Designing and Understanding Risk Communication in Human-AI Interaction: Perceptual Effects and Implications

Sarah Schömbs

ORCID: 0009-0001-3251-3199

PhD Thesis

October 2026

School of Computing and Information Systems
Faculty of Engineering & IT
The University of Melbourne

Submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy.

This page is intentionally left blank.

Abstract

AI systems increasingly inform and autonomously make decisions that carry real-world consequence, from medical guidance to autonomous physical assistance. These systems are non-deterministic and introduce uncertainty that users must infer and navigate, often by relying on cues embedded in the interaction itself that serve to form judgements about the 'riskiness' of a system and how much it can be trusted. Risk, defined as a function of *uncertainty* and *consequence*, is therefore not simply a property to be calculated but a message that must be communicated and perceived within the interaction. The design of these risk communication cues can shape whether users fail to register a risk, underestimate it, or develop appropriate risk perceptions. This thesis empirically explores the design of risk communication across varying AI systems and its effects on user perception. Drawing on four articles spanning empirical work across online experiments, a design workshop, and laboratory studies, we examine how communicative affordances, and the design mechanisms deployed within them, can encode risk information, and how these shape risk perception and system trustworthiness.

This thesis first investigates risk communication against established perception calibration strategies, positioning risk communication as a promising cue for reducing the gulf between perceived and actual trustworthiness without sacrificing AI utility. Secondly, we systematically investigate framing strategies of consequence-explicit risk communication, surfacing *human-centredness as a design desideratum* to enhance risk perceptions and for users to internalise the risk as personally relevant, while revealing drawbacks for systemic AI risk communication. We demonstrate decision-stakes and political orientation as important contextual and individual factors influencing system trustworthiness and risk perceptions respectively, factors that need to be accounted for when designing *effective* and *appropriate* risk communication. These findings reveal important caveats. Risk communication may be unwarranted in low decision-stakes contexts due to the cognitive costs it imposes, and there is no one-solution-fits-all risk communication design suitable across end users. Further, we conduct a design-led investigation realising human-centredness through anthropomorphism as a primary design mechanism, across the communicative affordances of robot platforms. We present our exploratory work turning to consequence design, alongside a systematic comparison of uncertainty communication design, surfacing crucial tensions for encoding risk information through anthropomorphism.

Across our work, we synthesise propositions to advance risk communication design in human-AI interaction. Among them, we conceptualise *risk mirroring* and its boundaries, and unpack the tension between designing for *risk assessment*, which necessitates preserving information granularity, versus *risk perception*, which aims to shape risk responses. We discuss how risk communication can potentially function as a cognitive forcing function without imposing procedural friction, though cognitive deliberation may be attenuated where the design prioritises heuristic processing, as can occur through anthropomorphic encoding. Our empirical findings motivate the adoption of a risk-centric lens when designing interventions to support trustworthy and transparent AI, and call for an interdisciplinary design approach for risk communication in human-AI interaction. We emphasise that risk communication can be exploited to manipulate users' risk perceptions and reflect on the ethical and societal implications of this work. We conclude by identifying pathways for future work, offering the HCI research community important avenues to advance user-facing in-situ risk communication in human-AI interaction.

This page is intentionally left blank.

Declaration

I hereby declare and confirm that:

1. this thesis comprises only my original work towards the degree of Doctor of Philosophy,
2. due acknowledgement has been made in the text to all other material used,
3. appropriate ethics procedure and guidelines have been followed to conduct this research;
4. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies, and appendices.
5. the thesis comprises 100% dissertation and 0% creative outputs as agreed by the advisory committee at confirmation.

Sarah Schömbs
July 2026

This page is intentionally left blank.

Preface

This thesis is submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Melbourne. The research presented in this thesis was undertaken in its entirety during my period of candidature within the School of Computing and Information Systems, under the supervision of Associate Professor Wafa Johal and Associate Professor Jorge Goncalves.

This thesis reports four articles of which I was the primary author. Article III and Article IV are peer-reviewed published articles; Article I and Article II are currently submitted and under review. Across the four research articles, I formulated the research questions, designed the study and experimental setup, implemented the necessary systems and software, deployed the data collection instruments, recruited the participants, conducted in-person studies where applicable, collected the data, and conducted the quantitative and qualitative data analysis. I prepared the manuscripts for submission and incorporated feedback from reviewers where applicable. All articles involved the contribution of co-authors, outlined below, who supported the refinement of the study design, the technical implementation, the data analysis and the writing of the manuscript, and who provided feedback. Their contribution is greatly appreciated, and I therefore adopt the term "we" to reflect and recognise their contribution and research as a collaborative process.

The research articles are included in full, as author approved versions, within their respective chapters. In the remainder of this thesis, each article is referred to by Roman numeral (Article I to IV). Each article is preceded by an introduction that motivates and contextualises the work, and followed by a reflection that situates the article within the broader objectives of this thesis. Further, all studies reported in Article I-IV received the ethical approval from the University of Melbourne Human Research Ethics Committee. (Ethics Application ID: 27586 (Article I), 33699 (Article II), 27391 (Articles III and IV)).

Article I Sarah Schömbs, Saumya Pareek, Jorge Goncalves, and Wafa Johal. 2026. How Communicating an AI's Relational Risk Influences Trust Perceptions in AI-Assisted Decision-Making. *Submitted for publication to The International Journal of Human-Computer Studies (IJHCS), Feb 2026.*

Article II Sarah Schömbs, Jarod Govers, Jorge Goncalves, and Wafa Johal. 2026. System, Human, or Society? How Risk Framing and Political Orientation Shape AI Risk Disclaimer Perceptions in LLM Applications. *Submitted for publication to The AAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES), May 2026.*

Article III Sarah Schömbs, Jorge Goncalves, and Wafa Johal. 2025. "I Can Feel the Risks by Looking at the Robot Face": Communicating Risk through a Physical Agent. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, 236–252. https://doi.org/10.1145/3715336.3735759

Article IV Sarah Schömbs, Saumya Pareek, Jorge Goncalves, and Wafa Johal. 2024. Robot-Assisted Decision-Making: Unveiling the Role of Uncertainty Visualisation and Embodiment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3613904.3642911

I would like to acknowledge the following sources of funding that supported my research:

- The Melbourne Research Scholarship;
- The Rowden White Scholarship;
- The Diane Lemaire Scholarship;
- The Faculty of Engineering and Information Technology Conference Travel Scholarship;
- The Google Travel Grant;

Generative AI was used in this thesis to improve clarity and expression. All generated output was consistently checked and verified. No other third-party editorial assistance was provided in the preparation of this thesis.

This page is intentionally left blank.

Acknowledgements

My journey over the past few years has been made possible by the countless people I have met along the way, people who not only crossed my path but enriched it in ways I did not expect or hope for. Looking back, I never assumed I would do a PhD, let's be honest, I did not even know it was an option. Turns out, it was one of the best decisions I have ever made, and I savoured every bit of it. In many ways, this is a testament to my two supervisors, Wafa Johal and Jorge Goncalves, who have not just guided me, but made this journey immeasurably joyful. Our supervision meetings and in-between chats were one of the best parts of my PhD. So my gratitude goes out, first and foremost, to Wafa Johal, who has inspired me every step of the way with her never-wavering strength, curiosity, and excitement for exploring new topics and diving into research. I am in continuous awe of your sheer span of expertise. You encouraged me to nurture my love of questioning everything and anything, to be creative, and explorative. Thank you for believing in me. I have always felt supported, while you never ceased to challenge me, and by doing so, you allowed me to grow into my own strength as a researcher. To Jorge Goncalves, our endless banter has often lifted the heaviness that may come with a PhD, and made it feel light and simply fun. Thank you. You were truly my Gandalf, guiding me through the endless seas of statistical models and my sometimes tangled thoughts - always ready to cut a paragraph in half.

I would also like to thank Eileen Roesler, without whom I would probably not have started a PhD in the first place, who introduced me to human-robot interaction, and who encouraged me to pursue this trajectory. Thank you, I hope you know how important your support was and still is to me.

To my best friend, Tereza Gross, who has always supported me in whatever I do, while never expecting something of me, and whose friendship has been one of the biggest constants in my life, thank you for always finding the right words when things get hard, and for letting me disappear during deadline times without ever holding it against me. You show me how love and friendship do not care about living on different continents. To my friend, Jacobe Klein, who has shown me what true generosity looks like, and who is so insanely cool, thank you for being exactly who you are, always. To my friend Lina Brockmeier, who I met at exactly the right time and in exactly the right place, and whose friendship I consider one of the biggest gifts of this PhD, thank you. Our morning runs along the beach and our shared book obsessions have been escapism par excellence. To my friend Charlotte Kobiella, one of the unexpected gifts of this research journey, who, all of a sudden, became our roomie and who continues to bring so much light into our life. I cannot wait for us to live on this side of the world together.

To my friends and lab mates Yan Zhang, Tharaka Sachintha Ratnayake, Jian Zhang, Farshad Haghgoo Daryakenari, Ramtin Tabatabaei, and Muhammad Bilal – I am so impressed with what you do, how you make it seem so effortless, and I treasure all our chats, meetings, research collaborations, coffee breaks, Yo-Chi runs, lab demos, and laughs. I cannot wait to see where life takes you and where our paths may cross again. Thanks to Saumya Pareek, who was my academic and matcha partner in crime. Thanks to Piumi Perera, for her kindness and calm. Thanks to Ilaria Torre, Sarah Webber, and my Wattle Fellowship crew, especially Linh Do, who have allowed me to merge my commitment to environmental protection with my research life during

this PhD.

Thanks to Cherie Sew, Jarod Govers, Suwani Gunasekara, Songyan Teng, Adélaïde Genay, Benjamin Tag, and Eduardo Velloso for all the laughs, the shared meals, and all the shenanigans in the lab, during Friday pub hangs, or conferences across the world. Thanks to all my amazing colleagues, peers, and friends at the HCI group who created an amazing and supportive work environment, and to the colleagues, friends, and co-authors I got to meet during conferences and research visits, among them Jan Leusmann and Leon Fröhling.

And finally, to my family. Thank you to my parents and brother, who have set an example for how important it is to love your profession, whatever it may be, and to never settle for less. I would also like to thank them for opening the door to music and movement since I was a child, which, more often than not, carried me through the tough times and helped me find my strength. My mum, who has always told me to spread my wings.

But most important of all, I would like to thank my partner, Sven Brzostowski, although gratitude does not do it justice. I am still in wonder of how you just packed our two suitcases and moved across the world with me to start this adventure, never once questioning me. You have fed me through endless deadlines, picked up our life when I was knee-deep in work, and made me laugh every single day, but honestly, you just make me love life. You have taught me fearlessness, ease, and what is important in life, and I know that together, we can do whatever we set our minds to.

Contents

Abstract iii

Declaration v

Preface vii

Acknowledgements x

1 Introduction 1

1.1 Research Questions and Contribution . . . 2

1.2 Research Methodology . . . 8

1.3 Ethical Considerations . . . 12

1.4 Thesis Outline . . . 14

2 Background 15

2.1 Theoretical Foundations of Risk Perception and Communication . . . 15

2.1.1 Defining Risk . . . 15

2.1.2 Risk Perception . . . 16

2.1.3 Judgement and Decision-Making about Risk . . . 17

2.1.4 Risk Communication . . . 20

2.2 Risk in Human-AI Interaction . . . 22

2.2.1 Defining and Classifying AI Risks . . . 23

2.2.2 Risk Communication in Human-AI Interaction . . . 24

2.2.3 Trust and Risk in Human-AI Interaction . . . 25

2.3 Summary . . . 29

3 How AI Risk Communication Shapes Perceived System Trustworthiness 30

3.1 Introduction . . . 30

3.2 Article I . . . 31
3.3 Chapter Reflection . . . 87

4 Framing Effects of AI Risk Consequences in LLM Interaction 89
4.1 Introduction . . . 89
4.2 Article II . . . 90
4.3 Chapter Reflection . . . 111

5 Encoding Risk Consequences through Anthropomorphism 112
5.1 Introduction . . . 112
5.2 Article III . . . 113
5.3 Chapter Reflection . . . 131

6 Designing Uncertainty Communication in Robot-Assisted Decisions 132
6.1 Introduction . . . 132
6.2 Article IV . . . 133
6.3 Chapter Reflection . . . 150

7 Discussion and Future Directions 152
7.1 The Message and The Medium . . . 152
7.2 Toward Human-Centred Risk Communication . . . 154
7.3 Humanising Risk through Anthropomorphism . . . 155
7.4 Why XAI Needs Risk Communication . . . 158
7.5 One Size Does NOT Fit All . . . 159
7.6 Risk Communication as a Cue . . . 161
7.7 Societal and Ethical Considerations . . . 164
7.8 Future Directions and Limitations . . . 166
7.8.1 Longitudinal Studies and Real-world Settings . . . 166
7.8.2 Behavioural Outcomes of Risk Communication - Beyond Reliance . . . 167
7.8.3 Two Pathways into Adaptive Policies for Risk Communication . . . 168
7.8.4 Risk Communication as a Bi-Directional Process . . . 169
7.8.5 Overlooked Determinants of Risk Perception in Human-AI Interaction . . . 170

8 Conclusion 172

List of Figures

1.1 Overview of the research questions guiding this thesis and their associated empirical Articles I–IV, illustrating the reciprocal relationship between the *design* and *perception* of risk communication across AI systems with varying communicative affordances. 3

List of Tables

1.1 Research questions and the chapters in which they are addressed. 4

Chapter 1

Introduction

> *"Citizens of modern industrial societies are presently learning a harsh and discomforting lesson—that the benefits from technology must be paid for not only with money, but with lives. Whether it be ozone depletion and consequent skin cancer from the use of spray cans, birth defects induced by tranquilizing drugs, or radiation damage from nuclear energy, every technological advance carries some risks of adverse side effects."*
>
> Fischhoff et al. (1978)

Technological advancements have always involved an intricate trade-off between benefits and risks. New technologies expand what individuals and societies are able to accomplish, but they also introduce new uncertainties and possibilities of harm.

