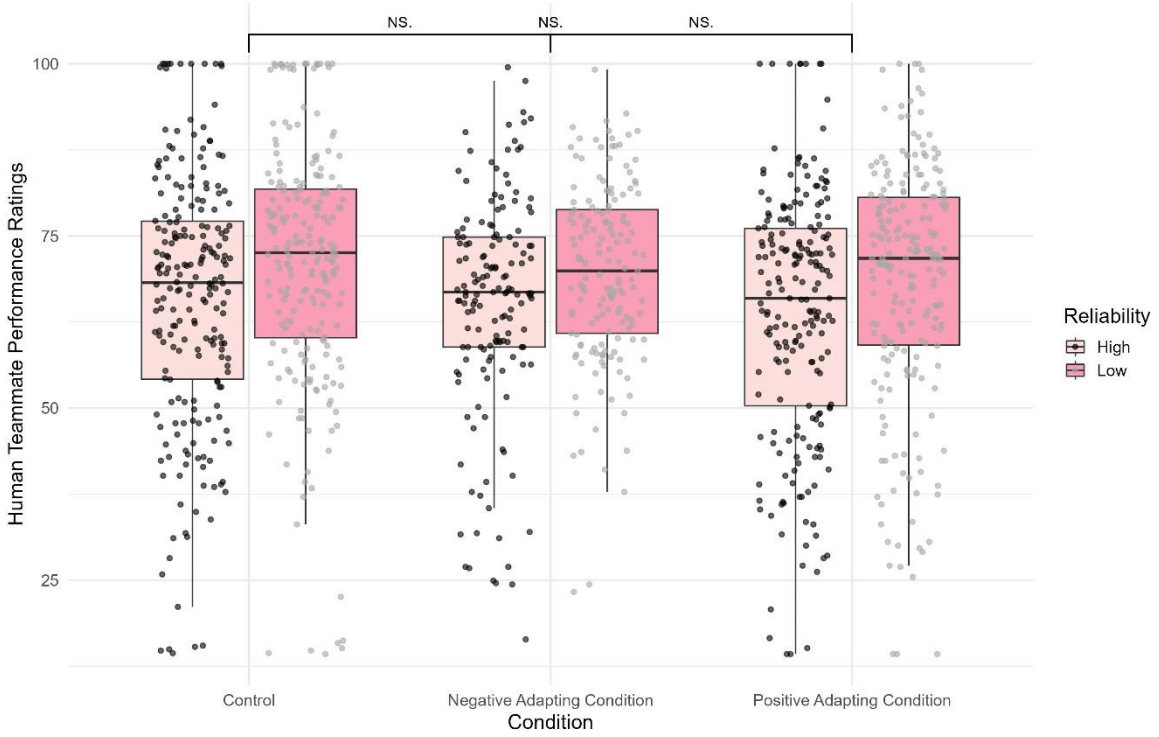
The analysis for AI ratings revealed that the main effect of Adaptability was not

significant (F(2, 27.03) = 2.52, p = 0.0990), suggesting no overall difference in AI ratings across the three conditions. However, the main effect of Reliability was significant (F(1, 1135.40) = 13.07, p < 0.001), with higher ratings observed in the High Reliability condition. The interaction between Adaptability and Reliability was not significant (F(2, 1135.39)= 1.52, p = 0.2193), indicating that the effect of Adaptability on AI ratings did not differ by Reliability level. Post-hoc comparisons revealed no significant differences.

5.3.6 Human Teammate Performance Ratings

A linear mixed-effects model was conducted to examine the effects of condition and reliability on Human performance ratings, with participants as a random effect. The model included a main effect for condition (Positive, Negative and Control), a main effect for reliability (High, Low), and their interaction. Figure 26 shows a box plot of these results. We applied ANOVA to the model to extract significant results and used Satterthwaite's method for degrees of freedom.

Figure 26. Boxplot representing human teammate performance scores across the conditions, with colours denoting reliability. Significant comparisons are marked, providing insight into differences in performance. The y-axis shows performance scores, while the x-axis categorises the conditions.



The model results for human performance ratings showed that neither Adaptability (F(2, 27.14) = 0.29, p = 0.747) nor the interaction between Adaptability and Reliability were significant (F(2, 1134.50) = 0.02, p = 0.9790), respectively,

indicating that performance ratings did not vary across conditions. However, Reliability had a significant main effect, ($F(1,1134.51) = 25.16$, $p < 0.001$), with higher ratings observed in the Low Reliability condition. Post-hoc comparisons revealed no significant differences.

These findings suggest that while human performance scores were generally higher under low reliability, the condition (Positive, Negative and Control) did not significantly influence human performance ratings.

5.3.7 The Godspeed Questionnaire

Our study conducted LMMs for the subsections of the Godspeed questionnaire as participants completed them four times throughout the experiment, so we still needed to control for the variability between participants

Figure 27. Mean Godspeed Scores for Anthropomorphism, Likeability, and Perceived Intelligence by Condition and Reliability

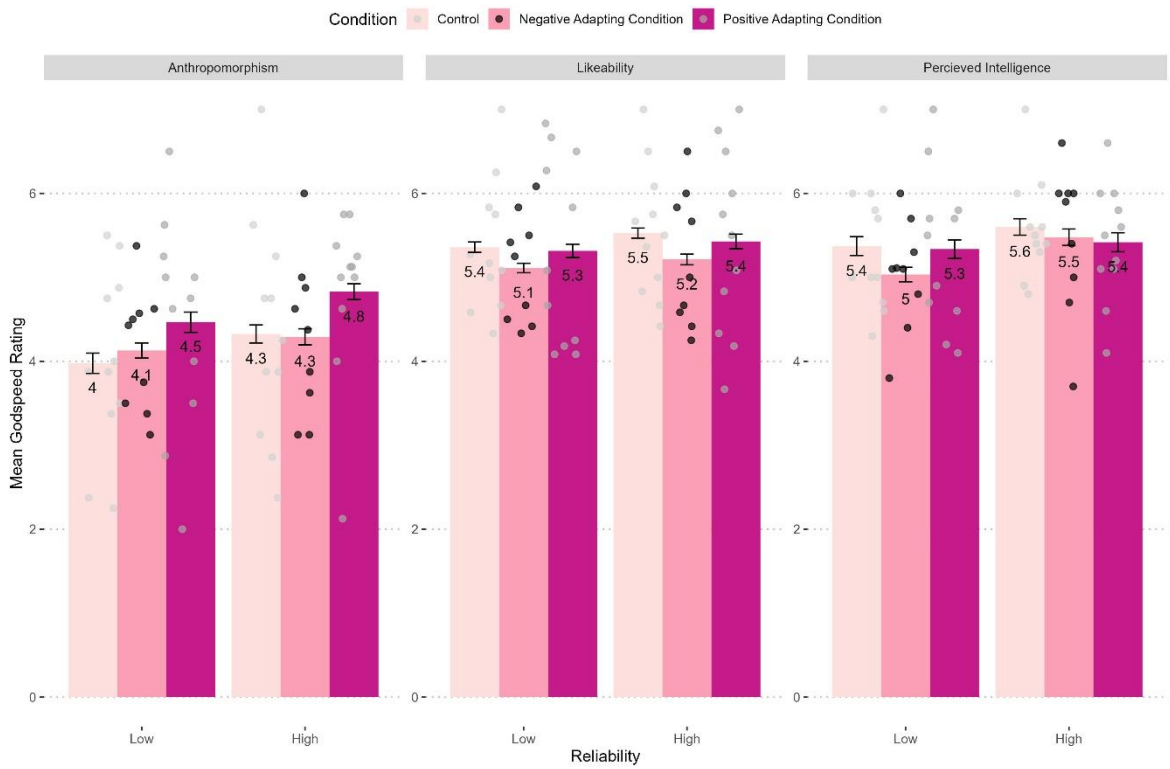


Figure 27 displays the mean Godspeed scores for three dimensions, Anthropomorphism, Likeability, and Perceived Intelligence, across different conditions and two levels of reliability (Low and High). For Anthropomorphism, scores are generally higher in the High-Reliability condition, with the Positive Adapting Condition showing the highest score (4.8). In the Likeability dimension,

scores are higher in the High-Reliability condition for all conditions, with the Control condition reaching the highest score (5.5). Perceived Intelligence scores are consistently high across conditions and reliability levels, with slightly higher values in the High-Reliability condition.

These results suggest that reliability and condition influence participants' perceptions of anthropomorphism, likeability, and perceived intelligence.

5.3.7.1 Likeability

A linear mixed-effects model was conducted to examine the effects of condition (Positive, Negative, and Control) and reliability (High, Low) on likeability ratings, with participants as a random effect. The model included main effects for condition and reliability, as well as their interaction. To gain p-values, we used a Type III ANOVA on the model.

The Type III ANOVA using Satterthwaite's method revealed a significant main effect of Reliability on likeability ratings, ($F(1, 705.01) = 5.25$, $p = 0.0222$). This suggests that likeability ratings were influenced by the level of reliability, with higher ratings in the High Reliability condition. However, the main effect of Adaptability was not significant ($F(2,28.00)=0.28$,$p=0.7568$), indicating no differences in likeability ratings across the three conditions. The interaction between Reliability and Adaptability was also not significant, ($F(2, 705.01) = 0.13$, $p = 0.8787$), suggesting that the effect of Reliability on likeability did not vary across conditions.

Post hoc comparisons revealed no significant differences between any of the conditions (all $p>0.05$).These findings suggest that while likeability ratings were slightly lower under low-reliability conditions, the specific conditions (Positive, Negative and Control) did not significantly impact likeability scores.

5.3.7.2 Anthropomorphism

A linear mixed-effects model was conducted to examine the effects of condition (Positive, Negative, and Control) and reliability (High, Low) on anthropomorphism ratings, with participants as a random effect. The model included the main effects

of condition and reliability, as well as their interaction. To gain p-values, we used a Type III ANOVA on the model.

The Type III ANOVA using Satterthwaite's method showed a significant main effect of Reliability on anthropomorphism ratings, ($F(1, 458.01) = 9.86$, $p = 0.0018$), indicating higher anthropomorphism ratings in the High Reliability condition. The main effect of Adaptability was not significant ($F(2, 27.98) = 0.72$, $p = 0.4957$), suggesting no differences in anthropomorphism ratings across the three conditions. Additionally, the interaction between Reliability and Adaptability was not significant ($F(2, 458.00) = 0.52$, $p = 0.5920$), indicating that the effect of Reliability on anthropomorphism did not depend on adaptability.

Post hoc tests did not reveal any significant pairwise differences between the conditions (all $p > 0.05$).These findings suggest that while anthropomorphism ratings were lower under low-reliability conditions, the specific condition (Positive, Negative and Control) did not significantly impact anthropomorphism scores.

5.3.7.3 Perceived Intelligence

A linear mixed-effects model was conducted to examine the effects of condition (Positive, Negative, and Control) and reliability (High, Low) on perceived intelligence ratings, with participants as a random effect. The model included main effects for condition and reliability, as well as their interaction. To gain p-values, we used a Type III ANOVA on the model.

The Type III ANOVA using Satterthwaite's method revealed a significant main effect of Reliability on perceived intelligence ratings ($F(1, 585.01) = 13.89$, $p < 0.001$), indicating that higher perceived intelligence ratings were associated with the High Reliability condition. The main effect of Adaptability was not significant ($F(2, 28.01) = 0.25$, $p = 0.7784$), showing no differences in perceived intelligence across conditions. The interaction between Reliability and Adaptabiluty was also not significant ($F(2, 585.01) = 2.34$, $p = 0.0977$).

Post hoc comparisons showed no significant differences between the conditions (all $p>0.05$).These findings suggest that while perceived intelligence ratings were

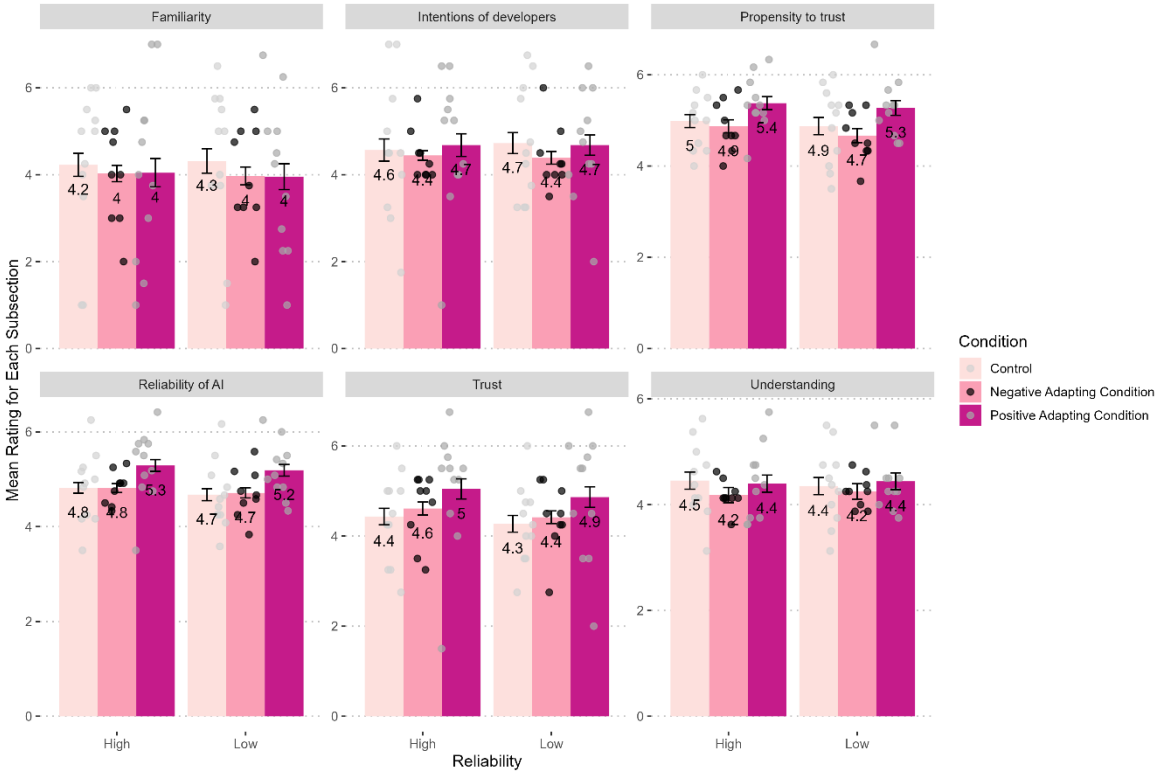
lower under low-reliability conditions, the specific condition (Positive, Negative and Control) did not significantly impact perceived intelligence scores.