The foundations of risk research were built in response to the societal challenges of their time, with nuclear power as one of its major catalysts in the 1960s, which gave rise to the subjective nature of risk perception when experts first recognised the discrepancy between objective risk communication and subjective risk perception [188]. While risk is constituted by two core components, *uncertainty* and *consequence* [10, 175, 221], simply presenting information about either is not sufficient. Individuals rarely engage in formal risk assessments by systematically quantifying probabilities and expected negative outcomes. Instead, risk perception is a latent, multidimensional, and inherently subjective construct [177, 181], shaped by cognitive and affective processes [210, 211]. Crucially, precisely because uncertainty is inherent to risk, individuals' judgements about risk are subject to the cognitive heuristics and biases that govern how people evaluate information and form judgements under uncertainty [88, 179, 198]. Risk communication is therefore not merely a matter of presenting information, but of accounting for the subjective perceptions of that information, which highlights the reciprocal relationship between *the design risk communication* and how it is *perceived* by individuals.

The aforementioned trade-off is mirrored today in the case of Artificial Intelligence (AI), which presents one of the most consequential technological shifts of our time, and with it, introduces non-trivial tension between utility and risk. While AI systems offer wide-ranging benefits, among them the reduction of language and knowledge barriers, the democratisation of access to information, and the acceleration of scientific progress, they also produce non-factual information, breach privacy, and reproduce harmful misrepresentations [217]. At the same time, robot platforms, often fuelled by advancing models, extend AI systems' capabilities in sensing, planning, and acting as physical agents [162], serving not only as potential sources of risk but as conduits through which risk information can be designed and communicated in-situ. Slattery et al. [176] describe how AI is evolving toward "greater autonomy, multi-modal capabilities, and agentic behaviors" and, by doing so, gives rise to new risk categories, noting that "the convergence of advanced language models with robotic systems may create novel risks". According to Slattery et al. [176], these risks span domains on

the system side, such as a lack of transparency and interpretability, or AI pursuing goals that conflict with human values, and on the user side, such as unsafe use and overtrust, or a loss of agency and autonomy. Over the past years, hundreds of AI incidents have been recorded by the AI Incident Database [126][1], revealing a diverse risk landscape spanning classifiers, large language models, and robotic systems across a wide range of deployment domains. In simple terms, these systems are imperfect, fail, and depending on their deployment context, carry risks with potentially detrimental real-world consequences. Recent work on decoding real-world AI incidents through the causal AI risk taxonomy developed by a research team at MIT [176] reveals that the burden of those consequences is disproportionately carried by *end-users*, in that the majority of AI incidents occur *post-deployment* [47], in the very contexts where end-users encounter and interact with these systems. The question, then, is not whether AI systems will fail, but whether users are equipped to anticipate the associated risks and respond to them appropriately.

Importantly, AI systems are no longer tools in the traditional sense, but function as cognitive extensions that influence how we think, decide, and act [37, 38]. Nearly three decades earlier, in 1997, Parasuraman and Riley painted the picture of how automation has "freed humans from many time-consuming and labor-intensive activities", but does not yet have the ability to automate cognitive processes or decision-making, eventually asking "Could machine displacement of human thinking become more commonplace in the future?" [139]. Arguably, we have reached that point. Within recent years, we shifted from automation systems confined to physical labour and deterministic tasks, to AI systems that enter everyday contexts where people seek its guidance and follow its recommendations — reshaping human cognition [37, 38]. Notably, AI adoption is accelerating faster than comprehension, and users are often not equipped to navigate its probabilistic nature or to evaluate the capabilities and limitations of the systems they interact with. At the individual level, this gap can manifest in distortions between actual and perceived system trustworthiness [165, 166] and over-reliance [29, 103, 141, 226], as well as in the threat of intellectual complacency [37] and an alignment of humans to machines [171], reinforcing the need to design cues that help users calibrate their perceptions of a system and the associated risks. Trust has emerged as the central construct through which human-AI interaction research has approached these challenges, with appropriate trust, defined as trust that reflects the actual capabilities and limitations of a system [111], positioned as a prerequisite for appropriate system use.

In the pursuit of trustworthy and transparent AI, research across ML (e.g. [13, 15, 31, 149]), human-AI interaction (e.g. [141, 203, 212, 213]), and human-robot interaction (e.g. [45, 101, 113, 133]) has investigated a range of explanation approaches and strategies for communicating innate uncertainty, on the premise that opening the black box of AI systems will lead users to calibrate their trust more appropriately and engage more critically with system output and behaviour. Such explanation approaches have, however, frequently proved insufficient (e.g. [76, 128, 212, 226]). Consequently, recent work advocates for new approaches to explainability [226] and new interaction paradigms [128], with some of this work turning to risk literature to inform novel intervention strategies [76] in an effort to address the representation biases and distorted judgements that misleading uncertainty formats elicit. Nevertheless, risk communication in human-AI interaction remains scarcely investigated, and informing its design requires drawing on the theoretical foundations of risk research and an empirical understanding of how risk information is subjectively perceived.

1.1 Research Questions and Contribution

In this thesis, we aim to advance both the *design* and *understanding* of user-facing risk communication in human-AI interaction. We enrich our understanding by investigating the perceptual effects of risk communication, the cognitive mechanisms through which risk information is processed and reasoned about, and the individual and contextual determinants that shape its effectiveness. We advance the design by exploring

[1] https://incidentdatabase.ai

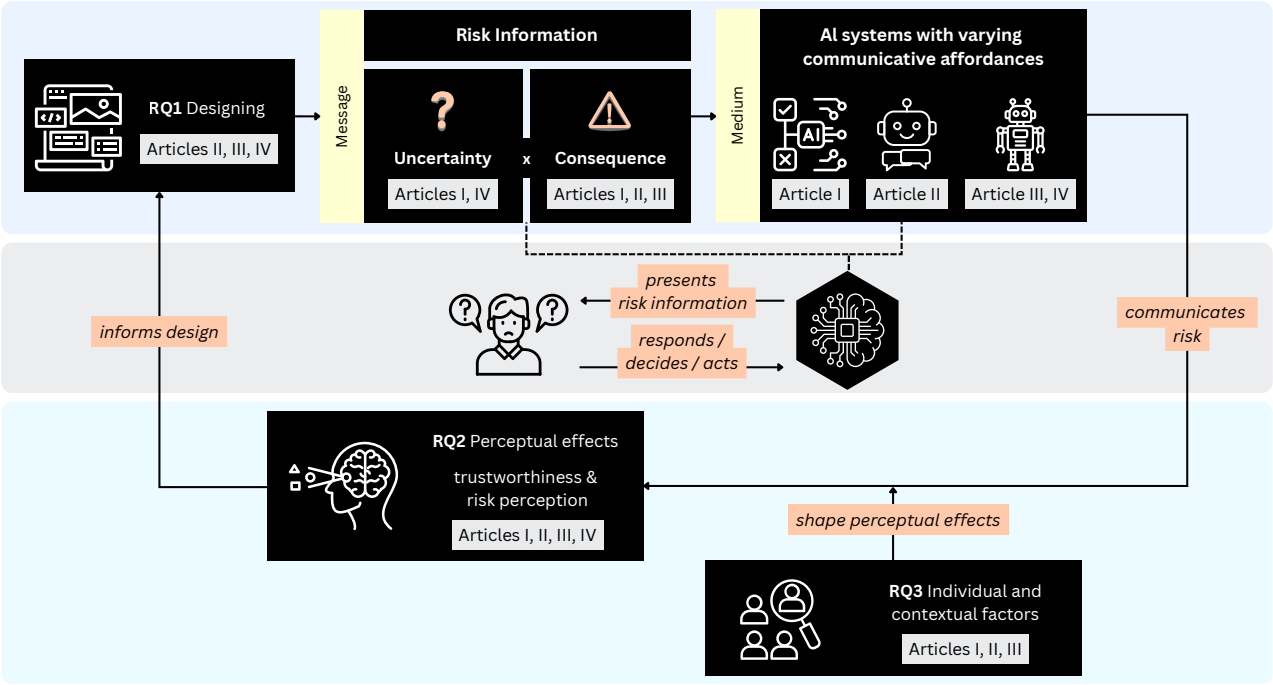


Figure 1.1: Overview of the research questions guiding this thesis and their associated empirical Articles I–IV, illustrating the reciprocal relationship between the *design* and *perception* of risk communication across AI systems with varying communicative affordances.

the risk communication landscape across AI systems with varying communicative affordances, surfacing the design considerations and implications that emerge across this breadth. Collectively, these aims are captured in and guided by the following overarching research question.

Overarching RQ

How can we improve the design of user-facing risk communication across AI systems with different affordances in human-AI interaction?

We further deconstructed the overarching research question into four sub-questions that are neither hierarchical nor sequential. Rather than building incrementally across the body of work, the outlined research questions are deliberately tensioned and operate across the articles presented in this thesis (Chapter 3 - 6). This is most pronounced in the relationship between RQ1 and RQ2, where questions of perception and design mutually inform one another. *How* risk core components are designed and communicated through AI affordances shapes how users *perceive* system trustworthiness and risk, and in turn, how users perceive such communication informs the design response. Table 1.1 provides an overview of RQ1–4 and the chapters in which they are respectively addressed.

RQ1, RQ2 - Design and Perception Risk communication is broadly defined as the "exchange or sharing of risk-related data, information and knowledge between and among different target groups" [10]. Across this thesis, we conceptualise risk communication as risk information communicated to an *end-user*, *in-situ*, through the *communicative affordances* of an AI system. Such risk information can be non-relational to the AI system itself, in that the system serves as a conduit for risk information about, for example, health risks in

Table 1.1: Research questions and the chapters in which they are addressed.

	Research Question	Addressed in
RQ1	How can we **design** the communication of risk core components, ***uncertainty*** and ***consequence***, in human-AI interaction, given varying **AI affordances**?	Chapter 4, 5, 6
RQ2	What are the **perceptual effects** of risk communication in human-AI interaction, in terms of users' perceived **system trustworthiness** and **risk perception**?	Chapter 3, 4, 5, 6
RQ3	What **individual** and **contextual factors** influence the perceptual effects of risk communication in human-AI interaction, and how?	Chapter 3, 4, 5
RQ4	What **recommendations** can we provide to promote effective user-facing risk communication through AI systems with varying affordances?	Chapter 3, 4, 5, 6

a care assistance context (see Chapter 5). The majority of our work, however, concerns relational risk [189] (Chapters 3, 4, 5), in that the risk information content pertains to a risk that stems from the AI system itself, and is communicated through the communicative affordances of the AI system. We define communicative affordances as the communication channels and encoding possibilities available for risk communication in human-AI interaction. Importantly, risk literature served as a guiding lens through which we approached both the design of risk communication and the interpretation of empirical findings across this body of work, drawing on related concepts such as cognitive biases and judgement under uncertainty. Risk, as conceptualised in this thesis, comprises two core components, *uncertainty* and *consequence*, and while all articles engage with risk communication broadly, individual works direct particular attention to the design and investigation of one component, with Articles 4 and 5 zooming into consequence-explicit risk communication and Article 6 foregrounding uncertainty-explicit risk communication.

In the following, we address **RQ1** and **RQ2** in tandem across Articles I–IV, reflecting the inherent interplay between the design of risk communication and its perceptual effects. Across these articles, we surface design tensions, trade-offs, and considerations that guide the design of uncertainty-explicit and consequence-explicit risk communication in human-AI interaction, and advance understanding of the perceptual effects that such communication elicits and the cognitive mechanisms through which risk information is processed and interpreted.

RQ 1 - Design

How can we design the communication of risk core components, (a) uncertainty and (b) consequence in human-AI interaction, given varying AI affordances?

RQ 2 - Perceptual Effects

What are the perceptual effects of risk communication through AI systems on user's system trustworthiness and risk perception?

Article I (Chapter 3) takes a first step toward positioning risk communication as a system trustworthiness calibration strategy in AI-assisted decision-making. Through a user study ($N = 300$), we contribute empirical insights into how risk communication shapes **system trustworthiness** perceptions across varying AI behaviour and decision-stakes, situated in a task-specific AI classifier context. We demonstrate that risk communication does not erode system trustworthiness nor diminish perceived AI utility; rather, it functions as a *useful* cue through which users gauge system trustworthiness, recalibrate expectations, and engage in more deliberate scrutiny of AI outputs. While risk communication design was not the primary contribution of this work, the risk communication stimulus was nevertheless risk-theoretically informed, entailing both uncertainty and consequence components and drawing on framing strategies from risk research situated in medical decision-making [6, 70, 127, 148]. Article I thus positions the communication of a system's relational risk as an important pathway for Explainable AI (XAI) research, and demonstrates the value of integrating risk theory into human-AI interaction. The full article presents the comprehensive empirical findings of this work.

Article II (Chapter 4) advances the understanding of both the design and perception or risk communication in human-LLM interaction through a user study ($N = 208$) investigating the *framing of consequence* in AI risk disclaimers. We systematically investigate the design of risk-consequence communication across three distinct framing lenses, whereby the same consequence is communicated through a system-capability, human-interaction, or a systemic lens. We demonstrate that the *presence* of consequence-explicit AI risk disclaimers decreases **system trustworthiness** perceptions and heightens **risk perception**. Critically, we unpack the cognitive mechanisms through which the AI risk information is processed, revealing that the systemic risk framing produces psychological distance and avoidance responses, while a human-interaction framing proves most effective by communicating *how a risk manifests for the user experience*, increasing its perceived severity and inducing negative affect such as worry and concern. Article II thus surfaces *human-centredness by design* as a *desideratum* for risk communication in human-AI interaction. We further demonstrate that system trustworthiness and risk perception constitute distinct and complementary constructs that respond to different aspects of risk communication, with framing driving risk perception but not system trustworthiness perceptions, warranting their joint consideration as outcome measures in XAI research. The empirical findings carry direct implications for the design of AI risk disclaimers, as well as for policy and regulation governing the risk communication of LLM systems as general-purpose technologies, which we elaborate on in the full article.

Article III (Chapter 5) extends the exploration into human-centred risk communication by moving toward social robot platforms as conduits of risk information and facilitators of risk communication through their unique AI affordances. In contrast to Articles I, II and IV, Article III focuses on *non-relational risk*, in that the robot serves as a conduit for risks that are relevant to the user but do not stem from directly from the system itself. Guided by the objective to advance *human-centredness by design* and HRI literature, we investigate anthropomorphism [156] and embodiment [158] as design mechanisms for risk communication, combining a design workshop with HCI experts ($N = 9$) and a user study evaluating the resulting design ($N = 28$) deployed on a real-world robot platform. We surface important design considerations and frictions, among them tensions around *information precision* and the risk of *over-inducing affect*, and reveal how *anthropomorphic morphology shapes the interpretive context of risk information*. Empirically, we demonstrate that anthropomorphically encoded risk communication elevates **risk perception** across its affect and severity dimensions, and facilitates *risk mirroring*, in that it elicits self-projection and empathetic responses to how a risk would manifest for oneself and for others. Article III thus extends the design landscape of consequence-explicit risk communication to embodied physical agents, further reinforcing *human-centredness by design* as a design desideratum, and reveals its application potential for health risk communication in public spaces.

Contextualised in a robot-assisted decision-making paradigm ($N = 36$), **Article IV** (Chapter 6) extends the investigation of anthropomorphic encoding as a design mechanism to uncertainty-explicit risk communication, examining how such encoding is perceived compared to an uncertainty display designed to miti-

gate cognitive biases and informed by health risk communication [3, 123, 137]. We surface a fundamental design tension between *information granularity* and *holistic uncertainty comprehension*, demonstrating that anthropomorphic encoding supports holistic understanding but affords less critical deliberation, whereas information visualisation informed uncertainty communication elicits more deliberate cognitive engagement with the communicated uncertainty information. We further reveal that the degree of uncertainty itself determines the need for information richness, with high uncertainty instances rendering the robot system as less transparent in users' perceptions, surfacing *the degree of uncertainty* as a consequential design parameter. Article IV thus contributes empirical insights into how users perceive, reason about, and interpret uncertainty information in robot-assisted decision-making, and derives design considerations and implications for uncertainty-explicit, instance-based risk communication, while bridging a critical gap between human-robot interaction and information visualisation research [191].

RQ3 - Context and Individual Factors Situational risk has traditionally been regarded as an important factor in trust models spanning organisational psychology [122] and human-automation interaction [111, 189]. Situational risk [189], by nature, characterises what is at stake. Situational risk presents an equally important factor in human-AI interaction, where AI systems are deployed across a wide variety of contexts and tasks, each carrying vastly different real-world consequences. A user interacting with a recommender system to identify the next film to watch faces fundamentally different stakes than one relying on an AI-assisted decision tool to support early detection of skin cancer, where the actions and decisions informed by the system may carry profound consequences for the individual. How risk communication should be designed across such contexts, and how *decision-stakes* moderate its effectiveness, represents a critical and underexplored dimension of risk communication in human-AI interaction. Prior work on AI-assisted decision-making has emphasised the importance of decision-stakes, but predominantly investigated either end of this spectrum - low situational risk (e.g., AI supporting the identification of animals or shapes [140] or answering factual tasks [96]) or high situational risks (e.g., AI supports skin cancer screening [32] or the assessment of jail time in criminal law cases [90]).

At the same time, *user-facing* risk communication is not a matter of public risk perception across homogeneous groups but targets the *individual* interacting with a system and receiving risk information in-situ, foregrounding a fundamental question: *who* is receiving this information, and through what lens do they perceive, weigh, and interpret it? Risk literature has consistently shown that risk perception is highly individual and subjective [25], pointing to a plethora of individual difference variables that shape how risk is perceived and responded to, among them prior experience with risk [92], cultural background [66], world views [46], and political orientation [39, 86]. The degree to which information aligns with one's moral values, for instance, determines its persuasiveness [42, 58]. Yet, which individual differences are relevant in the context of human-AI interaction, and how they may shape risk perceptions, remains largely unexplored.

Collectively, both dimensions demonstrate that risk communication in human-AI interaction is *contextually embedded* and received by *individuals* with differing characteristics, motivating **RQ3**.

RQ 3 - Individual and Contextual Factors

> What individual and contextual factors influence the perceptual effects of risk communication in human-AI interaction, and how?

In the following we address **RQ3**, turning to the contextual and individual factors that the empirical work presented in this thesis surfaces as determinants of risk communication effectiveness.

In **Article I**, we contribute empirical insights through a systematic comparison of perception calibration

strategies across low- and high-stakes decision contexts. Our findings demonstrate that *decision-stakes* significantly moderate their effects on users' perceptions of system trustworthiness and therefore constitute a crucial determinant of strategy *appropriateness*. Decision-stakes presents the very factor that illuminates the shortcomings of AI accuracy as an established XAI strategy, rendering it ill-suited in high-stakes contexts, whereas risk communication, while not subject to the same shortcomings, may nevertheless come with unwarranted cognitive costs in low-stakes contexts. Decision-stakes thus emerges as a design parameter that must be foregrounded in risk communication design decisions. In the full article, we provide detailed empirical findings and discuss how the moderating role of decision stakes shapes the benefits and drawbacks of each calibration strategy, and thus its appropriateness.

In **Article II**, we direct attention to *political orientation* as a central individual difference variable for risk communication in human-AI interaction, motivated by two converging rationales. First, AI is increasingly a subject of political discourse, from debates around AI regulation and oversight [35], technocracy [44], and democratic epistemic agency [43], to real-world instances of AI systems deployed to spread politically charged content [153]. Second, political orientation increasingly serves as a mental shortcut through which AI systems are appraised [42, 224]. Risk communication should purposefully counteract such emerging risks, yet its own effectiveness may be *conditioned* by how well the communicated risk aligns with a user's political orientation. We contribute empirical insights to **RQ3** and demonstrate that the effectiveness of AI risk disclaimer framing strategies is, indeed, *contingent* on an individual's political orientation, and that human-interaction framing emerges as a relatively robust strategy across orientations, raising important questions around *user-adaptive versus default risk communication design.* Our findings further unpack deeper evaluative mechanisms associated with individualistic, liberty-focused orientations on the one hand, and collectivist, egalitarian orientations on the other. The full article elaborates on the moderating role of political orientation across liberal, moderate, and conservative-leaning individuals, and deepens the ethical discussion around resulting design considerations, thus providing important implications for HCI and AI policy and regulation.

Article III, while not designed to investigate **RQ3** through a systematic comparison of individual or contextual factors, nevertheless surfaces *lived experience* with the communicated risk as an important individual difference shaping people's risk perceptions, consistent with risk literature demonstrating that prior experience shapes how individuals perceive and respond to risk [92]. Notably, lived experiences emerged as either a *catalyst* for risk perception or source of what can best be characterised as *desensitisation* effect through risk familiarity.

RQ4 - Recommendations Lastly, we turn our attention to **RQ4**, which asks what recommendations we can derive to promote effective user-facing risk communication across AI systems with varying affordance.

RQ 4 - Recommendations

> What recommendations can we provide to promote effective user-facing risk communication through AI systems with varying affordances?

We address **RQ4** across all four articles. Rather than constituting a dedicated empirical focus, recommendations emerge throughout the body of work as design consideration and implications derived from our empirical findings. These span practical design considerations for lightweight linguistic interventions such as framing in LLM and AI-assisted decision-making contexts; design tensions, benefits and pitfalls for anthropomorphic and embodied risk communication across social robot and robot-assisted decision-making platforms; contextual and individual determinants of risk communication effectiveness; and implications for the appropriate assessment of risk communication effectiveness. The work presented in this thesis thus informs the

HCI community, system designers and researchers, as well as AI policy and regulation.

1.2 Research Methodology

We employed a mixed-method research design across the four empirical studies presented in this thesis (Chapter 3–6). The studies in Article I and Article II were conducted as *online experiments* via the crowdsourcing platform Prolific[2], whereas the studies in Article III and Article IV were conducted *in-person* in a controlled laboratory environment. Article III further presents a *design workshop* (Study I), conducted *in-person* on-site at the University's laboratory. Across Articles I–IV, quantitative data were captured through validated questionnaires collected via self-report, and qualitative data were elicited through open-ended post-task responses (Articles I, II, IV), focus group discussions in Study I of Article III, and semi-structured interviews in Study II of Article III. The surveys deployed across Article I-IV were designed using the online platform Qualtrics[3].

Notably, we predominantly constructed Cumulative Link Mixed Models (CLMMs) for the quantitative data analysis across this thesis, as they have been increasingly advocated for in HCI research as an appropriate frequentist approach for ordinal data, particularly in repeated measures designs where random effects must be accounted for [204].

The methodology was carefully selected according to the objectives of each study, the characteristics and communicative affordances of the respective AI system, and broader methodological considerations, including sample size, experimental control, ecological validity, internal validity, and the reduction of potential confounds. The methodological breadth of this thesis also reflects the methodological richness of HCI research, in which different methods offer complementary ways of examining the design and perception of interactive systems. For instance, in-person laboratory studies were prioritised when the research involved the embodied and physical properties of robotic systems, whereas a design workshop was conducted where the research required generative and expert-led exploration of a risk communication design space. The rationale for these choices and their associated trade-offs is detailed in the study descriptions that follow. For each study, we outline the study design, experimental manipulations, operationalisation of tasks, stimuli and scenarios, measures, recruitment procedures, and data analysis approaches.

In Article I, we conducted a *scenario-based online user study* ($N = 208$) to investigate the perceptual effects of risk communication on *system trustworthiness* relative to established perception calibration strategies contextualised in *AI-assisted decision-making*, and its interaction with decision-stakes as a contextual factor. We employed a 4 (*Perception Calibration Strategy*: Risk Communication, AI Accuracy, AI Apology, Baseline) $\times$ 2 (*Decision-Stakes*: Low, High) between-subjects factorial design.

The experimental task constituted an AI-assisted diagnostic task, a well-established paradigm in human-AI interaction, presented as a threshold-based visual estimation task [55, 154]. We distributed $N = 9$ diagnostic cases across three Blocks (I-III), each encoding different AI performance profiles (I: correct, II: erroneous, III: correct) to operationalise varying observable AI performances, informed by the research objectives of this work and adopted from related works [90, 97, 140, 194]. The AI system was simulated with a fixed accuracy at 78% to ensure experimental control over the injected error type and its position within the trial sequence in Block II. The classifier stimuli and scenario narrative for the diagnostic task were adapted from prior work [55], in which they were validated for manipulating decision-stakes in human-AI interaction. The diagnostic interface resembled a typical classifier system, and we deliberately designed an AI processing indicator for each case to support the perceived authenticity of the AI system. We deployed a training block prior to the main experiment to support task familiarisation.

[2] https://www.prolific.com/
[3] https://www.qualtrics.com/

System trustworthiness served as the dependent measure, operationalised as a latent construct across the determining dimensions of perceived *AI transparency*, *performance*, and *utility*, assessed via the validated Multi-dimensional Trust Questionnaire [157] in a repeated measures design administered after each performance block. After each diagnostic case, we measured participants' agreement with the AI recommendation and self-reported confidence in that agreement [96, 140], which served as robustness checks to ensure our experimental manipulation of AI performance was perceived as intended. We captured the Trust in Automation – Propensity to Trust subscale [102] as a covariate to account for individual differences [141].

We recruited participants via Prolific and applied several pre-screening and exclusion criteria. Recruitment was restricted to participants with sufficient English language proficiency, a prior platform approval rate of $\geq$98%, and we applied Prolific's internal quota sampling to balance gender distribution. We deployed a CAPTCHA at study entry, along with attention and comprehension checks which served to ensure data quality, and each participant was permitted to take part only once to mitigate learning effects.