5.3.8 Trust in AI Scores

Our study conducted LMMs for the Trust in AI questionnaire subsections as participants completed them four times throughout the experiment, so we still needed to control for the variability between participants.

Figure 28 displays the mean trust scores for six dimensions, Familiarity, Intentions of Developers, Propensity to Trust, Reliability of AI, Trust, and Understanding, across different conditions and two levels of reliability. Error bars indicate standard errors. Although there is some difference between means the analysis revealed no significant main effects or interactions.

Figure 28. Bar plot depicting the mean trust scores across different reliability levels for various conditions. The data is further segmented by subsection, for comparison of trust levels across conditions and reliability ratings.



5.4 Discussion

The primary goal of this research chapter was to explore how AI social adaptability and reliability impact trust, influence, and team performance in HATs.

Specifically, the study examined how participants perceived and interacted with AI under three conditions: Positive Adaption (AI tailored to user preferences), Negative Adaptation (AI that operates counter to user preferences), and Control (AI without specific adaptation to user preferences). Additionally, the study considered the impact of high and low reliability in each condition, aiming to better understand trust dynamics in human-AI collaboration.

In this chapter, we observed nuanced interactions between AI adaptability, trust, and performance, leading to mixed support for our hypotheses. Hypothesis 1, which predicted higher performance accuracy in the Adapting condition compared to the Negative Adapting condition, received partial support as participants indeed scored higher in these conditions; however, performance ratings by participants did not consistently align with actual performance, limiting conclusive support for this hypothesis. Hypothesis 2, predicting lower trust in the Negative Adapting condition, was supported, with lower influence and trust ratings reflecting a sensitivity to AI misaligned with user preferences. Conversely, Hypothesis 3, which suggested that AI adaptability would enhance trust regardless of reliability, was not upheld, as trust scores were not significantly higher in the Positive Adapting Condition than in the Control condition. Interestingly, participants rated their human teammate's performance more favourably when the AI's reliability was low, pointing to the impact of team dynamics on perceived performance.

5.4.1 Trust Ratings

Our findings reveal a strong positive correlation between trust and influence ratings, indicating a meaningful relationship between implicit and explicit trust measures in this experiment. Specifically, influence ratings allow us to accept Hypothesis 2, as participants displayed lower trust in the AI in the Negative Adapting condition. Influence ratings were significantly higher in the Control and Positive Adapting Conditions compared to the Negative Adapting condition, suggesting that participants were more influenced by an AI aligned with their preferences. Interestingly, however, participants appeared even more sensitive to the AI's behaviour when it contradicted their preferences, as shown by more significant trust erosion in the Negative Adapting condition. This finding aligns with recent research suggesting that user-contradicting AI systems lead to trust erosion

due to cognitive dissonance and discomfort. For instance, Glikson and Woolley (2020) highlight that unexpected or contradictory AI behaviours can elicit negative emotional responses. This supports the idea that socially misaligned AI may disrupt trust and heighten user sensitivity to the AI's actions.

Our results also indicate that preference-opposing adaptation has a more significant impact on reducing trust than preference-supporting adaptation has on building it. This aligns with the trust asymmetry effect, where negative experiences disproportionately impact trust, making rebuilding challenging (Schaefer et al., 2016; Zhu et al., 2021). In our study, trust ratings closely mirrored influence ratings and, with participants expressing more trust in the Adapting and Control conditions than in the Negative Adapting condition. This supports the notion that participants were more reactive to an AI that contradicted their preferences, potentially because humans weigh negative experiences more heavily than positive ones (Jones-Jang & Park, 2023). These findings offer insights into the emotional dynamics when AI behaviour diverges from user expectations. This sensitivity may be especially pronounced in HATs operating in trust-critical environments, such as healthcare or autonomous vehicles.

Moreover, while these findings support Hypothesis 2, they do not conclusively support Hypothesis 3, which predicted that AI adaptability in the Positive Adapting Condition would lead to higher trust scores irrespective of reliability. Although the TiA questionnaire showed some marginally significant results, it ultimately failed to support Hypotheses 2 or 3 definitively. Nonetheless, mean trust scores across questionnaire subsections remained consistently lowest in the Negative Adapting condition, which warrants further investigation. The trend of decreased trust with misalignment aligns with findings by Rahwan et al. (2019) and Hengstler et al. (2016), who demonstrated that autonomy-supportive AI systems, those that align closely with user needs and behaviours, tend to enhance trust. Misaligned AI adaptations, conversely, can undermine users' perceived control and autonomy, eroding trust more significantly than reliability discrepancies alone. This sensitivity to AI alignment mirrors socialisation theories in human teams, where initial alignment with norms and expectations promotes rapport and trust (Chao et al., 1994; Saks & Gruman, 2014).

In this context, the principle of adaptive personality in human teams offers a valuable lens. As Chao et al. (1994) and Van Maanen & Schein (1977) suggested, adaptive socialisation involves modifying behaviour to align with team norms, enhancing integration and trust. Just as adaptive socialisation helps align team members' behaviours with organisational culture, socially adaptive AI in HATs could support alignment with user preferences, fostering smoother collaboration. This parallels research on adaptive autonomy in AI, where agents dynamically adjust their autonomy to match team needs (Hauptman et al., 2023). While our study focused on initial preference-matching rather than dynamic adaptation, the importance of alignment for trust building aligns well with these adaptive frameworks. Such an approach suggests that personalised, preference-matched AI may be a practical foundation for real-world HATs, where dynamic adaptation may not be feasible.

5.4.2 AI and Human Performance Ratings

Our findings for Hypothesis 1 reveal that the percentage of correct answers was highest in the Adapting and Control conditions and lowest in the Negative Adapting condition, particularly in the high-reliability condition. Specifically, participants in the Negative Adapting condition scored around 10% lower than in the Adapting and Control conditions, supporting Hypothesis 1 that preference alignment positively influences performance. This finding is consistent with studies showing that AI, which aligns with user preferences, can enhance performance by fostering smoother, more intuitive interactions (Liu et al., 2003; Strauss, 2017). The parallel between performance and trust results is particularly intriguing: both indicate that an AI's alignment, or lack thereof, with user preferences significantly impacts outcomes. This suggests that the mismatch between user expectations and the AI's behaviour drives declines in trust and performance.

Interestingly, participants' performance ratings did not vary significantly across conditions or by reliability, indicating a discrepancy between actual and perceived performance (Bansal et al., 2019; Chavaillaz et al., 2016). This suggests that participants may not accurately assess performance differences based on reliability, possibly due to cognitive biases or the influence of unrelated factors like AI likability and perceived intelligence. For instance, participants rated the AI more

favourably on traits like likability, anthropomorphism, and perceived intelligence in high-reliability conditions, suggesting that reliability influenced their perceptions more than adaptive behaviour. This aligns with studies on anthropomorphism and perceived intelligence in AI, which indicate that perceived reliability can influence user perceptions even when it does not necessarily reflect actual performance ((de Visser et al., 2016; Roy & Naidoo, 2021; Waytz et al., 2014). These discrepancies underscore that adaptive preferences alone may not enhance perceived performance, suggesting a nuanced interplay between reliability, adaptation, and perceived attributes.

However, trial AI performance ratings from the experiment did not fully support Hypothesis 1, indicating that the effects of adaptability on performance are complex and may not straightforwardly translate into perceived performance gains. Likewise, our findings did not provide evidence for Hypothesis 3, which posited that adaptability to user preferences or opposition to those preferences would directly impact perceived performance. This may suggest that adaptability alone, while valuable in specific contexts, might not significantly influence performance perceptions without concurrent reliability signals.

An unexpected finding was that participants rated their human teammate's performance higher when the AI's reliability was low. This suggests a compensatory effect where lower-performing AI may inadvertently enhance perceived human performance. This effect may stem from comparison bias or a shift in expectations: as the AI's performance dips, participants may adjust their perceptions, viewing human contributions as comparatively more substantial (Jones-Jang & Park, 2023). This result resonates with findings on team dynamics, where perceived performance can be influenced by contrasting behaviours within the team (Glikson & Woolley, 2020).

In summary, the percentage of correct answers supports Hypothesis 1 by indicating that preference alignment with the AI positively impacts performance, though this effect does not extend to participants' performance ratings. The observed discrepancies between actual and perceived performance highlight the complex role of AI adaptability, user expectations, and reliability in shaping

performance perceptions. Future research could explore how relative reliability within HATs influences perception and trust and whether training on adaptive AI behaviours could help mitigate these biases, ultimately creating more balanced perceptions of AI and human team members.

5.4.3 Limitations

While the findings of this chapter provide important insights into the impact of AI social alignment on trust and performance within human-AI teams, several limitations should be acknowledged. First, although participants were assigned to conditions based on their self-reported communication preferences, the distribution across preference types was not fully balanced across experimental groups. Specifically, the Adapted AI (PA) condition primarily included participants with friendly preferences, while the Non-Adapted AI (NoA) condition consisted mainly of participants with formal preferences. This uneven distribution may confound the interpretation of results, as some effects attributed to alignment could be partially influenced by baseline differences in interaction preferences.

Second, while the manipulation of AI alignment (adapted vs. non-adapted) was effective, it is possible that some participants may not have fully noticed or interpreted the AI's communication style as intended. Future research could incorporate manipulation checks to assess perceived alignment more directly.

Third, the measures of trust and performance were captured over a relatively short interaction period. As trust in AI is known to evolve over time, longitudinal or repeated-interaction designs may provide a more comprehensive picture of how social alignment affects sustained collaboration.

Finally, this study focused on only four binary interaction traits (e.g., friendly vs. formal, short vs. long). Real-world communication preferences are likely more nuanced and dynamic. Future studies could explore more flexible and adaptive AI models that adjust to individual communication patterns over time, rather than relying on static assignments.

Addressing these limitations in future work would strengthen the generalizability of the findings and help refine guidelines for designing socially aligned AI teammates.

5.4.4 Conclusions

This chapter's findings highlight that AI adaptability, when designed to align or misalign with user preferences, significantly affects trust and team dynamics within HATs. Results indicate that user preference alignment enhances trust and influence ratings, while misalignment detrimentally impacts trust, underscoring the asymmetrical effect of negative experiences on trust calibration. Furthermore, adaptability's impact on perceived human teammate performance, mainly when low AI reliability, emphasises the importance of relative reliability of AI teammates and role clarity in team dynamics.

These insights advance the thesis's exploration of how various AI design factors, reliability, human likeness, and adaptability influence HAT performance. This chapter also raises questions about adaptability's role as a foundational design strategy versus a dynamic trait, suggesting that adaptability alone may not suffice to build optimal HATs without reliable, context-aware support. The concluding chapter will synthesise findings from each experimental chapter, discussing practical implications for designing adaptive, reliable AI teammates that enhance trust and cohesion in real-world, high-stakes applications.

Chapter 6 Conclusions

6.1 Introduction

Integrating AI as a teammate within HATs represents a fundamental shift in how technology supports and collaborates with human users. This thesis has explored critical dimensions of this shift, focusing on the dynamics of trust, anthropomorphism, and SI within HATs. Through a detailed bibliometric analysis of trust research, experimental studies on the impact of AI characteristics on human collaboration, and the exploration of theoretical frameworks, this work has aimed to illuminate the complex relationship between humans and AI agents as evolving teammates rather than mere tools.

Central to this thesis is the idea that for AI to function as an effective teammate, it must move beyond transactional roles, incorporating qualities that foster trust, mutual understanding, and team cohesion. Chapters have progressively examined how HATs are shaped by trust calibration (de Visser et al., 2020; Lee & See, 2004) , the role of anthropomorphism in facilitating human-like interactions (de Visser et al., 2016; Glikson & Woolley, 2020), and the potential for socially intelligent AI systems to enhance collaborative outcomes (Dautenhahn, 1995; Kox et al., 2022; Nass et al., 1994a; Williams et al., 2022; Zadeh et al., 2019). By addressing these interrelated aspects, this thesis has highlighted the need for a human-centric approach in HAT design that emphasises technical capability, social adaptability, and ethical considerations.

This concluding chapter synthesises the primary findings across these domains, reflecting on the unique contributions and limitations of the research. It also identifies future directions essential for advancing the field of HATs. These include investigating long-term trust dynamics, refining AI's social capabilities and developing ethical frameworks to govern the growing role of AI in human teams. The insights drawn from this research are intended to guide theoretical and practical advancements in designing AI teammates that align with human expectations and values, ensuring productive and trustworthy HATs in a rapidly evolving technological landscape.

6.2 Summary of Key Findings

This research explored the dynamics of trust, reliability, perceived performance, and ethical considerations that influence HATs. We can comprehensively understand how these factors interact and shape user experiences and team outcomes by integrating the results across multiple chapters.

6.2.1 Trust

Trust emerged as the foundational component for successful collaboration in HATs. The literature review in Chapter 1 set the stage, emphasising that trust is central to human-machine collaboration, with calibrated trust being vital for productive interactions (de Visser et al., 2020; Hoff & Bashir, 2015; Lee & See, 2004; Muir, 1994). Trust is not static but dynamic, requiring users to continuously adjust their confidence based on the system's actions and experiences (Glikson & Woolley, 2020; Li et al., 2023; Reinhardt, 2023). This dynamic process was confirmed in the experimental chapters, where trust ratings were influenced by system reliability, adaptability, and the presence of relational cues.

Chapter 2's bibliometric analysis broadened our understanding of trust research by mapping the scope of existing studies, highlighting the central role of trust in various contexts, and contextualising human-AI teaming within this expansive field. This helped us better appreciate the position of human-AI trust research within the larger body of trust literature and understand the importance of defining and measuring trust appropriately.

In Chapter 3, trust was significantly enhanced in low-reliability conditions when AI systems were anthropomorphised. This suggests that in situations where AI reliability is compromised, human-like features can help users feel more comfortable and mitigate the negative impact of unpredictability, which supports previous research (Bittner et al., 2019; de Visser et al., 2016; Roy & Naidoo, 2021; Seymour & Van Kleek, 2021; Złotowski et al., 2015). Conversely, in high-reliability conditions, trust was less influenced by anthropomorphism, highlighting that trust is fundamentally rooted in the system's technical performance. These findings reinforce that users calibrate their trust based on system reliability,

adjusting their expectations according to the AI's perceived ability to perform effectively (Heyder et al., 2023; Jensen et al., 2021; Troshani et al., 2021).

Chapter 4 further explored the impact of EI on trust, using emojis to evoke emotional intelligence (Beattie et al., 2020; Fadhil et al., 2018). While these affective cues enhanced user perceptions of the AI's likeability, they did not significantly impact the actual trust ratings or performance outcomes in task settings. These findings suggest that cognitive rather than emotional factors primarily drive trust in HATs. However, the findings in Chapter 4 are not strong enough to conclude with confidence. Future work could use newly developed measures to investigate the role of emojis (Shang et al., 2024). It is also possible that emojis could potentially play a more significant role in fostering trust in specific contexts, such as high-stakes environments or emotionally charged tasks if emojis are aligned explicitly with an emotional state (Beattie et al., 2020; Boutet et al., 2021; Fadhil et al., 2018; Rajan et al., 2023).

Chapter 5 emphasised the importance of AI adaptability in trust formation. The research found that trust was significantly lower in the Negative Adapting condition, where the AI's behaviour conflicted with user preferences, highlighting the importance of alignment between user expectations and AI behaviour. This finding is aligned with the "trust asymmetry effect" (Poortinga & Pidgeon, 2004; Zhu et al., 2023), where negative experiences disproportionately impact trust compared to positive interactions. Overall, trust in AI systems requires a careful balance of technical reliability and adaptability, with misalignment causing more harm than alignment benefits.

In conclusion, trust is the cornerstone of effective collaboration in HATs, and this research highlights its dynamic and context-dependent nature. As established in the literature review, calibrated trust is essential for productive HATs, with users continuously adjusting their trust based on the AI system's reliability, preference alignment, and anthropomorphised features. This was corroborated across the experimental chapters, where trust ratings were found to be strongly influenced by system performance and the presence of anthropomorphism or preference alignment. In low-reliability conditions, anthropomorphic features helped mitigate

trust erosion, but in high-reliability contexts, trust was more firmly anchored in the AI's technical performance.

Additionally, while emotional cues, such as emojis, enhanced the AI's likeability, they did not significantly impact trust outcomes. This suggests that cognitive factors may outweigh emotional ones in driving trust in HATs; however, this area of work needs further investigation. Finally, adaptability was identified as a critical factor, with misalignment between the AI's behaviour and user expectations significantly reducing trust. Overall, these findings reinforce the idea that trust in AI systems requires a delicate balance of reliability, adaptability, and alignment with user needs, carefully considering when and how anthropomorphic cues can play a role in fostering trust (Chen & Park, 2021; Jensen et al., 2021; Kim & Song, 2021; Kulms & Kopp, 2019; Seymour & Van Kleek, 2021; Troshani et al., 2021).

6.2.2 Reliability

The research consistently identified reliability as a critical determinant of trust, performance, and user perceptions. In Chapter 3, system reliability was the most significant factor influencing trust and performance ratings. AI systems perceived as reliable boosted user confidence, regardless of whether they were anthropomorphic. This finding reaffirms that technical performance remains the cornerstone of trust in HATs (Glikson & Woolley, 2020; Henrique & Santos, 2024; Lahusen et al., 2024; Ryan, 2020).

Reliability emerged as a consistent and pivotal factor in shaping trust, performance, and user perceptions across the research. As highlighted in Chapter 3, system reliability was the most significant influence on trust and performance ratings, with reliable AI systems boosting user confidence regardless of whether they featured anthropomorphic traits. This underscores the central role of technical performance in fostering trust in HATs.

Chapter 5 further demonstrated that reliability remains crucial even when considering adaptability. In the Negative Adapting condition, where AI behaviour conflicted with user preferences, performance ratings were notably lower, particularly when the system's reliability was high. This finding suggests that

reliability not only forms the bedrock of trust but also serves as an anchor for performance evaluations, with misalignment between user expectations and AI behaviour detracting from trust and perceived effectiveness. These findings build upon older work that found reliability to be a driver of trust in automation and robotics (Chavaillaz et al., 2016; Desai et al., 2012).

Overall, these findings affirm that in HATs, reliability is a critical driver of trust and performance, with adaptability and anthropomorphism serving as complementary factors that enhance the experience without overshadowing the foundational importance of reliability, which builds on previous works in hot HATs and human teams (Hariri et al., 2015; Hauptman et al., 2023; Klarner et al., 2013; Pulakos et al., 2006; Zhao et al., 2022).

6.2.3 AI Perceived Performance

AI's perceived performance was intricately linked to system reliability and human-like features. Chapter 3 found that in low-reliability conditions, anthropomorphic features led to higher trust but also resulted in lower performance ratings. This indicates that while anthropomorphic designs can improve the initial perceptions of an AI's trustworthiness, they may also elevate expectations about its performance, which, when unmet, can lead to negative evaluations (Poortinga & Pidgeon, 2004; Zhu et al., 2023).

In contrast, high-reliability systems consistently achieved better task performance, regardless of whether they were anthropomorphic. This highlights a fundamental takeaway: while anthropomorphic cues can enhance perceived trustworthiness in specific contexts, technical accuracy and consistency are far more critical in shaping actual performance perceptions (Chavaillaz et al., 2016; de Visser et al., 2016; Glikson & Woolley, 2020). This reinforces findings from HRI that reliability drives actual performance outcomes, while human-like features may influence more subjective dimensions, such as likeability and familiarity (Honig & Oron-Gilad, 2018; Reeves et al., 2020).

Chapter 4's examination of emojis revealed that while these emotional cues influenced perceptions of human teammates' performance, they had no

substantial impact on AI performance ratings in task-based settings. This suggests that while emotional cues can improve team cohesion and perceptions of AI "likeability", they are not significant drivers of perceived AI performance, particularly in scenarios that demand high cognitive focus (Flathmann et al., 2023). It is evident that system reliability remains the most critical factor in determining how users evaluate the AI's performance in collaborative tasks.

6.2.4 Human Teammate Performance

The findings across these chapters, supported by broader research, offer valuable insights into human teammate performance in HATs , revealing how trust, adaptability, and team dynamics shape perceptions and outcomes. A notable observation is the compensatory effect of AI reliability: human performance ratings can increase when AI reliability is low. This phenomenon, noted in Chapter 5, aligns with research suggesting that when one team member (human or AI) underperforms, others are viewed as more capable by comparison (Glikson & Woolley, 2020; Endsley, 2023). This dynamic underscores that the performance of human teammates is closely tied to the relative performance of the AI, reflecting the nuanced interplay of team roles and expectations (McNeese et al., 2021).

Additionally, the relationship between trust and team dynamics plays a crucial role in shaping perceptions of human performance. When trust in the AI erodes, as in the Negative Adapting condition explored in Chapter 5, participants rely more heavily on their human teammates, enhancing their perceived performance. Trust dynamics influence how other team members are valued (McNeese et al., 2021). These findings align with broader studies emphasising the importance of trust calibration for effective collaboration in HATs; misplaced trust can distort perceptions and lead to over- or under-reliance on AI (de Visser et al., 2020; Hauptman et al., 2023).

While emotional cues like anthropomorphic design or affective communication from AI enhance its perceived likeability and trustworthiness, they have limited direct impact on human teammate evaluations. This aligns with research indicating that affective cues in AI influence trust in the AI itself rather than the wider team

(Waytz et al., 2014; Fadhil et al., 2018). Furthermore, these emotional cues often do not compensate for broader task-related dynamics like reliability or role clarity, which remain the primary determinants of trust and perceived performance (Janhunen et al., 2024; McNeese et al., 2021).

Human performance ratings remained consistent across experimental conditions regarding actual contribution, even when the AI's behaviour varied. This reflects the influence of comparative dynamics and role interdependence over objective measures. When AI reliability decreased, humans were perceived as stepping into a more critical role despite no changes in their actual behaviour. This finding aligns with team theories emphasising the importance of clear roles and interdependence for effective collaboration (Chai et al., 2017; Saks & Gruman, 2014). Moreover, as noted in recent reviews, transparency and role clarity are essential for fostering effective HAT performance (McNeese et al., 2018, 2021).

6.2.5 Conclusion of Findings

This research reveals that trust in HATs is primarily shaped by system reliability, with adaptability and relational cues playing supplementary roles. Reliability is the foundation for trust and perceived performance, while AI's adaptability helps maintain trust when aligned with user preferences. However, misalignment and overly anthropomorphic or emotional cues risk creating challenges by miscalibrating user trust. These findings highlight the need for balanced AI designs prioritising transparency, reliability, and ethical considerations to foster effective and sustainable human-agent collaborations in diverse contexts.

6.3 Contributions to the Field

This thesis significantly contributes to HATs, trust in AI, and human-AI collaboration by developing new theoretical insights, offering practical design guidance and introducing novel methodologies for studying and measuring human-AI interactions. By addressing trust, anthropomorphism, and social alignment, this work extends beyond existing literature to establish a comprehensive foundation for understanding AI as a teammate rather than a

mere tool. It also explores the impact of having multiple human team members, setting the stage for future advancements in collaborative workplace AI.

6.3.1 Theoretical Contributions

This thesis expands on existing frameworks of trust in HATs by examining trust as a dynamic, context-sensitive construct specifically suited to HATs. Previous studies, such as those by Lee and See (2004), focused on trust calibration in automation, emphasising the need for trust to align with system reliability and transparency. Building on this foundation, this thesis demonstrates that trust in AI teammates requires a nuanced approach considering the social and psychological dimensions of human-machine interactions.

Trust in HATs is shown to hinge not only on reliability and transparency but also on AI's anthropomorphism and preference alignment, which can lead to more collaborative and effective relationships. In expanding the notion of trust to reflect the unique demands of HATs, this work provides a more precise investigation to inform the design of AI systems capable of fostering sustainable trust with human collaborators. It also highlights the issues of humanising AI, as it can lead to increased levels of trust when reliability is low, resulting in overtrust in the system (Robinette et al., 2016). For these reasons, it is essential to take a steady approach to developing AI that is more human and capable of social alignment to ensure that the system's robustness justifies the positive impacts on trust and likeability.