We conducted an a priori power analysis accounting for interaction effects to determine the minimum required sample size using the R package `InteractionPoweR` [14]. We assessed the internal consistency among scale items via Cronbach's α. We analysed the quantitative data using CLMMs to account for the ordinal structure of the data, the repeated measures design, and random effects. The models were constructed in R using the statistical package `ordinal`[4]. In our models, we also accounted for the temporal dependency of system trustworthiness across the performance blocks [140]. The qualitative data were analysed using Thomas' general inductive approach [192], with the inter-rater reliability assessed via Cohen's $\kappa(= 0.81)$.

In Article II, we again conducted a *scenario-based online user study* ($N = 300$) to investigate the perceptual effects of the *consequence framing* in AI risk disclaimers embedded in *LLM systems*, and their interaction with people's political orientation, to inform their design respectively. We designed the study in a between-within factorial design, with *consequence framing* (capability, human-interaction, systemic framing) operationalised as the within subjects variable, and a between-subjects control condition in which no AI risk disclaimer was embedded in the LLM interface to investigate the effects of *disclaimer presence*.

Our experimental manipulation built on a sociotechnical evaluation framework [218]. The AI risk disclaimer consisted of a consequence-explicit component, whose framing was manipulated to either communicate the consequence through the lens of the *system's capability*, how the consequence manifests within *the human-interaction*, or through a *systemic lens*.

The study adopted a scenario-based design, with each scenario operationalising typical risk-containing LLM interactions [217] informed by the MIT AI Incident Tracker database [132]: seeking information about health, LLM assistance in the hiring process, and booking assistance. Rather than static depictions, we embedded simulated LLM interaction interfaces within the online survey. Each interaction was presented via a typical chat interface, following a hard-coded conversational structure which we adopted consistently across scenarios to mitigate confounds. To increase ecological validity, we designed the LLM interface in line with existing real-world systems, and positioned the AI risk disclaimer following industry examples (e.g. ChatGPT [136] and Claude [5]) which also reflects regulations for AI disclosures put forward by the EU AI Act (Article 50 (5) [52]). To reduce carry-over effects throughout the online experiment, we introduced each stimuli as a separate AI system that necessitated evaluation, deployed a Latin square design for the presentation order, and deployed a stroop task between the main experimental phase and the post task questionnaire.

We operationalised *risk perception* as the dependent measure, assessed across its perceptual dimensions *perceived severity* and elicited *negative affect* via a validated risk perception questionnaire [210], alongside a unidimensional holistic risk perception item [221]. We further assessed *system trustworthiness* via a reduced version [146, 163] of the validated Trust between People and Automation scale [85], measured across *perceived*

[4] https://cran.r-project.org/web/packages/ordinal/index.html

trust and *distrust*; we opted for this over the MTQ (Article I) given that this study was not contextualised in an AI-assisted decision-making scenario. We assessed political orientation across generic, social, and economic dimensions following prior work on political self-placement [118]. To account for individual differences, we included the Trust in Automation – Propensity to Trust subscale [102] and the AI literacy – Understand AI subscale [34] as covariates.

We assessed a priori power using the R package `InteractionPoweR` [14] to determine the minimum sample size while accounting for interaction effects. To maintain data quality, we deployed a CAPTCHA verification, a JavaScript browser test to verify that participants could access the LLM simulations, embedded an attention check and an AI risk disclaimer recognition check as additional exclusion criteria, and set the participants' approval rate to $\geq$ 98%. We applied additional Prolific pre-screening criteria to ensure that participants were based in the United States, proficient in English, and aged $\geq$ 18 years or older. Restricting the sample to US-based participants was necessary because the political orientation measure was designed for a US population.

Article III comprises two studies. **Study I** reports a *design workshop* to understand the anthropomorphic encoding of risk information through *robot platforms*, adopted as a HCI methodology well-suited to surface design tensions and outcomes [71]. The design workshop served as a preliminary step for **Study II**, which presents a *pre-post in-person user study* to gain in-depth insights into the perceptual effects and application potential of the resulting design approach.

Study I describes a design workshop structured around an exploratory, collaborative design generation, and the surfacing of design tension. The workshop drew on information visualisation (InfoVis) research [71, 106], with workshop stimuli derived from an affective visualisation design corpus [106]. The workshop consisted of a familiarisation phase with the robot platform and relevant theoretical concepts to reduce knowledge imbalance among participants, a design activity phase, and a semi-structured group discussion to identify design friction. The familiarisation phase further served to mitigate the 'novelty effect' common in human-robot interaction research [184], which emerges due to the interaction with a yet unfamiliar emerging technology. We utilised the Furhat[5] robot platform selected for its anthropomorphic morphology, customisable appearance, and facial expressiveness.

We recruited HCI experts ($N = 9$) via snowball sampling, with the expert pool determined through prior survey eligibility criteria to ensure both disciplinary diversity and HCI expertise [94]. The collected data comprised the design outcomes produced during the design activity phase and the qualitative data elicited from two sources: the participants' individual elaborations on their designs to ensure that interpretations remained grounded in participants' own rationale rather than externally imposed by the research team, and the group discussions. The qualitative data were analysed using a reflexive, inductive-leaning thematic analysis [23] given the exploratory nature of the work.

Study II details a pre-post in-person user study to assess anthropomorphism and embodiment as design mechanisms for risk communication, and to investigate a social robot as both a facilitator of those mechanisms and an embodied agent serving as a conduit for risk information. Building on the design considerations and outcomes of Study I, we selected two health risk scenarios and refined the corresponding designs for deployment on the Furhat robot. Concretely, we developed a Python script to generate robot faces that encoded risk information through customised texture overlays, which we exported as `.charpacks` and deployed on the robot. We deployed gaze behaviour and facial expressions to enhance the expressiveness of the risk communication. The experimental task constituted a risk information exploration task, in which participants navigated health risk data through a graphical user interface. Each data point, encoding a specific risk value, triggered the corresponding risk communication frame on the robot platform.

[5] https://www.furhatrobotics.com

We assessed *risk perception* as the dependent measure, operationalised across *perceived severity* and elicited *negative affect* [211], before and after participants' exposure to the robot. Three data comprehension items assessing value interpretation and overall trend interpretation [104] were included as a robustness check. We further assessed the disposition to trust technology [107] and participant's tendency to anthropomorphise [167, 216] as covariates to account for individual differences.

An a priori power analysis was conducted in *G*Power* [57] to determine the minimum required sample size. To counterbalance scenario order, we arrived at a final sample of $N = 28$ participants. The recruitment was conducted via the institutional notice board. Quantitative data were analysed using CLMMs to account for people's anthropomorphisation tendencies and dispositional trust in technology as covariates, and to model the nested structure of the repeated scenario-level observations within participants, given that each participant responded to two scenarios in a pre-post study design. We analysed the qualitative data using a reflexive thematic analysis [23].

In Article IV, we conducted an *in-person* between-within subjects user study to investigate the perceptual effects of anthropomorphic embodied cues for uncertainty communication in robot-assisted decision-making. Specifically, we examined how the type of uncertainty communication and degree of uncertainty influence system trustworthiness and user decision-making, employing a 2 (Uncertainty Communication: Embodied vs. Graphical User Interface) $\times$ 3 (Degree of Confidence: Low, High, 100%) design.

The experimental task constituted a collaborative medication sorting task situated in a high-stakes factory context, in which participants worked collaboratively with a collaborative robot (Franka Emika[6]) to identify defective medication at a packaging station. The robot was responsible for displaying its level of confidence in sensing defective pill boxes and for packing those at-risk boxes, with participants providing support in packaging and retaining final decision-making authority. The task was implemented as an autonomous pick-and-place task via ROS2 [121], with a recorded video displayed prior to the main experimental phase to support familiarisation with the task and the robot platform, and to mitigate novelty effects [184].

The two uncertainty communication conditions were both risk-informed but differed in their communicative affordances. The embodied condition mapped uncertainty linearly onto the trajectory function of the robot arm, whereas the GUI condition deployed an icon array uncertainty visualisation via a Flask web application integrated with ROS2. The appropriateness of the experimental manipulation was confirmed through an a priori manipulation check. The manipulation check was designed as a between-subjects online study ($N = 60$) deployed on Prolific to verify that the experimental manipulation was perceived as intended. Each confidence level condition consisted of $N = 10$ assessment trials and we operationalised the degree of uncertainty through instance-based confidence levels consistent with established uncertainty indicators in human-AI interaction [152].

User *decision-making* (shipment vs. additional testing) served as a behavioural trust measure, with recorded *alterations in judgement* as an indicator of cognitive deliberation. *System trustworthiness* was assessed via a unidimensional measure, and the Multidimensional Trust Questionnaire (MTQ) across the dimensions perceived *system transparency* and *performance* [157]. Dispositional trust in new technology [108] was assessed as a covariate. Average confidence estimation was included as a robustness check.

An a priori power analysis was conducted in *G*Power* [56] to determine the minimum required sample size, informed by which we recruited $N = 36$ participants. We recruited participants through the institutional notice board. Quantitative data were analysed using a mixed analysis of covariance, controlling for dispositional trust in new technology [108]. We analysed qualitative data following a deductive-leaning reflexive thematic analysis [23].

[6] https://franka.de

1.3 Ethical Considerations

HCI as a research discipline is not just about *humans* interacting with systems, or designing interactive systems *for* humans, but fundamentally also about conducting research *with* humans [81]. A recurring ethical tension in the empirical work presented in this thesis concerns the very topic of risk communication and risk perception itself: to investigate how people respond to risk information, some degree of risk must be simulated, which requires careful consideration to avoid exposing participants to genuine harm or distress. Throughout this thesis, we applied several approaches to ensure that the research presented in this work was ethically conducted, and that ethical considerations around human participants guided each study. Across this thesis, we identified no risks to non-participants. We describe our risk assessment and mitigation strategies in the sections that follow. In-person research (Article III (Study I,II), Article IV) was conducted under the supervision of at least one person from the research team who had completed the University of Melbourne's Research Integrity Online Training prior to data collection. Each study underwent a comprehensive ethics review and received the approval from the University of Melbourne Human Research Ethics Committee (Ethics Application ID: 27586 (Article I), 33699 (Article II), 27391 (Articles III and IV)).

Informed Consent, Anonymisation, and Compensation. Prior to participation, all participants received a plain language statement elaborating on the research objectives, the expected study task, anonymisation procedures, data storage, and the potential benefits and risks of participation. Participation was voluntary, and participants were informed that they could withdraw at any time without consequence. Following the plain language statement, participants were asked to provide active consent before proceeding. All studies included a debriefing at the conclusion of participation, in which participants were informed of the true nature of the study, and that the task, stimuli, and associated consequences were fictional.

We refrained from collecting any personally identifiable information across the studies reported in this thesis beyond demographic information (age, gender) and individual difference measures, such as dispositional trust in automation, tendency to anthropomorphise, and AI literacy, which served solely to address the research questions and were reported at the group level or as model parameters, with no data attributable to individual participants. To further ensure anonymity and prevent any breach of privacy, individual participants were assigned pseudonyms (e.g., P01) in the data analysis and in all research reporting. With respect to participant compensation, we adhered to Prolific's payment standards[7] for all online studies (Article I and II) to ensure that participants were paid fairly and that our recruitment practices were non-exploitative. For on-site in-person studies (Article III and IV), participants were compensated in accordance with the Australian minimum wage. Participants who failed attention, recognition, or comprehension checks outlined in Section 1.2 continued to receive full compensation for their time.

Sampling Considerations and Cultural Context. In Article II, we restricted recruitment to US-based participants via Prolific's internal pre-screening criteria, given that the political self-placement measure deployed in this study was designed for a US population, and that constructs such as liberal and conservative orientation carry different meanings across cultural contexts. We explicitly acknowledge that the findings of Article II are situated within a US context and may not generalise across different political landscapes or cultural experiences of AI.

More broadly, the recruitment criteria applied across this thesis (see Section 1.2), including English language proficiency and platform approval rate thresholds for crowdsourcing studies, were guided by considerations of experimental comprehension and data quality. We report all recruitment criteria transparently and acknowledge that these criteria inevitably exclude other populations, which presents a limitation with respect

[7] https://researcher-help.prolific.com/en/article/9cd998

to the generalisability of findings across, for instance, cultural contexts. In Article III, our expert sampling was restricted to HCI researchers and practitioners to ensure relevant disciplinary expertise and diversity of perspective within the field, reflecting the design-oriented objectives of the workshop (Study I). We acknowledge, however, that the exclusion of other stakeholder groups, such as end-users or health experts, constitutes a limitation of this work.

Experimental Manipulation, Stimuli, and Apparatus. A central ethical tension in this thesis concerns the operationalisation of risk without exposing participants to actual harm. Manipulating situational risk is inherently challenging as participants must sufficiently engage with the scenario and its associated consequences for the manipulation to be effective, yet the experimental setup must adhere to ethical guidelines [78, 83]. Researchers in HCI therefore often manipulate situational risks through tasks that include loss or gain (e.g. monetary, time) for participants (eg. [51, 220]) or the task context in which the AI is embedded (scenario-based) (eg. [55, 145]). We carefully designed stimuli and task scenarios across all studies to avoid overly sensitive or distressing content, while maintaining sufficient realism to support ecological validity. In Article I, we manipulated situational risk through a scenario-based task in which participants assessed microscope images of bacterial infestation, drawn from either human tissue (high stakes) or plant tissue (low stakes). This stimuli manipulation has been validated in prior work [55]. To enhance the authenticity of the scenario, we included detailed contextual briefs, institutional logos (a medical centre; plant nursery), and generated images depicting either a human patient or a plant in the scenario briefs. We identified a minimal risk of distress arising from engaging with the associated context. However, the microscope images depicted bacterial coverage in a technical manner rather than graphic or sensitive content, and given the simple nature of the decision-making task and the clear explanation of the AI's role, the psychological impact was expected to be minimal. This risk was further mitigated through a clear communication of the task context in the plain language statement prior to the experiment, the ability to pause or withdraw at any time, and a post-study debrief which re-emphasised the fictional nature of the task and stimuli.