This thesis also contributes to the theoretical understanding of anthropomorphism and user alignment within HATs, integrating insights from the CASA paradigm (Nass et al., 1994) to explore how human-like qualities in AI can foster trust. While anthropomorphism has been widely studied, this work extends the application of anthropomorphic cues to HATs, investigating the conditions under which such cues enhance collaboration. The findings reveal that moderate anthropomorphism, manifested through human-like language and nonverbal cues, can enhance user trust and promote collaboration. This study supports previous research on the benefits of SI in team settings (Boyatzis et al., 2017; Williams et al., 2022; A. Zhang & Patrick Rau, 2022) and establishes the need for a balanced

approach to anthropomorphism, where AI is designed to convey human-like qualities while avoiding discomfort or manipulation. This balanced perspective aligns with the work of Troshani et al. (2021) and Glikson and Woolley (2020), who emphasise anthropomorphism's contextual and ethical considerations, particularly in fostering emotional connections without misleading users.

Finally, this work is unique as it also considers the presence of a human teammate. Contextual factors like AI reliability, adaptability, and emotional presentation influence human teammate performance in HATs. Ratings of human performance often reflect relative AI performance and team dynamics rather than an objective assessment of human contributions akin to the halo effect (Lachman & Bass, 1985; Naquin & Tynan, 2003; Nicolau et al., 2020). This underscores the importance of designing AI systems that foster clear roles and balanced dynamics to enhance team collaboration.

These insights underscore the importance of anthropomorphism and social alignment in AI, adding a new dimension to HAT theory by emphasising that AI teammates should be designed for functional efficiency and to support emotional engagement and social cohesion within teams.

6.3.2 Practical Implications

This thesis offers actionable insights for designing and implementing AI in collaborative settings. The empirical findings suggest that specific design principles, such as optimal levels of anthropomorphic language and the strategic use of social alignment, can encourage trust and engagement. These recommendations provide practical information for developing AI teammates who align with human social expectations and optimise team dynamics.

From a practical perspective, AI developers and designers are encouraged to adopt anthropomorphic features judiciously, focusing on transparency and reliability as foundational design principles. For instance, the findings demonstrate that while anthropomorphic design can increase user trust and likeability, it must not obscure the AI's actual capabilities or limitations. This aligns with recommendations by Hauptman et al. (2023) and Schelble et al. (2022),

advocating for transparent communication of AI strengths and boundaries to avoid ethical dilemmas such as over-trust or misuse of AI in critical settings.

Ethically, this work underscores the responsibility of designers to ensure that anthropomorphic AI promotes informed collaboration rather than manipulation. By creating systems that appear "human-like", there is a risk of users overestimating AI capabilities or forming inappropriate emotional connections, as highlighted by Waytz et al. (2014). This calls for a balanced approach that respects user autonomy while enhancing team dynamics. These findings support a broader discourse on the ethical integration of AI, emphasising the need for systems that are both user-centric and grounded in ethical transparency, ensuring that anthropomorphism serves as a tool for collaboration rather than deception

As demonstrated in studies by Fussell et al. (2008) and Pelau et al. (2021), nonverbal cues can significantly enhance user engagement when applied judiciously. However, this research goes further by empirically testing the influence of these cues within the unique context of HATs, where the balance of trust and comfort is essential to avoid over-reliance or discomfort. These insights are particularly relevant for AI developers and designers, providing a roadmap for integrating anthropomorphic features and user alignment into AI systems to enhance trust calibration and overall team effectiveness.

Additionally, this thesis emphasises the need for user-centric adaptability in AI teammates, contributing to a growing body of research advocating for AI systems responsive to user preferences and team dynamics. By demonstrating that adaptable AI teammates who align their communication style to mimic user preferences are perceived as more trustworthy and practical, this work supports a user-centred approach to AI design. These findings echo calls from Berretta et al. (2023) and Schelble et al. (2022) for adaptive AI systems that enhance engagement and collaboration by aligning with team members' needs. The emphasis on adaptability highlights a practical pathway for future HATs, where AI's responsiveness to social and contextual cues strengthens team cohesion and reduces friction in collaborative environments.

6.3.3 Ethical Contributions

Across all chapters, ethical and contextual considerations emerged as critical when designing HATs. Using anthropomorphic features and emotional cues like emojis raised concerns about the potential for miscalibrating trust, particularly in low-reliability conditions. For instance, Chapter 4 demonstrated that emojis could soften negative perceptions of AI but did not improve trust or performance. Emotional anthropomorphism can foster pseudo-intimacy, encouraging users to form emotional attachments to AI that lack depth or authenticity, potentially leading to inappropriate reliance on critical domains such as healthcare or finance (Placani, 2024). Therefore, ethical design principles must prioritise transparency and reliability, ensuring that anthropomorphic cues or adaptability do not obscure an AI system's actual capabilities.

Additionally, educating users about the limitations of AI systems is essential. Through clear communication of AI's scope and constraints, user empowerment can reduce over-reliance and promote informed collaboration. For instance, incorporating ethical design features, such as explainable AI, can enhance user understanding and encourage the appropriate use of emotional cues in contextually relevant ways without overshadowing the system's technical attributes (Endsley, 2023; Kim et al., 2023; Nasir et al., 2024; Ribeiro et al., 2016).

6.3.4 Methodological Contributions

Methodologically, this thesis advances the study of trust in AI by using bibliometric analysis to map the interdisciplinary landscape of trust research. By tracing trust research across psychology, computer science, and organisational behaviour, this analysis reveals the evolution and interconnections within trust literature, highlighting shifts from interpersonal trust to trust in automated and AI systems (Rousseau et al., 1998; Glikson & Woolley, 2020). This approach provides a foundational understanding of trust research and equips future researchers with a framework to explore interdisciplinary linkages and trends in trust and AI. By identifying key themes and influential works, this analysis offers a comprehensive

view of the trust landscape, establishing a basis for future studies examining trust within HATs.

The experimental methodologies employed in this thesis also contribute to HAT research. These studies introduce a controlled framework for examining trust calibration in HAT contexts by systematically varying anthropomorphic features and nonverbal cues such as emojis. These experiments build on the work of Glikson and Woolley (2020) and Robinette et al. (2016) by providing replicable designs that other researchers can adapt to explore additional anthropomorphic design and SI variables.

This thesis advances theoretical frameworks, practical design principles, and methodologies for understanding and improving human-AI collaboration. By positioning trust, anthropomorphism, and AI social alignment as integral to practical HATs, this work not only addresses some of the critical challenges in AI teammate design but also sets a foundation for future research that aims to create AI systems that are trustworthy, adaptable, and ethically designed for productive HATs. Finally, the work focuses on the ratings of other human teammates to understand further how AI can influence team dynamics.

6.4 Limitations

While this thesis offers valuable insights into the development of HATs, trust in AI, and the roles of anthropomorphism and SI, several limitations should be considered. These limitations relate to the context and generalizability of findings, methodological constraints, and challenges inherent to studying complex human-AI interactions. Recognising these limitations provides a basis for refining future research and enhancing the practical application of these findings.

One significant limitation concerns the contextual constraints of the experimental studies, which were conducted in controlled environments that may not fully capture the complexity of real-world HATs. Rix (2022) and Berretta et al. (2023) highlight that studies in controlled settings often struggle to replicate the diverse, dynamic conditions encountered in practical applications. While controlled experiments allow for precision in examining specific variables (such as

anthropomorphism and nonverbal cues), they inherently limit the ability to generalise findings to broader contexts where human-AI collaboration is subject to changing environmental factors, varying task demands, and differing organisational cultures. Consequently, the applicability of these findings to highly dynamic settings, such as emergency response or complex team-based decision-making scenarios, may be restricted.

Another limitation relates to measuring trust and SI in AI, which are inherently complex constructs. Despite efforts to develop and apply specific metrics, such as adaptability and emotional cues, measuring these qualities remains challenging due to the absence of standardised frameworks. As outlined by Lee and See (2004), existing measures of trust in automation are generally designed for simple, transactional interactions. They may only partially capture the relational dynamics of trust in HATs, where trust is continually calibrated and influenced by social factors. This limitation underscores the need for more robust, context-sensitive tools to accurately measure trust in AI teammates, especially as trust in AI may evolve differently from trust in human teammates. Moreover, the development of standardised metrics for SI in AI remains in its infancy, which restricts the ability to conduct comparative studies and hinders the broader applicability of these findings.

Finally, a notable methodological limitation is the reliance on short-term interactions in experimental settings. Trust and SI in HATs are dynamic constructs that evolve through repeated interactions, feedback, and observed behaviour, as highlighted in studies by Robinette et al. (2016) and Williams et al. (2022). This thesis, however, primarily focuses on short-term trust calibration and immediate reactions to anthropomorphism and social cues. While these findings offer valuable insights into initial trust formation, they may not fully capture the long-term dynamics of trust in HATs, which is critical for applications where human-AI interactions occur over extended periods. Longitudinal studies are needed to explore how trust and perceptions of SI evolve in AI teammates over time, providing a more comprehensive understanding of sustained human-AI collaboration.

In summary, while this thesis substantially contributes to studying human-AI collaboration, these limitations highlight the field's complexity and scope. Addressing these limitations in future research will be essential for advancing the theoretical and practical understanding of HATs, particularly in creating AI teammates that are trustworthy, adaptable, and ethically designed to support effective human-AI collaboration across diverse settings.

6.5 Future Directions

The findings and limitations of this thesis highlight several promising avenues for further research in HATs, particularly in areas of trust dynamics, anthropomorphic design, SI, and ethical AI development. By pursuing these directions, future studies can deepen our understanding of human-AI collaboration and contribute to creating effective, adaptable, and ethically designed AI teammates.

One pressing area for future research is the exploration of trust dynamics over extended periods. This thesis has focused primarily on short-term interactions, offering insights into the initial stages of trust calibration and anthropomorphic influence. However, trust in HATs is not static; it evolves through repeated interactions and may fluctuate based on AI performance, adaptability, and reliability over time (Robinette et al., 2016; Williams et al., 2022).

Longitudinal studies that track trust over time and in varied real-world settings would provide a more comprehensive understanding of how sustained human-AI collaboration impacts trust. For instance, future research could examine how initial trust formation and subsequent trust breaches or repairs influence long-term collaboration. Such studies could utilise a mixed-methods approach, combining quantitative trust metrics with qualitative assessments to capture the complexity of trust evolution in HATs. These insights would be particularly valuable for designing AI teammates suited for long-term, high-stakes settings, such as healthcare and defence, where sustained trust is crucial.

SI in AI remains an underexplored area, particularly regarding how AI systems perceive and respond to subtle social cues from human teammates. Future studies should examine how AI can be equipped with context-sensitive SI, enabling it to

respond to human emotions, adapt its behaviour dynamically, and foster better alignment with team goals (Williams et al., 2022). For example, research could focus on developing AI systems that recognise user frustration or satisfaction and adjust their level of guidance or collaboration accordingly. Additionally, studies could investigate how AI systems that exhibit self-awareness about their limitations (e.g., expressing uncertainty when their confidence is low) affect user trust and reliance. This line of research would benefit from interdisciplinary approaches, drawing on insights from psychology, human-computer interaction, and machine learning to create socially intelligent AI that enhances team cohesion and effectiveness.

Finally, future research should aim to test the principles of HATs in real-world, dynamic environments to validate and expand upon the findings of this thesis. While controlled experiments offer valuable initial insights, field studies in operational settings, such as healthcare, emergency response, or remote teamwork, can reveal how AI teammates function under real-world pressures and unpredictability (Rix, 2022; Berretta et al., 2023). Field studies could assess how AI's adaptability, anthropomorphism, and SI impact team performance, situational awareness, and user trust in high-stakes environments. Such research could also examine how human teams adjust their behaviour and strategies based on AI actions, providing critical insights into the reciprocal dynamics of human-AI interaction. These studies would not only validate theoretical insights but also inform the design of AI systems that are robust and responsive in complex, dynamic team settings.

6.6 Conclusion

This thesis has undertaken an in-depth exploration of the evolving landscape of HATs, investigating the foundational elements necessary for AI to transition from functional tools to trusted teammates. By examining trust, anthropomorphism, and SI, this work contributes to a nuanced understanding of how AI can integrate meaningfully within human teams. Drawing on interdisciplinary insights, this research presents a comprehensive framework for human-AI collaboration,

underscoring the critical importance of calibrated trust, balanced anthropomorphism, and adaptable SI.

The findings of this thesis affirm that AI's effectiveness as a teammate relies on far more than technical capability alone. Trust is shown to be the cornerstone of successful HATs, requiring not only reliability but also dynamic responsiveness and transparency. By empirically examining how AI design elements, such as anthropomorphic language and nonverbal cues, impact trust and performance, this research provides actionable insights for developers, offering design guidelines that humanise AI without overstepping into discomfort or manipulation. These insights extend to ethical considerations, advocating for a principled approach to AI that fosters trust without undermining user autonomy or consent.