In Article II, the stimuli consisted of simulated human-LLM interactions across three everyday scenarios: seeking health advice, booking assistance, and support in a hiring process. Neither the AI risk disclaimer content nor the interaction scenarios were identified as potentially distressing, as they depicted situations participants are likely to encounter in their everyday lives. The scenarios remained non-graphic and did not introduce sensitive cultural or political content beyond what is commonly present in public discourse. In Article III, the stimuli for the design workshop (Study I) were drawn from an affective visualisation corpus [106] associated with health-related risks. We shared the stimuli with participants prior to the workshop so they were informed of the nature of the material before agreeing to participate. In Study II, the health risk information was encoded through anthropomorphic facial features deployed on the Furhat social robot platform[8]. The stimuli carried some potential to elicit discomfort. We mitigated this by avoiding overly provocative or sensitive imagery and by providing participants with information about the nature of the study prior to participation. Across Study I and Study II, participants retained the right to withdraw at any point and a debrief was administered following participation. We identified no risk of physical harm in either study, as the task was solely observational and the Furhat platform is not capable of physical movement beyond head motion.

In Article IV, the collaborative task involved a physically moving robot arm (Franka [9], 7 DoF), and risks related to the physical interaction were carefully assessed and mitigated. No physical contact between participants and the robot was permitted at any point during the study. The robot was positioned at an appropriate distance from the participant, with clearly marked safety zones delineating its workspace. Participants were instructed prior to the task not to touch the robot or violate the safety zone. The study was conducted in ac-

[8] https://www.furhatrobotics.com
[9] https://franka.de

cordance with the laboratory's Standard Operating Procedure, which included ensuring the workspace was de-cluttered and cleared. Throughout the study, a researcher remained in control of a release controller that could immediately lock the robot's joints and halt its motion in the event of an adverse incident.

For the in-person studies presented in Articles III and IV, we dedicated time prior to the experiment for participants to familiarise themselves with the robot, which is particularly important when engaging with novel embodied technology [184], and to introduce the laboratory environment to create a comfortable setting and provide the opportunity to ask questions.

1.4 Thesis Outline

We organised the remaining chapters of this thesis as follows.

In **Chapter 2**, we establish the theoretical foundations that ground and inform our work. We introduce key theories, frameworks, and concepts pertaining to risk and its perception, judgement and decision-making under uncertainty, and risk communication, before turning to how risk has been defined and treated in human-AI interaction, with particular attention to AI risk taxonomies, trust, and the communicative role of AI systems.

Chapters 3 – 6 present the four empirical articles that constitute the core of this thesis, each addressing the research questions introduced in Section 1.1. **Chapter 3** presents Article I, which investigates how communicating an AI's relational risk influences trust perceptions in AI-assisted decision-making. **Chapter 4** presents Article II, which examines how the framing of AI risk disclaimers shapes risk perception in LLM-based interaction. **Chapter 5** presents Article III, which explores anthropomorphism and embodiment as design mechanisms for risk communication through a social robot platform. **Chapter 6** presents Article IV, which investigates uncertainty communication in human-agent interaction through a collaborative robot task. Each empirical chapter opens with an introduction which briefly describes and situates the respective article within the thesis, followed by the article itself, and closes with a chapter reflection that draws out the key contributions and positions the work in relation to the research questions.

Chapter 7 synthesises the findings across all four articles in relation to the research questions guiding this thesis, and reflects on their collective contribution through an HCI lens. Drawing on this synthesis, the chapter puts forward a set of propositions to advance the design of risk communication in human-AI interaction. The chapter further addresses important societal and ethical considerations that arise from this research. Lastly, it concludes by drawing on the collective findings to outline future research directions through which the HCI community can extend these contributions and further advance the understanding and design of risk communication in human–AI interaction.

Chapter 2

Background

This chapter reviews the main concepts underlying this thesis to provide the foundations necessary to understand how people perceive risk, process information under uncertainty, and the challenges this poses for the design of risk communication in human-AI interaction. The chapter is structured into two parts, reflecting its twofold objective. In Section 2.1, we first outline foundational literature on risk, its perception, and risk communication, drawing on decades of evolving risk theory, judgment and decision-making research, and cognitive and psychological science. In Section 2.2, we then turn to more recent work on risk and uncertainty in the context of human-AI interaction, reviewing relevant taxonomies of AI risk, approaches to communicate uncertainty, and the interplay between risk perception and system trustworthiness.

2.1 Theoretical Foundations of Risk Perception and Communication

Risk communication is ubiquitous in everyday life. We encounter it on our way to work, passing campaigns warning about the consequences of tired driving, in the medication we take to combat migraines, and in risk communication during critical events such as natural hazards or pandemics. Risk information helps us make informed decisions and take protective actions. The study of how people *perceive* risk information experienced a pivotal moment when nuclear technology emerged as a major technological change [188, 206]. Contrary to prior assumptions by governmental organisations and regulators, the public attributed far greater risk to nuclear technology than experts anticipated, often weighing the perceived risks more heavily than the technology's utility. This demonstrated one of the central challenges in risk communication: a gulf between the actual risk information provided and how that information is perceived and interpreted by the individual. Similar to nuclear technology, artificial intelligence now presents an equivalent challenge, raising urgent questions around how risk information should be communicated to individuals, and how it is perceived.

Beginning with theoretical foundations of risk and its perception, this section reviews how risk is defined, conceptualised and how people form risk judgements, drawing on dual process theories, heuristics, and cognitive biases to review the psychological mechanisms underlying risk perception. We then review relevant literature on risk communication to identify its core components, and factors relevant to designing effective risk communication.

2.1.1 Defining Risk

Before we dive into the psychological mechanisms that determine how individuals evaluate risk information, and ultimately, how to therefore best present information about risk, we need to define risk to clarify the main

construct this thesis is concerned with. The construct of risk has been in the forefront for decades across a multitude of research disciplines, including but not exhaustively, social science, communication science, and psychology. It has been defined in an array of definitions, broad and narrow, which Breakwell [25] breaks down to two fundamental dimensions that make up the common denominator, and therefore fundamentally denote risk: *probability* and *effect*. According to Breakwell [25], probability describes the likelihood of an adverse event, whereas effect describes the consequence of the respective event, often attempted to be quantified through its extend or severity. In its essence, a consequence can be positive and negative. In this thesis, nevertheless, we focus on negative consequences, following the definition provided by Wilson et al. [221], describing risk consequences as "the negative outcomes to something of value given exposure to a hazard".

Returning to the very definition of risk, we note that not all risks can be reliably quantified through probability estimates, especially not in an emergent complex context such as AI. The definitional tension between uncertainty and risk has been debated for decades. Frank Knight, in his early work *Risk, uncertainty and profit* (1921) [99], later critically discussed by LeRoy and Singell [114], provides one of the earliest influential accounts of this distinction. Contextualised within insurance markets, Knight distinguishes between objective probabilities, those that are quantifiable and thus would be widely accepted, and subjective probability estimates, which are individual and subjectively derived estimates from personal judgement. According to Knight, the former defines a situation as risk [114], whereas the latter constitutes a defining element of situations characterised by uncertainty. The definition of risk, and its relation to uncertainty, has since been subject to considerable debate and reinterpretation, moving away from Knight's original framing toward broader conceptualisations in which uncertainty is understood as inherent to risk. To date, no unified definition of risk has been agreed upon, but its core components have persisted. The Society for Risk Analysis Glossary (SRA) [10], which has also been consulted to define AI risks [175], clearly demonstrates this unresolved debate. Rather than converging on a single definition, the SRA presents several qualitative definitions of risk across disciplines and contexts. Among these definitions, however, uncertainty and consequence emerge as the shared core components of risk. Notably, the SRA presents probability as a possible metric for operationalising uncertainty, rather than a definitional element of risk itself. This is consistent with ISO 31000 [84], which places uncertainty at the core of risk, defining risk as "the effect of uncertainty on objectives", with likelihood described as expressions of risk. Interestingly, the term likelihood has been deliberately used with the intention to move away from the mathematical narrowness of probability, and to acknowledge that it can be both objective and subjective in the context of risk. In this thesis, we define the two dimensions of risk as *uncertainty* and *consequence*.

Two relevant concepts, which are often used interchangeably with risk [25], are *hazard* and *harm*. Hazard describes the source of a risk [25], and harm occurs when a risk materialises [80]. To illustrate in the context of this thesis, interacting with an AI system, in the widest sense, presents the hazard, whereas the experience of a consequence due to an AI behaviour, for instance acting on false information presented by the AI, describes the respective harm.

2.1.2 Risk Perception

While it is useful to define risk as precisely as the literature allows, it is important to understand that risk information does not equal perception. The formulation of risk as two components of *uncertainty* and *consequence* might evoke a false sense of security, in that, as long as those two components are communicated, the risk will be perceived accordingly, and in that sense 'correctly'. However, in the case of risk, it could not be further from the truth. Risk perception is a multi-dimensional construct [210, 211, 221] that gained attention due to the fact that how we perceive risks often diverges from what the information objectively presents. One of the founding figures in risk perception research, Paul Slovic, goes as far as to describe that "there is no objective risk out there", as risk is determined by its perception, which is subjective by nature [178]. In the

subsequent sections, we review important determinants of risk perception relevant to grounding this thesis, beginning with how risk perception is measured as a latent construct, serving as a means to understand the construct itself.

Measuring Risk Perception The subjective nature of risk perception is reflected in the evolution of its measurement instruments. In 1969, Starr [188] examined the foundational question of "How safe is safe enough?" by investigating acceptable risk-benefit trade-offs to inform policy. In response, Fischhoff et al. [62] proposed a psychometric approach to assessing risk perceptions in 1978. Following the underlying theoretical assumption that risk is *subjectively perceived*, rather than merely the product of a probability calculation, the approach was later formalised as the 'psychometric paradigm' [177, 180, 181]. The paradigm describes a methodological approach to quantify risk perceptions based on a set of risk characteristics, such as the perceived controllability of the risk or the perceived severity of consequence [62, 177, 206], which are subjectively appraised by the individual. One of the objectives of this method is to generate "cognitive maps" of hazards by plotting their perceptual ratings along underlying risk dimensions, typically derived through a factor analysis [180]. The psychometric paradigm further recognises that an individual's risk perception can be shaped by a multitude of influencing factors, spanning psychological processes to social and cultural determinants [178].

More recently, Robyn S. Wilson and colleagues conducted a series of empirical studies [210, 211, 221] to develop an instrument to measure risk perceptions that is applicable across domains. Their work re-emphasises risk perception as a multi-dimensional construct in which each dimension constitutes a determinant of risk perception. Informed by a review of the existing literature, their initial works captured risk perception as *individual beliefs* about (a) the *likelihood* (b) the *severity of consequences*, and (c) the *affect* evoked by a hazard [221]. Their results revealed multiple insights that are particularly noteworthy in the context of this thesis. First, their findings suggest that the perceived likelihood is less important to people's holistic risk perceptions than the other two perceptual components, i.e. perceived severity of the consequence and affect. Secondly, their findings led them to propose a fourth dimension that compromises the perceived *vulnerability* in respect to the hazard, which was subsequently assessed and confirmed as an important fourth perceptual dimension by Walpole and Wilson [210], who re-termed it as *susceptibility* for greater conceptual precision. Thirdly, their findings suggest that the perceived severity of a consequence emerged as the strongest predictor of general risk perception, whereas the elicited affect emerged as the strongest predictor for the degree to which a risk feels personally relevant [210].

Their work offers a fundamental contribution to risk perception research, proposing that risk perception as a latent construct can be expressed as [210, 221]:

$$\text{Risk} = \text{Probability}(\text{Exposure} + \text{Susceptibility}) \times \text{Consequences}(\text{Severity} + \text{Affect}). \tag{2.1}$$

2.1.3 Judgement and Decision-Making about Risk

Before exploring how information, particularly information about uncertainty, is processed and interpreted, it is important to briefly review the dual process theory. With roots in philosophy and early psychology, the theory emerged in cognitive science through the early work of Wason and Evans [215] in the 1970s, and was subsequently advanced in the context of judgement and decision-making by Evans, Stanovich and colleagues (e.g. [54, 186, 187]), before being more broadly popularised by Kahneman [87]. The theory has since been drawn upon across a multitude of research fields, including risk perception literature [206] and HCI (e.g., [18, 19]), and recently has been extended by human-AI interaction research due to the advancing capabilities of AI systems [37, 38, 169].

Fundamentally speaking, dual process theory posits that there are two qualitatively distinct types of cognitive processing at play when forming judgements. It is widely accepted that one process is automatic,

operating quickly, and heuristically, without much cognitive effort, whereas the other is deliberative and places greater demands on an individual's working-memory [54]. These two processes have been referred to in various ways across the literature, most commonly as *System 1* and *System 2*, or *Type 1* and *Type 2* thinking, and more broadly as *intuition* and *reflection*. A stream of research has associated different attributes with each type, such as autonomous versus controlled, fast versus slow, and non-conscious versus conscious processing [54, 186]. It is worth noting that Evans and Stanovich [54] themselves prefer the Type 1 and Type 2 terminology over the System-framing, as the latter may incorrectly imply the workings of two distinct cognitive systems. More importantly, they argue that many of the attributes commonly associated with each type are correlates rather than defining features, and identify the involvement of working memory as the defining characteristic differentiating the two.

The dual process theory has also been applied to risk perception research, for instance in the scientific discourse around the experiential and the analytical system [183]. In a series of theoretical and empirical work, Visschers and colleagues [206, 208] describe them as primary and secondary evaluative processes, and, by doing so, draw attention to *risk associations* as a central mechanism in risk perception. The primary evaluative process denotes quick, intuitive evaluations of risk, rooted in the evolutionary human tendency to rapidly assess whether to approach or avoid a given risk, which the authors also refer to as "gut feeling" [207]. According to Visschers and colleagues, this process relies on *general* associations, that is, a broad approach-avoidance response that does not require prior knowledge of the specific risk. The secondary evaluative process denotes a more conscious and controlled effort to evaluate a risk by drawing on *specific* associations, provided that the individual has sufficient motivation, time, and cognitive resources. These specific associations draw on shared characteristics and semantic similarities with other risks, and consequently play a crucial role when interpreting unknown risks.

Heuristics In situations characterised by uncertainty, which often tend to involve some form of risk, people rarely have access to complete set of reliable information on which they can rely to make decisions, and even when they do, they tend to avoid the cognitively demanding effort of carefully assessing probability and consequence parameters. In their seminal work, Tversky and Kahneman [198] posit that instead of carefully calculating probabilities to form judgements about risks, people tend to rely on *heuristics*. These heuristics represent mental shortcuts or rules of thumb that simplify judgement but can in turn distort people's risk assessments, leading to systematic errors in the process. In their original work, Tversky and Kahneman [198] distinguish between three overarching cognitive biases resulting from people's use of heuristics in judging probabilities. *Representativeness*, which describes peoples' tendency to judge by similarity, in that, if characteristics appear to be similar, their probability to 'belong together' is estimated more highly. *Adjustment and Anchoring*, which describes people's tendency to judge a probability based on an anchor, from which on the adjustment happens, which can be an initial association or a concrete value. And lastly, *Availability*, which describes people's tendency to judge probabilities based on the ease with which they can be imagined, or recalled. Drawing on a collection of fundamental work on heuristics in judgement [88], Slovic et al. [179] contextualise many of the judgemental biases in the context of risk perception. For the purposes of this thesis, the availability heuristic and its related biases are particularly interesting. Tversky and Kahneman [198] describe multiple factors that can contribute to the availability heuristic, among them two that are of particular relevance here. Biases due to *retrievability* describe how easily a risk comes to mind, which can be influenced by, for instance, a person's familiarity with the risk or the salience of a risk, and *imaginability* biases, which describe the degree to which a person can envision a risk, which can be influenced by, for instance, how vividly a risk is depicted.

Affect heuristic Affect has long been overlooked in research on risk judgement and decision-making, despite "dread" being one of the major factors identified to characterise risk perception in early psychometric

research referred to in section 2.1.2 [181]. A programme of research by Slovic, Finucane and colleagues, alongside the influential framing of "risk as feelings" by Loewenstein et al. [116], sparked a substantial body of research into the crucial role of affect in risk judgement, well captured by the notion that "people base their judgments of an activity or a technology not only on what they think about it but also on what they feel about it" [183]. In the context of dual process theories, affect has been identified as a key driver for the experiential processing, associated with intuitive, heuristic-based judgements of risk [183]. Finucane et al. [60] coined the term "affect heuristic" to describe how, when confronted with a risk stimulus, affective associations can be triggered from what they term an individual's "affect pool." This affect pool contains both negative and positive feelings associated with the risk stimulus, and is drawn upon to form quicker judgements about a risk, particularly under conditions that constrain more deliberate risk evaluations. In a series of empirical work, the authors show that affect plays an important role in people's risk-benefit assessments and contributes to explain the inverse relationship between risk and benefit perceptions. *How* the risk stimuli - in our case any cue related to a risk - is presented, e.g. its affective salience, can determine the degree to which individuals *feel* rather than reason about the risk. This is reflected in the recent operationalisation of risk measures described in Section 2.1.2, which emphasises the importance of assessing affective responses, that is, the degree to which people worry about, feel concerned by, or fear a risk, and the degree to which information presented about a risk induces such affect [210, 221]. The important role of affect in people's judgement and decision-making has also permeated adjacent fields, with recent research in data reasoning and information visualisation increasingly acknowledging the role of affective responses in how people process and interpret visual information (e.g. [93, 106]). The question of how risk communication can be deliberately designed to account for, or appeal to, affect will be addressed in the subsequent Section 2.1.4.

Framing Theory So far, we have firmly established that people's subjective evaluation of risk can diverge from its objective information content by applying heuristics. The language we use to describe a risk, whether framed as a matter of life or death, gain or loss, can itself shape people's risk judgements and attitudes, and more importantly, provoke or attenuate the application of heuristics. Framing theory describes this as a "phenomenon of choice" [201]. It denotes that a different framing of information, even though rationally equivalent in its information content, affects the way people judge the information. It therefore contradicts the previously held assumption of description invariance [200]. The framing theory [199] is widely known through its example of mortality versus survival rates, in which the risk of, for instance surgery, can be either expressed through a mortality or a survival rate, which presents different framing strategies of the same information, and, by doing so, influences peoples risk-seeking or aversion attitudes. A closely related phenomenon is loss aversion, which describes people's tendency to subjectively experience losses (e.g. losing $100) differently compared to equivalent gains (e.g. winning $100), in that losses *feel* more substantial in people's judgements [201]. In response, Tversky and Kahneman [201] formulate the cumulative prospect theory, which, in essence, posits that probabilities are subjectively weighted, and that people evaluate outcomes relative to a reference point rather than in absolute terms.

As illustrated in the aforementioned example about survival and mortality rates, the way options are framed has received considerable attention in medical decision-making (e.g., [6, 127]). Interestingly, information framing has been identified as both one of the reasons for the "collective statistical illiteracy" in the context of medical decision-making [68], and as a remedy, with the potential to serve as a mechanism to overcome cognitive biases. It is important to note here that the traditional heuristics and biases work by Tversky and Kahneman has been criticised for relying on single-event probability formats, which are argued to obscure the probabilistic nature of the information [67, 68]. According to Gigerenzer [67], single-event probability formats, such as presenting a survival rate as simple percentages, conceals the conditional probability structure required for correct diagnostic inference. Natural frequency formats, by contrast, would make this structure accessible and attenuate cognitive biases, which is why work by Gigerenzer posits that the aforementioned cognitive biases are products of representational choice rather than fixed cognitive limitations [67, 68].

2.1.4 Risk Communication

The preceding section establishes the theoretical foundation for understanding why effective risk communication must account for how individuals perceive risk, judge information, and make decisions through heuristics, cognitive biases, and affective processes [206]. Before we dive further into the topic of designing effective risk communication, it is important to define what this thesis understands as *risk communication*.

The Society for Risk Analysis Glossary defines risk communication as an "exchange or sharing of risk-related data, information and knowledge between and among different target groups (such as regulators, stakeholders, consumers, media, general public)" [10]. While this definition captures the technical and informational dimension of risk communication, its objective often extends beyond objectively informing to actively *shaping* how people perceive, evaluate, and respond to a risk. It is thus often persuasive in nature [9], as one party (e.g. a governmental institution or a health organisation) seeks to shift another's beliefs and attitudes about a risk, often motivated by the fact that people may have limited resources to make informed decisions or to take protective actions [190]. Fischhoff illustrated this point vividly in a 1995 essay, describing the phases of risk communication as evolving from "All we have to do is explain what we mean by the numbers" to "All we have to do is treat them nice", concluding that risk communication ultimately necessitates "All of the above" [61], demonstrating that effective risk communication is not a straightforward but rather a multifaceted problem that continues to evolve.

A stream of research in the 1990s identified differences in mental models as a fundamental problem in risk communication (e.g., [9, 20, 21, 22]). In simple terms, to design effective risk communication, the communication must meet the individual's understanding of and relationship to a specific risk. Mental models in the context of risk are build on cognitive schemas [25], which comprise beliefs and the individuals' understanding of the risk. The mental model approach to risk communication aims to complete the risk information, to correct information accuracy by reducing misconceptions about the risk, and to make the information specific enough to enable individuals to formulate informed decisions about risks [9]. Mental model approaches have been employed, for example, to assess how people conceptualise risks associated with climate change and to inform risk communication accordingly [22].

Breakwell [24] described three important pillars of risk communication: First, the *source*, which describes from what or whom the risk communication stems or originates. The source can be any entity perceived as sending the risk information, e.g. governmental institutions, health organisations, or the GP. Second, the *message*, which describes the structure, design and content of the information itself, to which we return in greater detail as it poses the focus of this thesis. And lastly, the audience, which alludes to the fact that risk communication needs to account for the *audience* receiving the message. In fact, individual and group differences have been at the forefront of risk perception research for decades, ranging from prior experience with a risk [92] to the influence of people's ideological orientation on risk perceptions [39]. Slovic [178] and Breakwell [25] both provide comprehensive overviews of the breadth of individual factors that shape risk perceptions. It is important to note that, while individual factors interact with risk perceptions, they do not constitute components of risk perception as a construct [210].

In addition to the three factors source, message, and audience [24], Breakwell [25] later also describes the message *medium* as an important pillar, referring to the means through which a risk message is presented to an individual. Traditional message media include news articles, television commercials, and pamphlets placed in doctors' offices, while technological advances have expanded these channels to include social media and AI systems, the latter of which are conceptualised as the message medium in this thesis. The inclusion of medium as a relevant factor of risk communication is also reflected in early communication theories, such as the mathematical theory of communication put forward by Shannon [168], or Lasswell [109], who described the act of communication as "Who Says What In Which Channel To Whom With What Effect."

Designing effective Risk Communication Narrowing the lens to the message content, and how risk information is represented, it is important to re-emphasise that, ideally, the communicated risk captures its two core components - *uncertainty* and *consequence* [25]. As the preceding review of heuristics and cognitive biases has demonstrated, probability information proves especially difficult for people to interpret and reason about correctly. This has important implications for the design of risk communication, as the way uncertainty is represented, for instance through the choice of format, constitutes a key design consideration for effective risk messages. Visschers et al. [209] reviewed several decades of empirical research on how probability information is presented in risk communication, from the 1980s to the mid-2000s. Their work provides an overview of probability representation formats used in risk contexts (e.g., percentages, frequencies, cumulative probabilities) and formulates recommendations for representing uncertainty in risk message, such as to use a consistent denominator across a risk message to preserve people's ability to compare probability information, even if they are neglecting the denominator and relying on the numerator as an anchor (see anchoring heuristic, Section 2.1.3).

Another design mechanism that is drawn upon in risk communication is the deliberate induction of affect. We reviewed affect as a heuristic beforehand, and it is unsurprising that *affect cues* can be used to influence risk perceptions, and responses. An individual's affective response can be elicited through various design choices concerning both the content and representation of a risk message. The work by Visschers et al. [205] distinguishes between *incidental* and *integral* affect cues. Importantly, their distinction is less about the communicative intent, but refers to the design of the risk message. Cues that are considered incidental induce or inhibit affect implicitly through the risk message format or context of the communication. According to Visschers et al. [205] examples include the use of linguistic cues such as technical jargon or the utilisation of specific probability presentation formats. Slovic et al. [182], for instance, demonstrated that when the risk of patient violence is expressed in frequency formats it is perceived as more threatening than equivalent probabilities expressed as percentages, attributed to the affective images that frequency formats more readily conjure, as imagining a concrete number of individuals who may turn violent (X out of Z patients) triggers stronger affective associations than an abstract percentage (Y%).

When affect is instead evoked through how the risk is directly described or depicted, i.e. the design of the content of the message, Visschers et al. [205] term this *integral affect induction*, with *fear appeals* representing a common overarching design strategy [161, 205]. Fear appeals can be realised through a range of mechanisms, such as narrative risk descriptions or risk stories that recount a risk scenario or provide evocative anecdotal examples of its consequences [77, 205] and vivid imagery such as photographs [92, 131] or videos [33] that depict risk consequences.

In 1989, Hendrickx et al. [77] demonstrated that narrative, scenario-based risk descriptions dominated statistical frequency information in people's risk judgements, with the perceived concreteness of the scenario emerging as a key factor in this effect. Notably, concrete and vivid narratives not only elicit affective responses but also render the risk more cognitively available, as the scenario is easier to mentally construct or retrieve, pointing to an important connection between integral affect induction and the availability heuristic [92], the latter of which we reviewed in a previous Section 2.1.3. A broad body of work on affect-inducing risk messages, spanning traditional risk perception research [60, 92, 161, 179, 205] and application domains such as natural hazard communication [131, 190] and medical decision-making [33], demonstrates that deliberate design choices, ranging from simple representation formats to narrative framing and vivid imagery, can draw on both affective and cognitive mechanisms to shape risk perceptions. Going back to Visschers et al. [205]' distinction, it should be noted, however, that the boundary between incidental and integral affect induction is not always self-evident and can be blurry in practice, particularly when design decisions about format and content are intertwined. Nowadays, risk communication campaigns draw explicitly on these insights, employing affect-laden imagery, narratives, and risk stories to appeal to people's risk judgements. Familiar examples include anti-smoking campaigns that depict the consequences of smoking in-situ through graphic

imagery on cigarette packages, or road safety campaigns on highways that recount the stories of lives lost or the experiences of surviving relatives.