In addition to its theoretical and practical contributions, this thesis introduces novel methodological approaches, including a bibliometric analysis that maps trust research across disciplines, and experimental frameworks that quantify the effects of social cues on human-AI interactions. These methodologies strengthen the findings presented here and provide a replicable basis for future studies aiming to refine the roles of trust and SI in HATs. By setting a foundation in these areas, this work opens pathways for future research to examine trust dynamics over time, refine adaptive AI behaviour, and consider diverse cultural perspectives in HAT design.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*. https://doi.org/10.1109/ACCESS.2018.2870052

Agarwal, R., & Prasad, J. (1999). Are individual differences germane to the acceptance of new information technologies? *Decision Sciences*, *30*(2). https://doi.org/10.1111/j.1540-5915.1999.tb01614.x

Ahanin, Z., & Ismail, M. A. (2022). A multi-label emoji classification method using balanced pointwise mutual information-based feature selection. *Computer Speech and Language*, *73*. https://doi.org/10.1016/j.csl.2021.101330

Ahmad, M. I., Mubin, O., & Orlando, J. (2017). A systematic review of adaptivity in human-robot interaction. *Multimodal Technologies and Interaction*, *1*(3). https://doi.org/10.3390/mti1030014

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and Explanation in AI-Informed Decision Making. *Machine Learning and Knowledge Extraction*, *4*(2). https://doi.org/10.3390/make4020026

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4). https://doi.org/10.1016/j.joi.2017.08.007

Bailey, M. E., & Pollick, F. E. (2023). Social Intelligence towards Human-AI Teambuilding. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, *37*.

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*(1). https://doi.org/10.1007/s12369-010-0082-7

Bainbridge, W. S., Brent, E. E., Carley, K. M., Heise, D. R., Macy, M. W., Markovsky, B., & Skvoretz, J. (1994). Artificial Social Intelligence. *Annual Review of Sociology*, *20*(1), 407–436. https://doi.org/10.1146/annurev.so.20.080194.002203

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2021). Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, *13A*. https://doi.org/10.1609/aaai.v35i13.17359

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*. https://doi.org/10.1609/hcomp.v7i1.5285

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. https://doi.org/10.1609/aaai.v33i01.33012429

Bansal, G., Wu, T., & Zhou, J. (2021). Does the whole exceed its parts? The efect of ai explanations on complementary team performance. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445717

Barczak, G., Lassk, F., & Mulki, J. (2010). Antecedents of team creativity: An examination of team emotional intelligence, team trust and collaborative culture.

Creativity and Innovation Management, *19*(4). https://doi.org/10.1111/j.1467-8691.2010.00574.x

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00328

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. In *International Journal of Social Robotics* (Vol. 1, Issue 1). https://doi.org/10.1007/s12369-008-0001-3

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beattie, A., Edwards, A. P., & Edwards, C. (2020). A Bot and a Smile: Interpersonal Impressions of Chatbots and Humans Using Emoji in Computer-mediated Communication. *Communication Studies*, *71*(3). https://doi.org/10.1080/10510974.2020.1725082

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., & Kluge, A. (2023). Defining human-AI teaming the human-centered way: a scoping review and network analysis. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1250725

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? In *Quarterly Journal of Economics* (Vol. 119, Issue 1). https://doi.org/10.1162/003355304772839588

Birt, L., Scott, S., Cavers, D., Campbell, C., & Walter, F. (2016). Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? *Qualitative Health Research*, *26*(13). https://doi.org/10.1177/1049732316654870

Bittner, E. A. C., Oeste-Reiß, S., & Leimeister, J. M. (2019). Where is the bot in our team? Toward a taxonomy of design option combinations for conversational agents in collaborative work. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2019-January*. https://doi.org/10.24251/hicss.2019.035

Bornstein, B. H., Pytlik Zillig, L. M., & Shockley, E. (2016). Inspiring and advancing the many-disciplined study of institutional trust. In *Interdisciplinary Perspectives on Trust: Towards Theoretical and Methodological Integration*. https://doi.org/10.1007/978-3-319-22261-5_1

Botsman, R. (2015). The Changing Rules of Trust in the Digital Age. *Harvard Business Review*.

Bottery, M. (2003). The Management and Mismanagement of Trust. *Educational Management & Administration*, *31*(3). https://doi.org/10.1177/0263211x03031003003

Boutet, I., LeBlanc, M., Chamberland, J. A., & Collin, C. A. (2021). Emojis influence emotional communication, social attributions, and information processing. *Computers in Human Behavior*, *119*. https://doi.org/10.1016/j.chb.2021.106722

Boyatzis, R. E., Thiel, K., Rochford, K., & Black, A. (2017). Emotional and Social Intelligence Competencies of Incident Team Commanders Fighting Wildfires. *Journal of Applied Behavioral Science*, *53*(4). https://doi.org/10.1177/0021886317731575

Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European Management Journal*, *40*(5). https://doi.org/10.1016/j.emj.2022.06.001

Cannizzaro, S., Procter, R., Ma, S., & Maple, C. (2020). Trust in the smart home: Findings from a nationally representative survey in the UK. *PLoS ONE*, *15*(5). https://doi.org/10.1371/journal.pone.0231615

Casper, W. J., Vaziri, H., Wayne, J. H., DeHauw, S., & Greenhaus, J. (2017). The jingle-jangle of work-nonwork balance: A comprehensive and meta-analytic review of its meaning and measurement. *Journal of Applied Psychology*, *103*(2). https://doi.org/10.1037/apl0000259

Cavalcante, W. Q. de F., Coelho, A., & Bairrada, C. M. (2021). Sustainability and tourism marketing: A bibliometric analysis of publications between 1997 and 2020 using vosviewer software. *Sustainability (Switzerland)*, *13*(9). https://doi.org/10.3390/su13094987

Chai, D. S., Hwang, S. J., & Joo, B. K. (2017). Transformational Leadership and Organizational Commitment in Teams: The Mediating Roles of Shared Vision and Team-Goal Commitment. *Performance Improvement Quarterly*, *30*(2). https://doi.org/10.1002/piq.21244

Chao, G. T., O'Leary-Kelly, A. M., Wolf, S., Klein, H. J., & Gardner, P. D. (1994). Organizational Socialization: Its Content and Consequences. *Journal of Applied Psychology*, *79*(5). https://doi.org/10.1037/0021-9010.79.5.730

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6). https://doi.org/10.1037/0022-3514.76.6.893

Chaudhuri, A., & Holbrook, M. B. (2001). The chain of effects from brand trust and brand affect to brand performance: The role of brand loyalty. *Journal of Marketing*, *65*(2). https://doi.org/10.1509/jmkg.65.2.81.18255

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied Ergonomics*, *52*. https://doi.org/10.1016/j.apergo.2015.07.012

Chen, Q. Q., & Park, H. J. (2021). How anthropomorphism affects trust in intelligent personal assistants. *Industrial Management and Data Systems*, *121*(12). https://doi.org/10.1108/IMDS-12-2020-0761

Chen, S., Zhang, Y., Dai, W., Qi, S., Tian, W., Gu, X., Chen, X., Yu, W., Tian, J., & Su, D. (2020). Publication trends and hot spots in postoperative cognitive dysfunction research: A 20-year bibliometric analysis. In *Journal of Clinical Anesthesia* (Vol. 67). https://doi.org/10.1016/j.jclinane.2020.110012

Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023). Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2). https://doi.org/10.1145/3610219

Chiocchio, F., & Essiembre, H. (2009). Cohesion and performance: A meta-analytic review of disparities between project teams, production teams, and service teams. *Small Group Research*, *40*(4). https://doi.org/10.1177/1046496409335103

Chiou, E. K., Lee, J. D., & Su, T. (2019). Negotiated and reciprocal exchange structures in human-agent cooperation. *Computers in Human Behavior*, *90*. https://doi.org/10.1016/j.chb.2018.08.012

Costa, A. C., Fulmer, C. A., & Anderson, N. R. (2018). Trust in work teams: An integrative review, multilevel model, and future directions. In *Journal of Organizational Behavior* (Vol. 39, Issue 2). https://doi.org/10.1002/job.2213

Costantini, S., De Gasperis, G., & Olivieri, R. (2019). Digital forensics and investigations meet artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, *86*(1–3). https://doi.org/10.1007/s10472-019-09632-y

Dacey, M., & Coane, J. H. (2023). Implicit measures of anthropomorphism: affective priming and recognition of apparent animal emotions. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1149444

Dalton, A., Wolff, K., & Bekker, B. (2021). Multidisciplinary Research as a Complex System. *International Journal of Qualitative Methods*, *20*. https://doi.org/10.1177/16094069211038400

Dang, J., King, K. M., & Inzlicht, M. (2020). Why Are Self-Report and Behavioral Measures Weakly Correlated? In *Trends in Cognitive Sciences* (Vol. 24, Issue 4). https://doi.org/10.1016/j.tics.2020.01.007

Dautenhahn, K. (1995). Getting to know each other-Artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, *16*(2–4). https://doi.org/10.1016/0921-8890(95)00054-2

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3). https://doi.org/10.1037/xap0000092

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2). https://doi.org/10.1007/s12369-019-00596-x

Derrick, D. C., & Elson, J. S. (2019). Exploring automated leadership and agent interaction modalities. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2019-January*. https://doi.org/10.24251/hicss.2019.027

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., & Yanco, H. (2012). Effects of changing reliability on trust of robot systems. *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/2157689.2157702

Dimoka, A. (2010). What does the brain tell us about trust and distrust? evidence from a functional neuroimaging study. *MIS Quarterly: Management Information Systems*, *34*(SPEC. ISSUE 2). https://doi.org/10.2307/20721433

Dodgson, M. (1993). Learning, Trust, and Technological Collaboration. *Human Relations*, *46*(1). https://doi.org/10.1177/001872679304600106

Doney, P. M., & Cannon, J. P. (1997). An examination of the nature of trust in buyer-seller relationships. *Journal of Marketing*, *61*(2). https://doi.org/10.2307/1251829

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*. https://doi.org/10.1016/j.jbusres.2021.04.070

Döppner, D. A., Derckx, P., & Schoder, D. (2019). Symbiotic co-evolution in collaborative human-machine decision making: Exploration of a multi-year design science research project in the air cargo industry. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2019-January*. https://doi.org/10.24251/hicss.2019.033

Duffy, A. (2015). Friends and fellow travelers: Comparative influence of review sites and friends on hotel choice. *Journal of Hospitality and Tourism Technology*, *6*(2). https://doi.org/10.1108/JHTT-05-2014-0015

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3–4). https://doi.org/10.1016/S0921-8890(02)00374-3

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative Content Analysis: A Focus on Trustworthiness. *SAGE Open*, *4*(1). https://doi.org/10.1177/2158244014522633

Endsley, M. R. (2023). Supporting Human-AI Teams:Transparency, explainability, and situation awareness. *Computers in Human Behavior*, *140*. https://doi.org/10.1016/j.chb.2022.107574

Fadhil, A., Schiavo, G., Wang, Y., & Yilma, B. A. (2018). The effect of emojis when interacting with conversational interface assisted health coaching system. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3240925.3240965

Fang, R., Duffy, M. K., & Shaw, J. D. (2011). The organizational socialization process: Review and development of a social capital model. In *Journal of Management* (Vol. 37, Issue 1). https://doi.org/10.1177/0149206310384630

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data and Society*, *6*(1). https://doi.org/10.1177/2053951719860542

Flathmann, C., Schelble, B. G., Rosopa, P. J., McNeese, N. J., Mallick, R., & Madathil, K. C. (2023). Examining the impact of varying levels of AI teammate influence on human-AI teams. *International Journal of Human Computer Studies*, *177*. https://doi.org/10.1016/j.ijhcs.2023.103061

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020). Fostering human agency: A process for the design of user-centric xai systems. *International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global*.

Freeman, F. S., & Kelley, T. L. (1928). Interpretation of Educational Measurements. *The American Journal of Psychology*, *40*(3). https://doi.org/10.2307/1414476

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, *33*(2). https://doi.org/10.1287/isre.2021.1079

Gallery, E., & Mitchell, C. J. (2009). Trusted computing: Security and applications. *Cryptologia*, *33*(3). https://doi.org/10.1080/01611190802231140

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, *1*(1). https://doi.org/10.30658/hmc.1.5

Garbarino, E., & Johnson, M. S. (1999). The different roles of satisfaction, trust, and commitment in customer relationships. *Journal of Marketing*, *63*(2). https://doi.org/10.2307/1251946

Ghosh, R., Shuck, B., & Petrosko, J. (2012). Emotional intelligence and organizational learning in work teams. *Journal of Management Development*, *31*(6). https://doi.org/10.1108/02621711211230894

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, *59*(2). https://doi.org/10.1037/0003-066X.59.2.93

Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, *24*(2). https://doi.org/10.1016/j.nedt.2003.10.001

Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems*, *144*. https://doi.org/10.1016/j.dss.2021.113515

Grimmelikhuijsen, S., & Knies, E. (2017). Validating a scale for citizen trust in government organizations. *International Review of Administrative Sciences*, *83*(3). https://doi.org/10.1177/0020852315585950

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Gulati, R. (1995). Does Familiarity Breed Trust? The Implications of Repeated Ties for Contractual Choice in Alliances. *Academy of Management Journal*, *38*(1). https://doi.org/10.5465/256729

Gulati, S., Sousa, S., & Lamas, D. (2017). Modelling trust: An empirical assessment. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10516 LNCS*. https://doi.org/10.1007/978-3-319-68059-0_3

Gulati, S., Sousa, S., & Lamas, D. (2018). Modelling trust in human-like technologies. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3297121.3297124

Hackman, J. R. (1987). Group-level issues in the design and training of cockpit crews. *NASA. Ames Research Center Cockpit Resource Management Training*.