2.2 Risk in Human-AI Interaction

Having established the foundations of risk perception, and risk communication, the second part of this chapter narrows the lens to human-AI interaction research. It reviews existing definitions and taxonomies of AI risk, established approaches to uncertainty communication, and models of trust in human-automation and AI interaction, thereby establishing the conceptual foundations for user-facing risk communication in this thesis. While the broader risk literature has long examined how people perceive and respond to risk, research in human-AI interaction has predominantly centred around the construct of trust. We therefore give particular attention to how trust models have evolved from interpersonal trust research, through human factors and human–automation research, to recent models developed specifically for human–AI interaction. The chapter further examines the interplay between risk and trust that emerges across this body of work.

The introduction of deep learning has brought a fundamental shift in the nature of system behaviour and its associated risks. Whereas traditional software and automation systems were characterised by determinism, explicit rule-based logic, and predictable behaviour, AI system outputs are probabilistic, learned from data, and thus inherently uncertain. This shift introduces unique challenges for human-AI interaction and risk communication respectively. Errors may be difficult to detect, outputs can be presented confidently albeit erroneous, and users need to be supported in gauging system limitations and failability. Within the broader landscape of AI systems, the nature of these challenges varies considerably. AI classifiers, such as decision-support tools applied to, for instance, detect skin cancer or flag fraudulent transactions, are probabilistic and data-driven, but remain narrow in the sense that they are task-specific, with bounded failures modes (false positive, false negative [30]), and quantifiable uncertainty. Generative AI systems, by contrast, are task-agnostic and open-ended in the sense that they can produce a wide variety of outputs across many different tasks and domains, and their failures are correspondingly diffuse and difficult to detect, with uncertainty innate to the system. More recent developments in increasingly agentic system introduce additional layers of opacity and complexity due to the system's increased autonomy and agency, with users being able to delegate tasks with little user specification, systems acting in open-ended environments with little user oversight, and the system's ability to achieve a broad-range of tasks with little dependencies [35].

As AI systems have become more capable and widely deployed, understanding and managing their risks has become a central concern across research, policy, and industry. Living databases such as the AI Incident Database [125], which aggregates AI incident reports from media, journalism, community submissions, and academic research, provide a continuously updated empirical record of AI-related harms as they occur in the real world, and have served as a foundation for further classification efforts such as the AI Incident Tracker [132]. Browsing these databases illustrates the sheer variety of AI risks observed in practice, ranging from accidents caused by autonomous vehicles, to wrongful arrests resulting from facial recognition systems, to large language models providing harmful health advice. What these databases make evident is that AI risks are not straightforward but depend on a complex interplay of factors. Among them is the type of AI system, that is, whether it is a more traditional classifier addressing a binary classification problem, an embodied AI system navigating the physical world, or an agentic system acting in open-ended environments. The risk landscape is further shaped by the deployment context and its associated stakes, for instance whether the AI system is deployed to recommend music, to identify skin cancer, or to sell items on a marketplace, as well as the role the AI system holds, whether it is used to inform decisions, seek information, deployed as a collaborator, or leveraged as a user proxy or surrogate.

2.2.1 Defining and Classifying AI Risks

Before examining the landscape of AI risks in greater detail, it is worth establishing what is understood by the term *AI risk* itself. The work conducted by the MIT Risk Initiative emphasises the inconsistency of terminology and classification when it comes to AI risks across research, policy and regulations, and industry [176], and emphasises that in order to collectively tackle AI risks across disciplines, there is need for a common language to describe the AI risk landscape. A useful starting point is the distinction drawn by the AI Incident Database [125], differentiating between AI incidents and AI issues, a distinction borrowed from prior risk assessment practice in aviation. An AI incident refers to a harm or a near harm, where the latter describes circumstances that were contained before causing serious damage. An AI issue, by contrast, describes a situation where a risk of harm exists, but has not yet materialised. As established in Section 2.1, this distinction maps onto the broader conceptual difference between harm as the manifestation of a negative event, and risk as the potential for the manifestation of a negative consequence. Risk, as defined beforehand, is inherently a combination of uncertainty and consequence, and this is reflected in the definition adopted by the MIT AI Risk Repository [175], derived from the Society for Risk Analysis [10], which defines AI risk as "the possibility of an unfortunate occurrence associated with the development or deployment of artificial intelligence."

Efforts to identify AI risks and establish a common language have led to the development of a growing number of taxonomies and frameworks, predominantly aimed at informing policy and regulation. These taxonomies vary considerably in their scope and focus. Some differ in the specificity of the AI system they address, from more narrowly focused taxonomies targeting specific model types such as large language models [217], to broader ones addressing generative AI [59] or general purpose AI systems [175, 202]. A number of taxonomies adopt a sociotechnical or systemic lens, examining how AI risks manifest across social groups and societal systems [170, 202], or provide frameworks for the safety evaluation of AI systems [218]. Others concentrate on specific factors relevant to the "pathways of dangerous AI," facilitating the assessment of *when* and *how* an AI system became harmful [223].

A fundamental step forward in establishing a shared language for AI risk is the AI Risk Repository, led by Slattery et al. [175], which synthesises a total of 74 existing frameworks into two complementary AI risk taxonomies. The first presents a *causal taxonomy*, which classifies the causality of an AI risk occurrence along three dimensions, each with an additional "other" category to account for cases that are not clearly classifiable: the *entity* dimension captures the source from which the risk originates due to an action or a decision of that entity, distinguishing between human and AI system; the *timing* dimension distinguishes between whether a risk occurred pre- or post-deployment; and the *intention* of the entity denotes whether the resulting risk was intentional (expected or anticipated) or not. As such, the causal dimensions address the questions of by *whom/how, when*, and *why* an AI risk arises. The second taxonomy is termed *domain taxonomy*, which adopts a consequence lens and captures what is more commonly associated with 'types of AI risks'. It organises AI risks into seven overarching domains, each with more detailed sub-domains: *Discrimination and Toxicity*, *Privacy and Security*, *Misinformation*, *Malicious Actors and Misuse*, *Human-Computer Interaction*, *Socioeconomic and Environmental Harm*, and *AI System Safety, Failures and Limitations*. The domain taxonomy builds on and extends the taxonomy of risks posed by language models developed prior by Weidinger et al. [217], which identified six categories: *Discrimination*, *Hate Speech and Exclusion*; *Information Hazards*; *Misinformation Harms*; *Malicious Uses*; *Human-Computer Interaction Harms*; and *Environmental and Socioeconomic Harms*.

Consider a person using a large language model to seek health advice. From a domain perspective, this scenario carries the risk of *Misinformation*, as the system may provide inaccurate health information, and the risk of *Privacy and Security*, as the system may inadvertently expose confidential health information shared by the user. The example illustrates a case in which the risk stems from the *AI as the entity*, *unintentional* in nature, and occurring *post-deployment*. By contrast, a case where a malicious actor deliberately exploits a

publicly available LLM to generate and spread harmful misinformation would be classified as a *human entity*, *intentional*, *post-deployment* risk, with *Malicious Actors and Misuse* as its primary risk domain.

2.2.2 Risk Communication in Human-AI Interaction

Within HCI research, efforts to address AI risk span several areas, including *reporting* and *documenting* existing risks, *anticipating* potential risks, and *communicating* risks to users post deployment. Reporting, documenting, and anticipation efforts have predominantly focused on developers and deployers, whereas post-deployment approaches include user-facing interventions such as AI disclosures, disclaimers, and uncertainty communication. It is worth noting that these areas are not mutually exclusive, and many empirical contributions span multiple parts of this landscape. Together, they provide the broader context within which this thesis situates its focus on user-facing risk communication.

From Reporting to Communicating AI Risks to End-users Reporting and documenting efforts have focused on producing structured artefacts that make AI risks transparent to developers and deployers. Many of these approaches build on the broader idea of AI nutrition labels by presenting relevant system characteristics, risks, and limitations in an accessible and standardised format. RiskCards [48], for example, document relevant characteristics and risk categories for LLM-based systems. Impact Assessment Cards [17], developed for classification and recommender systems, provide a broader assessment that includes expected benefits, identified risks, and corresponding hazard levels. RiskRAG [151] offers a more detailed reporting structure by documenting exemplary model uses, summarising risks associated with each use case, and outlining corresponding mitigation measures. In the context of agentic systems, Chan et al. [36] discuss mechanisms such as "agent identifiers" to make AI agents and agent characteristics more visible for developers and deployers. Beyond documenting known risks, a central challenge lies in enabling individuals to anticipate risks that have not yet materialised. Anticipating AI risk has been addressed primarily at the development stage, with tools designed to help developers foresee potential harms prior to the deployment of AI systems. Wang et al. [214] developed Farsight, a graphical interface designed to support the anticipation of potential harms during LLM prototyping. The tool enables developers to explore possible downstream use cases and misuse scenarios, to identify affected stakeholders and associated harms through a dedicated dashboard, and to connect model outputs to relevant news articles illustrating comparable real-world harms. In the context of AI classification systems, Buçinca et al. [30] propose AHA!, a tool that supports the anticipation of potential harms associated with an AI deployment. The tool generates illustrative risk scenarios from a matrix that maps different stakeholders against potential false-positive and false-negative system behaviours.

Although efforts to anticipate AI risks have predominantly focused on the development stage, the need to make potential risks foreseeable extends to users once AI systems are deployed. User-facing risk communication aims to support users in anticipating potential harms and forming appropriate perceptions of the system. Recent risk communication efforts range from AI disclosure labels that signal AI presence and use, as mandated by the EU AI Act Article 50 [52], to AI disclaimers adopted across industry that suggest potential AI fallibility and encourage users to verify AI outputs, such as those deployed in ChatGPT [136] and Claude [5], and more recently examined as "reliance disclaimers" in human-AI interaction research [16]. Some empirical work investigated the effects of AI disclosures on the perception of AI-generated content [193], or on the perception of the entity that used the AI system [26, 164]. Comparatively little work, however, has examined the design of such user-facing AI risk communication or their effects on users' perceptions of the AI system itself, including its trustworthiness and associated risks [16].

Importantly, the design of user-facing risk communication depends on the system through which users encounter it. In human-AI interaction, the type of AI system determines the interface through which users interact with it, and the interface in turn sets the affordances available for risk communication. This contrasts

with traditional risk communication, where the *source*, *medium*, and *message* are often distinct. For example, a health institution may mandate the communication of information about a seasonal influenza vaccination, which is then presented through a pharmacy webpage encountered by an individual while grocery shopping. An LLM presented as a conversational chatbot interface, a computer vision classifier embedded in a decision-support dashboard, or a robotic system navigating a physical environment each afford very different interaction paradigms and, consequently, very different design opportunities and constraints for risk communication. This gives rise to an important human-AI interaction challenge, that of *in-situ* risk communication, where risk information must be conveyed to users in the moment of interaction, embedded within the very system that often poses the risk itself.

Uncertainty Communication in Human-AI Interaction In Chapter 2.1, we established that risk, from a theoretical standpoint, comprises two components, *uncertainty* and *consequence*, both of which ought to be communicated to users [25]. A considerable body of literature on uncertainty communication has emerged within Explainable AI (XAI) research, emphasising that uncertainty, as an inherent property of AI systems, should be made transparent to support the formation of appropriate mental models and, consequently, appropriate system trust [12]. In AI decision-support, uncertainty has commonly been communicated through AI confidence scores presented alongside individual outputs as a local explanation strategy (e.g., [32, 152, 226]), or through AI accuracy as a holistic performance indicator and global explanation strategy (e.g., [76, 138, 225]). Both metrics are frequently expressed as simple percentages. However, mixed and inconsistent findings have raised questions about the effectiveness of such numerical representations. The concerns related to the limitations of numerical representations align with established research on "collective statistical illiteracy" [68] and the difficulties individuals face when incorporating numerical information into judgement and decision-making [147]. As established in the preceding Section, these limitations can increase reliance on heuristics and contribute to systematic cognitive biases. Therefore, recent work in XAI shifted towards representational formats beyond percentage-based expressions. For instance, presenting AI confidence in frequency formats has been suggested to reduce confirmation bias [32]. Other work conducted by He et al. [76] examined everyday analogies as supplementary descriptions of AI accuracy, intended to communicate system uncertainty more accessibly by activating familiar associations, such as those related to weather forecasts.

Whereas uncertainty can be operationalised through quantifiable metrics for classification systems, uncertainty is less directly quantifiable for LLM systems. This necessitates the identification and design of alternative representational formats and communicative mechanisms for conveying uncertainty. In line with the natural language affordances of such systems, Kim et al. [96] investigated linguistic cues and the use of generic (e.g. "There is uncertainty, but it seems like…") versus first-person (e.g. "I am not certain, but I would guess…") narrative uncertainty expressions. Their findings suggest that embedded linguistic uncertainty cues may serve to improve appropriate system reliance. More recently, empirical work conducted by Bo et al. [16] investigated so-called "uncertainty highlighting" as a design mechanism to communicate the uncertainty of LLM outputs. In their design, specific words within the LLM output were highlighted in colour to reflect token-level uncertainty, with the degree of colour opacity, ranging from light to strong pink, indicating the degree of uncertainty associated with each token. However, this approach has been found to be insufficient as a means of calibrating human trust [16], compared to, for instance, an AI disclaimer warning users to verify the information.