Hamidah, I., Sriyono, & Hudha, M. N. (2020). A bibliometric analysis of COVID-19 research using vosviewer. *Indonesian Journal of Science and Technology*, *5*(2). https://doi.org/10.17509/ijost.v5i2.24522

Hamza, A. (2016). Are Emojis Creating a New or Old Visual Language for New Generations? A Socio-semiotic Study. *Advances in Language and Literary Studies*, *7*(6). https://doi.org/10.7575/aiac.alls.v.7n.6p.56

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5). https://doi.org/10.1177/0018720811417254

Hariri, N., Mobasher, B., & Burke, R. (2015). Adapting to user preference changes in interactive recommendation. *IJCAI International Joint Conference on Artificial Intelligence*, *2015-January*.

Hauptman, A. I., Schelble, B. G., McNeese, N. J., & Madathil, K. C. (2023). Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. *Computers in Human Behavior*, *138*. https://doi.org/10.1016/j.chb.2022.107451

Hendriks, F., Distel, B., Engelke, K. M., Westmattelmann, D., & Wintterlin, F. (2021). Methodological and Practical Challenges of Interdisciplinary Trust Research. In *Trust and Communication*. https://doi.org/10.1007/978-3-030-72945-5_2

Henrique, B. M., & Santos, E. (2024). Trust in artificial intelligence: Literature review and main path analysis. *Computers in Human Behavior: Artificial Humans*, *2*(1). https://doi.org/10.1016/j.chbah.2024.100043

Hetherington, M. J. (1998). The Political Relevance of Political Trust. *American Political Science Review*, *92*(4). https://doi.org/10.2307/2586304

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. In *Journal of Strategic Information Systems* (Vol. 32, Issue 3). https://doi.org/10.1016/j.jsis.2023.101772

Heyselaar, E. (2023). The CASA theory no longer applies to desktop computers. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-46527-9

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3). https://doi.org/10.1177/0018720814547570

Honig, S., & Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. In *Frontiers in Psychology* (Vol. 9, Issue JUN). https://doi.org/10.3389/fpsyg.2018.00861

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9). https://doi.org/10.1177/1049732305276687

Hu, Y., Wang, D., Pang, K., Xu, G., & Guo, J. (2015). The effect of emotion and time pressure on risk decision-making. *Journal of Risk Research*, *18*(5). https://doi.org/10.1080/13669877.2014.910688

Ingram, M., Moreton, R., Gancz, B., & Pollick, F. (2021). Calibrating trust toward an autonomous image classifier. *Technology, Mind, and Behavior*, *2*(1). https://doi.org/10.1037/tmb0000032

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445923

Janhunen, E., Toivikko, T., Blomqvist, K., & Siemon, D. (2024, August 16). Trust in Digital Human-AI Team Collaboration: A Systematic Review. *AMCIS 2024 Proceedings*.

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, *1*(4). https://doi.org/10.1016/j.joi.2007.07.004

Jehn, K. A. (1995). A Multimethod Examination of the Benefits and Detriments of Intragroup Conflict. *Administrative Science Quarterly*, *40*(2). https://doi.org/10.2307/2393638

Jensen, T., Khan, M. M. H., Fahim, M. A. Al, & Albayram, Y. (2021). Trust and Anthropomorphism in Tandem: The Interrelated Nature of Automated Agent Appearance and Reliability in Trustworthiness Perceptions. *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere*. https://doi.org/10.1145/3461778.3462102

Jing, Y., Cai, H., Bond, M. H., Li, Y., Stivers, A. W., & Tan, Q. (2020). Levels of interpersonal trust across different types of environment: The micro–macro interplay between relational distance and human ecology. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000997

Johnson, M., Bradshaw, J. M., Feltovich, P., Jonker, C., Van Riemsdijk, B., & Sierhuis, M. (2012). Autonomy and interdependence in human-agent-robot teams. *IEEE Intelligent Systems*, *27*(2). https://doi.org/10.1109/MIS.2012.1

Jones-Jang, S. M., & Park, Y. J. (2023). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, *28*(1). https://doi.org/10.1093/jcmc/zmac029

Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems*, *43*(2). https://doi.org/10.1016/j.dss.2005.05.019

Jung, E. S., Dong, S. Y., & Lee, S. Y. (2019). Neural Correlates of Variations in Human Trust in Human-like Machines during Non-reciprocal Interactions. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-46098-8

Kamar, E. (2016). Directions in hybrid intelligence: Complementing AI systems with human intelligence. *IJCAI International Joint Conference on Artificial Intelligence*, *2016-January*.

Kaur, S., & Sharma, R. (2021). Emotion AI: Integrating Emotional Intelligence with Artificial Intelligence in the Digital Workplace. In *Advances in Science, Technology and Innovation*. https://doi.org/10.1007/978-3-030-66218-9_39

Kelly, J. R., & Karau, S. J. (1999). Group decision making: The effects of initial preferences and time pressure. *Personality and Social Psychology Bulletin*, *25*(11). https://doi.org/10.1177/0146167299259002

Kerrigan, G., Smyth, P., & Steyvers, M. (2021). Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration. *Advances in Neural Information Processing Systems*, *6*.

Khan, W. Z., Aalsalem, M. Y., Khan, M. K., & Arshad, Q. (2019). Data and Privacy: Getting Consumers to Trust Products Enabled by the Internet of Things. *IEEE Consumer Electronics Magazine*, *8*(2). https://doi.org/10.1109/MCE.2018.2880807

Kihlstrom, J. F., & Cantor, N. (2000). Social intelligence. In *Handbook of intelligence.* (pp. 359–379). Cambridge University Press. https://doi.org/10.1017/CBO9780511807947.017

Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems*, *44*(2). https://doi.org/10.1016/j.dss.2007.07.001

Kim, D., Song, Y., Kim, S., Lee, S., Wu, Y., Shin, J., & Lee, D. (2023). How should the results of artificial intelligence be explained to users? - Research on consumer preferences in user-centered explainable artificial intelligence. *Technological Forecasting and Social Change*, *188*. https://doi.org/10.1016/j.techfore.2023.122343

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, *61*. https://doi.org/10.1016/j.tele.2021.101595

Klarner, P., Sarstedt, M., Hoeck, M., & Ringle, C. M. (2013). Disentangling the effects of team competences, team adaptability, and client communication on the performance of management consulting teams. *Long Range Planning*, *46*(3). https://doi.org/10.1016/j.lrp.2013.03.001

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. *Advances in Intelligent Systems and Computing*, *823*. https://doi.org/10.1007/978-3-319-96074-6_2

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, *435*(7042). https://doi.org/10.1038/nature03701

Kox, E. S., Siegling, L. B., & Kerstholt, J. H. (2022). Trust Development in Military and Civilian Human–Agent Teams: The Effect of Social-Cognitive Recovery Strategies. *International Journal of Social Robotics*, *14*(5). https://doi.org/10.1007/s12369-022-00871-4

Kozlowski, S. W. J., & Ilgen, D. R. (2006). The Science of Team Effectiveness Enhancing the Effectiveness of Work Groups and Teams. *Psychological Science in the Public Interest*, *7*(3).

Krop, P., Koch, M. J., Carolus, A., Latoschik, M. E., & Wienrich, C. (2024). The Effects of Expertise, Humanness, and Congruence on Perceived Trust, Warmth, Competence and Intention to Use Embodied AI. *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613905.3650749

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., Heinecke, A., & Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(50), 20084–20089. https://doi.org/10.1073/pnas.0710103104

Kuc-Czarnecka, M., & Olczyk, M. (2020). How ethics combine with big data: a bibliometric analysis. *Humanities and Social Sciences Communications*, *7*(1). https://doi.org/10.1057/s41599-020-00638-0

Kulms, P., & Kopp, S. (2019). More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human–agent cooperation. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3340764.3340793

Kunar, M. A., & Watson, D. G. (2023). Framing the fallibility of Computer-Aided Detection aids cancer detection. *Cognitive Research: Principles and Implications*, *8*(1). https://doi.org/10.1186/s41235-023-00485-y

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/JSS.V082.I13

Lachman, S. J., & Bass, A. R. (1985). A direct study of halo effect. *Journal of Psychology: Interdisciplinary and Applied*, *119*(6). https://doi.org/10.1080/00223980.1985.9915460

Lahusen, C., Maggetti, M., & Slavkovik, M. (2024). Trust, trustworthiness and AI governance. *Scientific Reports*, *14*(1), 20752. https://doi.org/10.1038/s41598-024-71761-0

Lang, D. J., Wiek, A., Bergmann, M., Stauffacher, M., Martens, P., Moll, P., Swilling, M., & Thomas, C. J. (2012). Transdisciplinary research in sustainability science: Practice, principles, and challenges. *Sustainability Science*, *7*(SUPPL. 1). https://doi.org/10.1007/s11625-011-0149-x

Larsen, K. R., & Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Quarterly: Management Information Systems*, *40*(3). https://doi.org/10.25300/MISQ/2016/40.3.01

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, *84*(3). https://doi.org/10.1007/s11192-010-0202-z

Lawton, T., Grace, K., & Francisco, J. I. (2023). When is a Tool a Tool? User Perceptions of System Agency in Human–AI Co-Creative Drawing. *DIS '23: Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 1978–1996.

Lee, C., & Wong, C. S. (2019). The effect of team emotional intelligence on team process and effectiveness. *Journal of Management and Organization*, *25*(6). https://doi.org/10.1017/jmo.2017.43

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. In *Human Factors* (Vol. 46, Issue 1). https://doi.org/10.1518/hfes.46.1.50_30392

Lee, S., Li, M., Lai, B., Jia, W., Ryan, F., Cao, X., Kara, O., Boote, B., Shi, W., Yang, D., & Rehg, J. M. (2024). *Towards Social AI: A Survey on Understanding Social Interactions*.

Lenth, R. V. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R Package Version 1.10.2.090002.
Leong, Y. R., Tajudeen, F. P., & Yeong, W. C. (2021). Bibliometric and content analysis of the internet of things research: a social science perspective. In *Online Information Review* (Vol. 45, Issue 6). https://doi.org/10.1108/OIR-08-2020-0358
Lewicki, R., & Bunker, B. (1996). Developing and Maintaining Trust in Working Relations. In *Trust in organizations: Frontiers of theory and research* (pp. 114–139). https://doi.org/10.4135/9781452243610.n7
Lewicki, R. J., & Brinsfield, C. (2017). Trust Repair. In *Annual Review of Organizational Psychology and Organizational Behavior* (Vol. 4). https://doi.org/10.1146/annurev-orgpsych-032516-113147
Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From Principles to Practices. In *ACM Computing Surveys* (Vol. 55, Issue 9). https://doi.org/10.1145/3555803
Li, X., & Sung, Y. (2021). Anthropomorphism brings us closer: The mediating role of psychological distance in User–AI assistant interactions. *Computers in Human Behavior*, *118*. https://doi.org/10.1016/j.chb.2021.106680
Liang, C., Proft, J., Andersen, E., & Knepper, R. A. (2019). Implicit Communication of Actionable Information in Human-AI teams. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300325
Liang, H., & Shi, X. (2022). Exploring the structure and emerging trends of construction health management: a bibliometric review and content analysis. *Engineering, Construction and Architectural Management*, *29*(4). https://doi.org/10.1108/ECAM-01-2021-0080
Lins, K. V., Servaes, H., & Tamayo, A. (2017). Social Capital, Trust, and Firm Performance: The Value of Corporate Social Responsibility during the Financial Crisis. *Journal of Finance*, *72*(4). https://doi.org/10.1111/jofi.12505
Liu, J., Wong, C. K., & Hui, K. K. (2003). An Adaptive User Interface Based on Personalized Learning. *IEEE Intelligent Systems*, *18*(2). https://doi.org/10.1109/MIS.2003.1193657
Lopez, J., Textor, C., Lancaster, C., Schelble, B., Freeman, G., Zhang, R., McNeese, N., & Pak, R. (2023). The complex relationship of AI ethics and trust in human–AI teaming: insights from advanced real-world subject matter experts. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00303-7
Lu, B., Zeng, Q., & Fan, W. (2016). Examining macro-sources of institution-based trust in social commerce marketplaces: An empirical study. *Electronic Commerce Research and Applications*, *20*. https://doi.org/10.1016/j.elerap.2016.10.004
Lyons, J. B., Sycara, K., Lewis, M., & Capiola, A. (2021). Human–Autonomy Teaming: Definitions, Debates, and Directions. In *Frontiers in Psychology* (Vol. 12). https://doi.org/10.3389/fpsyg.2021.589585
Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3544548.3581058
Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology*, *111*(2). https://doi.org/10.1037/edu0000281
Mathur, L., Liang, P. P., & Morency, L.-P. (2024). Advancing Social Intelligence in AI Agents: Technical Challenges and Open Questions. In Y. Al-Onaizan, M. Bansal, &

Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 20541–20560). Association for Computational Linguistics.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Academy of Management Review*, *20*(3). https://doi.org/10.5465/amr.1995.9508080335

McAllister, D. J. (1995). Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal*, *38*(1). https://doi.org/10.5465/256727

McAllister, D. J., Lewicki, R. J., & Chaturvedi, S. (2006). Trust in developing relationships: From theory to measurement. *Academy of Management 2006 Annual Meeting: Knowledge, Action and the Public Concern, AOM 2006*. https://doi.org/10.5465/ambpp.2006.22897235

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2). https://doi.org/10.1145/1985347.1985353

McKnight, D. H., & Chervany, N. L. (2001). Conceptualizing trust: A typology and e-commerce customer relationships model. *Proceedings of the Hawaii International Conference on System Sciences*. https://doi.org/10.1109/HICSS.2001.927053

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors*, *60*(2). https://doi.org/10.1177/0018720817743223

McNeese, N. J., Schelble, B. G., Canonico, L. B., & Demir, M. (2021). Who/What Is My Teammate? Team Composition Considerations in Human-AI Teaming. *IEEE Transactions on Human-Machine Systems*, *51*(4). https://doi.org/10.1109/THMS.2021.3086018

Mead, J., Fisher, Z., & Kemp, A. H. (2021). Moving Beyond Disciplinary Silos Towards a Transdisciplinary Model of Wellbeing: An Invited Review. In *Frontiers in Psychology* (Vol. 12). https://doi.org/10.3389/fpsyg.2021.642093

Mehta, S., North, C., & Luther, K. (2016). An exploratory study of human performance in image geolocation tasks. *HCOMP 2016 GroupSight Workshop on Human Computation for Image and Video Analysis*.

Merritt, S. M., Heimbaugh, H., Lachapell, J., & Lee, D. (2013). I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3). https://doi.org/10.1177/0018720812465081

Merritt, T., & McGee, K. (2012). Protecting artificial team-mates: More seems like less. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/2207676.2208680

Merritt, T. R., Tan, K. B., Ong, C., Thomas, A., Chuah, T. L., & McGee, K. (2011). Are artificial team-mates scapegoats in computer games. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. https://doi.org/10.1145/1958824.1958945

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers Robotics AI*, *4*(MAY). https://doi.org/10.3389/frobt.2017.00021

Montag, C., Becker, B., & Li, B. J. (2024). On trust in humans and trust in artificial intelligence: A study with samples from Singapore and Germany extending recent

research. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100070. https://doi.org/https://doi.org/10.1016/j.chbah.2024.100070
Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., Chen, Y., Haas, B. W., & Becker, B. (2023). Trust toward humans and trust toward artificial intelligence are not associated: Initial insights from self-report and neurostructural brain imaging. *Personality Neuroscience*, *6*. https://doi.org/10.1017/pen.2022.5
Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. In *Profesional de la Informacion* (Vol. 29, Issue 1). https://doi.org/10.3145/epi.2020.ene.03
Morgan, R. M., & Hunt, S. D. (1994). The Commitment-Trust Theory of Relationship Marketing. *Journal of Marketing*, *58*(3). https://doi.org/10.1177/002224299405800302
Mori, M. (2012). The Uncanny Valley: The Original Essay by Masahiro Mori. *IEEE Robotics & Automation Magazine*, *12*(Figure 1).
Mori, M., MacDorman, K., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, *19*(2). https://doi.org/10.1109/mra.2012.2192811
Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11). https://doi.org/10.1080/00140139408964957
Musick, G., O'Neill, T., Schelble, B., McNeese, N., & Henke, J. (2021). Human-Autonomy Teaming: What Happens When Humans Believe Their Teammate is an AI? *Computers in Human Behavior*.
Naquin, C. E., & Tynan, R. O. (2003). The team halo effect: Why teams are not blamed for their failures. *Journal of Applied Psychology*, *88*(2), 332–340. https://doi.org/10.1037/0021-9010.88.2.332
Nasir, S., Khan, R. A., & Bai, S. (2024). Ethical Framework for Harnessing the Power of AI in Healthcare and Beyond. *IEEE Access*, *12*. https://doi.org/10.1109/ACCESS.2024.3369912
Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human Computer Studies*, *45*(6). https://doi.org/10.1006/ijhc.1996.0073
Nass, C., Steuer, J., & Tauber, E. R. (1994a). Computer are social actors. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/259963.260288
Nass, C., Steuer, J., & Tauber, E. R. (1994b). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78.
Nicolau, J. L., Mellinas, J. P., & Martín-Fuentes, E. (2020). The halo effect: A longitudinal approach. *Annals of Tourism Research*, *83*. https://doi.org/10.1016/j.annals.2020.102938
Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, *16*(1). https://doi.org/10.1177/1609406917733847
Oh, C., Song, J., Choi, J., Kim, S., Lee, S., & Suh, B. (2018). I lead, you help but only with enough details: Understanding the user experience of co-creation with artificial intelligence. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*. https://doi.org/10.1145/3173574.3174223
Ong, C., McGee, K., & Chuah, T. L. (2012). Closing the human-AI team-mate gap: How changes to displayed information impact player behavior towards computer

teammates. *Proceedings of the 24th Australian Computer-Human Interaction Conference, OzCHI 2012*. https://doi.org/10.1145/2414536.2414604
OpenAI. (2024). *ChatGPT (3.5)*.
Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce*, *7*(3). https://doi.org/10.1080/10864415.2003.11044275
Peeters, M. M. M., van Diggelen, J., van den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human–AI society. *AI and Society*, *36*(1). https://doi.org/10.1007/s00146-020-01005-y
Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1). https://doi.org/10.3758/s13428-018-01193-y
Pelau, C., Dabija, D. C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, *122*. https://doi.org/10.1016/j.chb.2021.106855
Pinski, M., Adam, M., & Benlian, A. (2023). AI Knowledge: Improving AI Delegation through Human Enablement. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3544548.3580794
Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI and Ethics*, *4*(3). https://doi.org/10.1007/s43681-024-00419-4
Poondy Rajan, Y., Aiswarya, B., & Jenifer Arokia Selvi, A. (2023). Embracing emojis: Bridging the gap in workplace technology adoption and elevating communication effectiveness. *Journal of Information Technology Teaching Cases*. https://doi.org/10.1177/20438869231220676
Poortinga, W., & Pidgeon, N. F. (2004). Trust, the asymmetry principle, and the role of prior beliefs. *Risk Analysis*, *24*(6). https://doi.org/10.1111/j.0272-4332.2004.00543.x
Pulakos, E. D., Dorsey, D. W., & White, S. S. (2006). Adaptability in the Workplace: Selecting an Adaptive Workforce. In *Advances in Human Performance and Cognitive Engineering Research* (Vol. 6). https://doi.org/10.1016/S1479-3601(05)06002-9
Reeves, B., Hancock, J., & Liu, X. (2020). Social Robots Are Like Real People: First Impressions, Attributes, and Stereotyping of Social Robots. *Technology, Mind, and Behavior*, *1*(1). https://doi.org/10.1037/tmb0000018
Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. *AI and Ethics*, *3*(3). https://doi.org/10.1007/s43681-022-00200-5
Revelle, W. (2016). Procedures for Personality and Psychological Research, NorthwesternUniversity, Evanston, Illinois, USA. *R Package Published through CRAN*, *1.6.12*.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*. https://doi.org/10.1145/2939672.2939778
Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2005). The mechanics of trust: A framework for research and design. *International Journal of Human Computer Studies*, *62*(3). https://doi.org/10.1016/j.ijhcs.2005.01.001

Rix, J. (2022). From Tools to Teammates: Conceptualizing Humans' Perception of Machines as Teammates with a Systematic Literature Review. *Proceedings of the 55th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2022.048

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*. https://doi.org/10.1109/HRI.2016.7451740

Rodgers, W., Murray, J. M., Stefanidis, A., Degbey, W. Y., & Tarba, S. Y. (2023). An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. *Human Resource Management Review*, *33*(1). https://doi.org/10.1016/j.hrmr.2022.100925

Rompf, S. A. (2015). Trust and rationality: An integrative framework for trust research. In *Trust and Rationality: An Integrative Framework for Trust Research*. https://doi.org/10.1007/978-3-658-07327-5

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, *35*(1). https://doi.org/10.1037/0003-066X.35.1.1

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. In *Academy of Management Review* (Vol. 23, Issue 3). https://doi.org/10.5465/AMR.1998.926617

Roy, R., & Naidoo, V. (2021). Enhancing chatbot effectiveness: The role of anthropomorphic conversational styles and time orientation. *Journal of Business Research*, *126*. https://doi.org/10.1016/j.jbusres.2020.12.051

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, *26*(5). https://doi.org/10.1007/s11948-020-00228-y

Salimzadeh, S., He, G., & Gadiraju, U. (2023). A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. *UMAP 2023 - Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. https://doi.org/10.1145/3565472.3592959

Salovey, P., & Mayer, J. D. (1990). Emotional Intelligence. *Imagination, Cognition and Personality*, *9*(3), 185–211. https://doi.org/10.2190/DUGG-P24E-52WK-6CDG

Sanneman, L., & Shah, J. A. (2022). The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. *International Journal of Human-Computer Interaction*, *38*(18–20). https://doi.org/10.1080/10447318.2022.2081282

Satyanarayana, V., Shankar, S., Sruthi, V., & Das, B. (2018). A Study of Artificial Social Intelligence in Conversational Agents. *Proceedings of the 3rd International Conference on Inventive Computation Technologies, ICICT 2018*. https://doi.org/10.1109/ICICT43934.2018.9034313

Schelble, B. G., Flathmann, C., McNeese, N. J., Freeman, G., & Mallick, R. (2022). Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams. *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP). https://doi.org/10.1145/3492832

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2024). Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming. *Human Factors*, *66*(4). https://doi.org/10.1177/00187208221116952

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4). https://doi.org/10.1080/12460125.2020.1819094

Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology*, *78*. https://doi.org/10.1016/j.joep.2020.102253

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. In *Academy of Management Review* (Vol. 32, Issue 2). https://doi.org/10.5465/AMR.2007.24348410

Sedighi, M. (2016). Application of word co-occurrence analysis method in mapping of the scientific fields (case study: the field of Informetrics). *Library Review*, *65*(1–2). https://doi.org/10.1108/LR-07-2015-0075

Seymour, W., & Van Kleek, M. (2021). Exploring Interactions between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2). https://doi.org/10.1145/3479515

Shang, R., Hsieh, G., & Shah, C. (2024, July 25). Trusting Your AI Agent Emotionally and Cognitively: Development and Validation of a Semantic Differential Scale for AI Trust. *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES2024)*. https://doi.org/10.48550/arXiv.2408.05354

Sharma, K., Schoorman, F. D., & Ballinger, G. A. (2023). How Can It Be Made Right Again? A Review of Trust Repair Research. In *Journal of Management* (Vol. 49, Issue 1). https://doi.org/10.1177/01492063221089897

Shen, Z., Li, L., Yan, F., & Wu, X. (2010). Cloud computing system based on trusted computing platform. *2010 International Conference on Intelligent Computation Technology and Automation, ICICTA 2010*, *1*. https://doi.org/10.1109/ICICTA.2010.724

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*. https://doi.org/10.1016/j.chb.2019.04.019

Shockley-Zalabak, P., Ellis, K., & Winograd, G. (2000). Organizational trust: What it means, why it matters. *Organization Development Journal*, *18*(4).

Sicari, S., Rizzardi, A., Grieco, L. A., & Coen-Porisini, A. (2015). Security, privacy and trust in Internet of things: The road ahead. In *Computer Networks* (Vol. 76). https://doi.org/10.1016/j.comnet.2014.11.008

Siemon, D. (2022). Elaborating Team Roles for Artificial Intelligence-based Teammates in Human-AI Collaboration. *Group Decision and Negotiation*. https://doi.org/10.1007/s10726-022-09792-z

Song, S. W., & Shin, M. (2024). Uncanny Valley Effects on Chatbot Trust, Purchase Intention, and Adoption Intention in the Context of E-Commerce: The Moderating Role of Avatar Familiarity. *International Journal of Human-Computer Interaction*, *40*(2). https://doi.org/10.1080/10447318.2022.2121038

Srikanth, P., Kumar, A., & Hedabou, M. (2022). An Uncertainty Trust Assessment Scheme for Trustworthy Partner Selection in Online Games. *IEEE Access*, *10*. https://doi.org/10.1109/ACCESS.2022.3230148

Sterelny, K. (2007). Social intelligence, human intelligence and niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480). https://doi.org/10.1098/rstb.2006.2006

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(11). https://doi.org/10.1073/PNAS.2111547119

Strauss, A. L. (2017). Mirrors and masks: The search for identity. In *Mirrors and Masks: The Search for Identity*. https://doi.org/10.4324/9781315124582

Sufyan, N. S., Fadhel, F. H., Alkhathami, S. S., & Mukhadi, J. Y. A. (2024). Artificial intelligence and social intelligence: preliminary comparison study between AI models and psychologists. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1353022

Sung, Y. T., & Wu, J. S. (2018). The Visual Analogue Scale for Rating, Ranking and Paired-Comparison (VAS-RRP): A new technique for psychological measurement. *Behavior Research Methods*, *50*(4). https://doi.org/10.3758/s13428-018-1041-8

Syrdal, D. S., Dautenhahn, K., Woods, S., Walters, M. L., & Koay, K. L. (2006). "Doing the right thing wrong" - Personality and tolerance to uncomfortable robot approaches. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*. https://doi.org/10.1109/ROMAN.2006.314415

Tanevska, A., Rea, F., Sandini, G., Cañamero, L., & Sciutti, A. (2020). A Socially Adaptable Framework for Human-Robot Interaction. *Frontiers in Robotics and AI*, *7*. https://doi.org/10.3389/frobt.2020.00121

Tang, J., Gao, H., Liu, H., & Das Sarmas, A. (2012). eTrust: Understanding trust evolution in an online world. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2339530.2339574

Terziev, V., & Stoyanov, E. (2018). (Conceptual Framework for Social Adaptation). *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3162839

Thorndike, E. L. (1920). Intelligence and its uses. *Harper's Magazine*, *140*, 227–235.

Troshani, I., Rao Hill, S., Sherman, C., & Arthur, D. (2021). Do We Trust in AI? Role of Anthropomorphism and Intelligence. *Journal of Computer Information Systems*, *61*(5). https://doi.org/10.1080/08874417.2020.1788473

Trösterer, S., Meschtscherjakov, A., Mirnig, A. G., Lupp, A., Gärtner, M., McGee, F., McCall, R., Tscheligi, M., & Engel, T. (2017). What we can learn from pilots for handovers and (de)skilling in semi-autonomous driving: An interview study. *AutomotiveUI 2017 - 9th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, Proceedings*. https://doi.org/10.1145/3122986.3123020

Ulfert, A. S., Georganta, E., Centeio Jorge, C., Mehrotra, S., & Tielman, M. (2023). Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework. *European Journal of Work and Organizational Psychology*. https://doi.org/10.1080/1359432X.2023.2200172

Uslaner, E. M. (2008). The foundations of trust: Macro and micro. *Cambridge Journal of Economics*, *32*(2). https://doi.org/10.1093/cje/bem039

Vakkari, P. (2008). Perceived influence of the use of electronic information resources on scholarly work and publication productivity. *Journal of the American Society for Information Science and Technology*, *59*(4). https://doi.org/10.1002/asi.20769

Van Eck, N. J., & Waltman, L. (2007). VOS: A new method for visualizing similarities between objects. *Studies in Classification, Data Analysis, and Knowledge Organization*. https://doi.org/10.1007/978-3-540-70981-7_34

Van Maanen, J. E., & Schein, E. H. (1977). *Toward a theory of organizational socialization*.

van Pinxteren, M. M. E., Wetzels, R. W. H., Rüger, J., Pluymaekers, M., & Wetzels, M. (2019). Trust in humanoid robots: implications for services marketing. *Journal of Services Marketing*, *33*(4). https://doi.org/10.1108/JSM-01-2018-0045

Vernon, P. E. (1933). Some Characteristics of the Good Judge of Personality. *Journal of Social Psychology*, *4*(1). https://doi.org/10.1080/00224545.1933.9921556

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy and Technology*, *34*(4). https://doi.org/10.1007/s13347-021-00477-0

Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*. https://doi.org/10.1109/HRI.2016.7451741

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*. https://doi.org/10.1016/j.jesp.2014.01.005

Weis, S., & Süß, H.-M. (2005). Social Intelligence--A Review and Critical Discussion of Measurement Concepts. In *Emotional intelligence: An international handbook.* (pp. 203–230). Hogrefe & Huber Publishers.

Weisman, W. D., & Peña, J. F. (2021). Face the Uncanny: The Effects of Doppelganger Talking Head Avatars on Affect-Based Trust Toward Artificial Intelligence Technology are Mediated by Uncanny Valley Perceptions. *Cyberpsychology, Behavior, and Social Networking*, *24*(3). https://doi.org/10.1089/cyber.2020.0175

Weiss, A., Michels, C., Burgmer, P., Mussweiler, T., Ockenfels, A., & Hofmann, W. (2021). Trust in everyday life. *Journal of Personality and Social Psychology*, *121*(1). https://doi.org/10.1037/pspi0000334

Westby, S., & Riedl, C. (2023). Collective Intelligence in Human-AI Teams: A Bayesian Theory of Mind Approach. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, *37*. https://doi.org/10.1609/aaai.v37i5.25755

Weymouth, R., Hartz-Karp, J., & Marinova, D. (2020). Repairing political trust for practical sustainability. *Sustainability (Switzerland)*, *12*(17). https://doi.org/10.3390/su12177055

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43). https://doi.org/10.21105/joss.01686

Wienrich, C., Reitelbach, C., & Carolus, A. (2021). The Trustworthiness of Voice Assistants in the Context of Healthcare Investigating the Effect of Perceived Expertise on the Trustworthiness of Voice Assistants, Providers, Data Receivers, and Automatic Speech Recognition. *Frontiers in Computer Science*, *3*. https://doi.org/10.3389/fcomp.2021.685250

Wiethof, C., Tavanapour, N., & Bittner, E. A. C. (2021). Implementing an intelligent collaborative agent as teammate in collaborative writing: Toward a synergy of humans and AI. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2020-January*. https://doi.org/10.24251/hicss.2021.047

Wilder, B., Horvitz, E., & Kamar, E. (2020). Learning to complement humans. *IJCAI International Joint Conference on Artificial Intelligence*, *2021-January*. https://doi.org/10.24963/ijcai.2020/212

Williams, J., Fiore, S. M., & Jentsch, F. (2022). Supporting Artificial Social Intelligence With Theory of Mind. In *Frontiers in Artificial Intelligence* (Vol. 5). https://doi.org/10.3389/frai.2022.750763

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004). https://doi.org/10.1126/science.1193147

Wynne, K. T., & Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, *19*(3). https://doi.org/10.1080/1463922X.2016.1260181

Yam, K. C., Goh, E. Y., Fehr, R., Lee, R., Soh, H., & Gray, K. (2022). When your boss is a robot: Workers are more spiteful to robot supervisors that seem more human. *Journal of Experimental Social Psychology*, *102*. https://doi.org/10.1016/j.jesp.2022.104360

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *International Conference on Intelligent User Interfaces, Proceedings IUI*. https://doi.org/10.1145/3377325.3377480

Yi, Y., Wu, Z., & Tung, L. L. (2005). How individual differences influence technology usage behavior? Toward an integrated framework. *Journal of Computer Information Systems*, *46*(2).

Yin, M., Vaughan, J. W., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300509

Zadeh, A., Chan, M., Liang, P. P., Tong, E., & Morency, L. P. (2019). Social-IQ: A question answering benchmark for artificial social intelligence. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*. https://doi.org/10.1109/CVPR.2019.00901

Zanatto, D., Patacchiola, M., Goslin, J., & Cangelosi, A. (2016). Priming anthropomorphism: Can the credibility of humanlike robots be transferred to non-humanlike robots? *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*. https://doi.org/10.1109/HRI.2016.7451847

Zhang, A., & Patrick Rau, P. L. (2022). Tools or peers? Impacts of anthropomorphism level and social role on emotional attachment and disclosure tendency towards intelligent agents. *Computers in Human Behavior*, *138*. https://doi.org/10.1016/j.chb.2022.107415

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3).

Zhang, Y., Liao, V., & Bellamy, R. (2020). *Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making*. 295–305. https://doi.org/10.1145/3351095.3372852

Zhao, M., Simmons, R., & Admoni, H. (2022). The Role of Adaptation in Collective Human–AI Teaming. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12633

Zhou, J., Arshad, S. Z., Luo, S., & Chen, F. (2017). Effects of uncertainty and cognitive load on user trust in predictive decision making. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10516 LNCS*. https://doi.org/10.1007/978-3-319-68059-0_2

Zhu, J., Yao, Y., & Jiang, S. (2023). Vulnerability or resilience? Examining trust asymmetry from the perspective of risk sources under descriptive versus experiential decision. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1207453

Zloteanu, M., Harvey, N., Tuckett, D., & Livan, G. (2018). Digital identity: The effect of trust and reputation information on user judgement in the sharing economy. *PLoS ONE*, *13*(12). https://doi.org/10.1371/journal.pone.0209071

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics*, *7*(3). https://doi.org/10.1007/s12369-014-0267-6

Zupic, I., & Čater, T. (2015). Bibliometric Methods in Management and Organization. *Organizational Research Methods*, *18*(3). https://doi.org/10.1177/1094428114562629

Accompanying Material

Supplementary Material 1. The Godspeed Questionnaire

Subsection	Question		
Anthropomorphism	Please rate your impression of the AI on these scales	Fake	Natural
	Please rate your impression of the AI on these scales	Machinelike	Humanlike
	Please rate your impression of the AI on these scales	Unconscious	Conscious
	Please rate your impression of the AI on these scales	Artificial	Lifelike
Likeability	Please rate your impression of the AI on these scales	Cold	Warm
	Please rate your impression of the AI on these scales	Dislike	Like
	Please rate your impression of the AI on these scales	Unfriendly	Friendly
	Please rate your impression of the AI on these scales	Unkind	Kind
	Please rate your impression of the AI on these scales	Unpleasant	Pleasant
	Please rate your impression of the AI on these scales	Awful	Nice
Perceived Intelligence	Please rate your impression of the AI on these scales	Incompetent	Competent
	Please rate your impression of the AI on these scales	Ignorant	Knowledgeable
	Please rate your impression of the AI on these scales	Irresponsible	Responsible
	Please rate your impression of the AI on these scales	Unintelligent	Intelligent

Please rate your impression of the AI on these scales	Foolish	Sensible

Supplementary Material 2. Trust in Automation Questionnaire with Changes

Subsection	Question		
Familiarity	I know similar AIs.	Strongly Disagree	Strongly Agree
	I have already worked with a similar AI.	Strongly Disagree	Strongly Agree
Intention of Developers	The developers are trustworthy.	Strongly Disagree	Strongly Agree
	The developers take my well-being seriously.	Strongly Disagree	Strongly Agree
Propensity to Trust	I'd rather trust an AI than not.	Strongly Disagree	Strongly Agree
	One should be careful with unfamiliar AIs.	Strongly Disagree	Strongly Agree
	AI generally works well.	Strongly Disagree	Strongly Agree
Reliability/ Competence	The AI is capable of interpreting situations correctly.	Strongly Disagree	Strongly Agree
	The AI works reliably.	Strongly Disagree	Strongly Agree
	An AI malfunction is likely.*	Strongly Disagree	Strongly Agree
	The AI is capable of taking over complex tasks.	Strongly Disagree	Strongly Agree
	The AI might make random errors.	Strongly Disagree	Strongly Agree
	I am confident about the AI's capabilities.*	Strongly Disagree	Strongly Agree
Trust in AI	I can rely on the AI.	Strongly Disagree	Strongly Agree
	I trust the AI.	Strongly Disagree	Strongly Agree
Understanding /Predictability	The AI's state was always clear to me.	Strongly Disagree	Strongly Agree
	The AI reacts unpredictably. *	Strongly Disagree	Strongly Agree
	I understand why things happen.	Strongly Disagree	Strongly Agree
	It is difficult to identify what the AI will do next. *	Strongly Disagree	Strongly Agree

Supplementary Material 3. User Preference Questionnaire

Subsection	Questions		
Short	I prefer teammates who communicate their ideas in brief and concise messages.	Strongly Disagree	Strongly Agree
	When receiving feedback, I appreciate short and to-the-point comments.	Strongly Disagree	Strongly Agree
	In team meetings, I value when discussions are kept short and focused.	Strongly Disagree	Strongly Agree
	I believe effective communication often means saying less, not more.	Strongly Disagree	Strongly Agree
	Quick, succinct responses in team chats or emails are more productive for me.	Strongly Disagree	Strongly Agree
Long	I appreciate when teammates provide detailed explanations in their communications.	Strongly Disagree	Strongly Agree
	When receiving feedback, I find more value in thorough and elaborate comments.	Strongly Disagree	Strongly Agree
	In team meetings, I prefer detailed discussions that cover topics extensively.	Strongly Disagree	Strongly Agree
	I believe that comprehensive communication prevents misunderstandings.	Strongly Disagree	Strongly Agree
	I prefer receiving emails or messages from teammates that are detailed and informative.	Strongly Disagree	Strongly Agree
Friendly	I feel more comfortable in a team when my teammates are open and approachable.	Strongly Disagree	Strongly Agree
	I appreciate teammates who make an effort to engage in casual conversations.	Strongly Disagree	Strongly Agree
	I believe that sharing personal stories strengthens a team's bond.	Strongly Disagree	Strongly Agree
	I prefer working with teammates who show warmth and friendliness.	Strongly Disagree	Strongly Agree
	Teammates who joke and laugh make the work environment more enjoyable for me.	Strongly Disagree	Strongly Agree
Formal	I value professionalism and a formal tone in all team communications.	Strongly Disagree	Strongly Agree
	I believe that keeping personal and professional lives separate improves team efficiency.	Strongly Disagree	Strongly Agree
	I prefer teammates who focus strictly on work-related topics during discussions.	Strongly Disagree	Strongly Agree
	I respect teammates more when they maintain a formal demeanour.	Strongly Disagree	Strongly Agree

A clear distinction between work and personal interaction with teammates is important to me.	Strongly Disagree	Strongly Agree