2.2.3 Trust and Risk in Human-AI Interaction

Trust in human-AI interaction has received considerable attention in the research community, with early work rooted in human factors and human-automation research. In this section, we review foundational trust models across relevant disciplines and illuminate where risk is situated within them. We further describe

trust as a latent multidimensional construct, and present a recent model of human-AI trust that centres perceptions of system trustworthiness, which is of particular relevance to this thesis given its focus on perceptual trust dimensions. Across this thesis, we define trust following Lee and See [111], who describe trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability", which, since then, has been frequently adopted as the guiding trust definition in human-AI interaction.

Historically, research on trust in human-automation interaction has been investigated through a human factors lens, driven by operational incidents in aviation and related domains that exposed the consequences of miscalibrated trust between operators and automated systems. These incidents were not confined to a single failure pattern but encompassed a range of breakdowns in the human-automation relationship, from operators placing excessive reliance on autopilot systems and neglecting active oversight, to the disregard of automated alerting systems. Importantly, these incidents provided the empirical impetus for early theoretical work on how trust, when misaligned with a system's actual capabilities, produces consequential failures.

In 1997, Parasuraman and Riley [139] provided an early and influential conceptualisation of these failures modes, framing them as consequences of *inappropriate trust* in human-automation interaction. Their framework identifies three forms of *trust miscalibration*. The first, termed *misuse*, refers to circumstances in which "excessive trust can lead operators to rely uncritically on automation without recognizing its limitations or fail to monitor the automation's behavior", a form of overreliance the authors attribute to decision biases and failures of human oversight. The second, *disuse*, describes the inverse pattern in which operators ignore a system and its capabilities in circumstances where use would be warranted, which the authors contextualise within the human factors challenge of false alarm rates in automated alerting systems, where poorly calibrated decision thresholds can lead operators to distrust and disregard the system. The third, *abuse*, sits somewhat apart from the preceding two in that it does not describe an operator failure, but rather circumstances in which automation is applied without appropriate consideration for the human operator, reflecting an automation-centric approach that disregards human factors.

Building on Parasuraman and Riley [139], Lee and See [111] later situated these failure modes more explicitly within a framework of system capabilities, reconceptualising (in)appropriate trust as a function of the degree to which trust *(mis)matches a system's capabilities*. Their work distinguishes between *overtrust*, in which trust surpasses a system's actual capabilities and gives rise to misuse; *distrust*, in which trust falls below what a system's capabilities would warrant, producing disuse; and *calibrated trust*, in which trust aligns with system capabilities and supports appropriate use.

The idea that trust should match a system's capabilities has motivated decades of subsequent research, from human factors to human-AI interaction, and remains particularly consequential in situations where miscalibrated trust carries significant risk.

The Role of Risk in Traditional Trust Models This section turns to the question how risk is situated within established theoretical trust models. While some of these models originate in human-human interaction or have their roots in human-automation research, they can be, and often are, applied to the context of human-AI interaction. As the central concern of this thesis is risk communication in human-AI interaction, we review these trust models through the lens of risk, with particular attention to the conceptualisation of risk within each framework, rather than offering an exhaustive description of all theoretical dimensions these models address. In the following, the trustor denotes the entity extending trust, and the trustee the subject of that trust. For the purpose of this thesis, the user assumes the role of the trustor and the AI system that of the trustee.

In 1995, Mayer et al. [122] developed the *integrative model of organisational trust*, which is rooted in interpersonal trust, but has been the foundation for trust models and research in human-AI interaction [165,

189]. In their model, the trustor's perception of the trustee's ability, benevolence, and integrity, collectively referred to as the "Factors of Perceived Trustworthiness", combined with the trustor's general propensity to trust, determine the level of subjective trust held toward the trustee. This level of subjective trust is then compared against the *perceived risk of the situation*. The perceived risk is thus positioned as a situational factor that functions as a threshold gateway between subjective trust and behavioural trust, and which the level of subjective trust must exceed in order to lead to the manifestation of trusting behaviour. Translated to human-AI interaction, this means that only when the user's subjective trust in the AI system surpasses the perceived risk, e.g. the decision stakes associated with the task, will trusting behaviour (e.g. delegating a task to an AI agent or acting on an LLMs system's output) follow. The conceptualisation of risk in Mayer et al. [122]'s model is somewhat in contrast to the trust model developed by Lee and See [111], which situates trust in human-*automation* interaction. In the latter, risk is an integral part of the very definition of trust, due to trust being characterised by uncertainty and vulnerability. Beyond this foundational role, perceived risk here too represents a contextual factor that influences the formation of the intention to rely on automation, which sits between what Lee and See [111] term trust evolution and reliance action.

Foundational work conducted by Hoff and Bashir [79] proposes a three-layered framework of trust in human-automation research. Their work distinguishes between different sources of trust in human-automation interaction across dispositional trust, situational trust, and learned trust. Within this framework, the authors describe dispositional trust as orthogonal to contextual factors and of the system itself, representing a human's stable predisposition to trust an automation system. In contrast, situational trust is most simply described as in-the-moment trust, shaped by contextual factors and factors of the human user themselves (e.g. current expertise). Lastly, learned trust originates from the human's initial learned trust prior to the interaction due to, for instance, pre-existing knowledge with a system, and the dynamically learned trust, which evolves during interaction in response to system characteristics such as design features and the nature of system failures. Notably so, perceived risk is understood as a contextual factor, referred to as "external variability", and directly influences situational trust between the human and the automation system, alongside the influence of other contextual factors such as task difficulty and task framing.

Across all three models, perceived risk is treated predominantly as a situational factor, either feeding directly into the formation of situational trust, as in Hoff and Bashir [79], or sitting between trust and its behavioural expression, moderating whether trust converts into behavioural expressions, as proposed by Mayer et al. [122] and Lee and See [111].

Trust as a Multi-Dimensional Construct determining System Trustworthiness Having examined how risk perception is conceptualised and measured in Section 2.1.2, the same question is now turned to trust, which, much like risk perception, has been theorised as a latent, multi-dimensional construct whose determinants and dimensionality have been the subject of considerable theoretical debate [100, 110, 111, 122]. Understanding how trust is measured is inseparable from understanding what it is constituted by, and thus what drives its perception. The foundational work by Lee and See [111], rooted in early empirical works by Lee and Moray [110], identified three dimensions that constitute trust between humans and machines. The first dimension is termed *performance* and refers to the way a user perceives system capabilities in regard to a task-specific goal. The second denotes the perceived *process*, which describes an individual's perception of the appropriateness of the system's behaviour in order to achieve an objective; and lastly, the perceived *purpose* of the system, which denotes the degree to which the system designer's intent is perceived as benevolent and oriented toward the user's interests. Importantly, Kohn et al. [100] situate these three dimensions as a direct operationalisation of the factors of perceived trustworthiness in the human-human trust model by Mayer et al. [122], in that performance, process, and purpose map onto ability, benevolence, and integrity respectively, translating them from a human-human to a human-automation context. As such, Kohn et al. [100] emphasise that these dimensions constitute *perceptions of system trustworthiness-related characteristics* rather than 'trust'

as a subjective attitude itself. This has broader implications for how trust is measured. Kohn et al. [100] argue that self-report measures of trust, by virtue of asking participants to reflect on their perceptions of a system, are in practice largely capturing *system trustworthiness perceptions*. More recent work has extended these trustworthiness dimensions to account for the increasing agency and opacity of advancing AI systems. In an effort to design a multi-dimensional measure to assess people's trust in human-AI interaction, Roesler* et al. [157] argue that, as systems move beyond rule-based automation toward higher degrees of autonomy, a user's perceived transparency of an AI system gains relevance as an important constituent of trust. As a result of their validation efforts, their multi-dimensional trust questionnaire encompasses four dimensions of trust in AI systems, i.e. *transparency*, *purpose*, *performance*, and perceived system *utility*.

System Trustworthiness Assessment and Risk-Benefit The trust models discussed thus far provide foundational perspectives on the pathways of trust in interactions between human and machines, and how risk is situated within that process. More recent work in human-AI interaction has shifted attention more explicitly toward perceptions of AI system trustworthiness and their relationship with risk, which is of particular relevance to the present thesis.

Motivated by the need to expand on the role of risk in existing trust models for human-AI interaction, Stuck et al. [189] propose a framework that distinguishes three types of risk. In line with prior trust models, their model describes *perceived situational risk* as a contextual factor [111, 122], representing a threshold that subjective trust must surpass in order to manifest as behavioural trust. Further, their model elaborates on the individual's *risk-taking propensity*, which directly influences people's behavioural trust as a dispositional determinant. Most importantly though, their model of trust extends traditional work by describing an individual's perceived *relational risk*, which reflects risk perceptions directed at *the AI system itself*, and which has a bi-directional influence on subjective trust, in that the perceived relational risk of an AI system influences the subjective trust and vice versa.

Other recent work in human-AI interaction has moved toward a narrower focus on perceptions of system trustworthiness and its interplay with risk perceptions. Schlicker et al. [165] identify the discrepancy between a user's *perceived* and *actual* system trustworthiness as a crucial determinant of trustworthy AI. As a response, the authors developed the Trustworthiness Assessment Model (TrAM), building on the organisational trust model proposed by Mayer et al. [122]. To understand how this discrepancy arises and how it might be overcome, the model introduces the concept of *cues* at its heart, denoting information utilised for a user's system appraisal, and described as the interface between actual and perceived system trustworthiness. Trustworthiness cues may take the form a displayed uncertainty indicator as reviewed in Section 2.2.2, explanations of system outputs, or user interface design elements - or, as in this thesis, user-facing risk communication. As their interpretation is inherently subjective, so too is the resulting perception of system trustworthiness. The TrAM conceptualises the assessment of these cues across two levels, the micro level as the psychological process through which individual users assess system trustworthiness, and the macro level as the systemic propagation of those assessments across broader sociotechnical contexts. In detail, at the micro level, the model describes how users assess a system's actual trustworthiness through cues on two sides - the system and human side. On the system side, the gap between the perceived and actual system trustworthiness depends on a cue's *relevance* (the degree to which a cue is "indicative" of the actual system trustworthiness, such as AI accuracy as a performance indicator of the AI system) and cue *availability* (whether cues are detectable and accessible to the user). On the human side, the accuracy of a user's AI trustworthiness assessments depends on cue *detection* and their *utilisation*, describing whether the user notices, appropriately interprets, and assigns importance to a given cue. However, cue assessment alone does not determine how trustworthiness is perceived. Besides this mechanism, the model positions *individual standards* as an equally central factor, in that each user holds a personal set of requirements for what makes a system trustworthy, which, in turn, determines which cues are sought and how they are weighted. At the macro level, the model further ac-

knowledges a coined *trustworthiness propagation process*, whereby assessments do not occur in a vacuum but spread across stakeholders within a sociotechnical system, with secondary cues produced by actors such as certification institutions or media further shaping individual trustworthiness assessments.

Returning to the role of risk within this model, an empirical follow-up study offers further insight into how risk operates in users' trustworthiness assessments. To empirically ground the TrAM, Schlicker et al. [166] examined users' AI trustworthiness assessments in the context of a large language model-based virtual doctor. Their findings reveal that users engage in risk-benefit appraisals as a central strategy for evaluating system trustworthiness, suggesting that risk is not merely a contextual factor but an active component of the trustworthiness assessment process. The authors propose that this risk-benefit assessment may be related to users' individual standards, though the precise nature of this relationship remains to be further investigated and specified.

2.3 Summary

The literature reviewed across risk perception and communication, judgement under uncertainty, and decision-making establishes the theoretical foundation of this thesis. Accordingly, we conceptualise risk as constituted of two components, *uncertainty* and *consequence*, while recognising that the communicated risk information does not translate directly into perceived risk. Heuristics and cognitive biases shape how individuals evaluate risk information, and different perceptual dimensions further determine how risk information is subjectively experienced. Crucially, the manner in which information is presented can distort perception and judgement, but it can also be designed to evoke these effects deliberately or to mitigate them. The landscape of AI risks is broad and evolving, and we narrow our focus to post-deployment strategies, where risks must be communicated to users directly at the point of interaction. Within this landscape, human-AI interaction has positioned trust as the central construct for evaluating the effectiveness of such strategies toward appropriate system use, with uncertainty communication constituting one prominent stream of this work. *Risk communication* constitutes such a post-deployment strategy, but remains scarcely investigated in human-AI interaction. We further position both *the message* and *the medium* as central components of risk communication, given that different AI systems afford different communicative possibilities, and it is precisely this intersection that this thesis addresses. Bridging established risk research with human-AI interaction thus enables this thesis to investigate the design and perception of in-situ user-facing risk communication.

Chapter 3

How Risk Communication Shapes Perceived System Trustworthiness in AI-Assisted Decision-Making

3.1 Introduction

The user's ability to anticipate AI harm and to adjust their system perceptions accordingly is crucial when collaborating with AI systems in assisted decision-task across contexts. Similar to everyday risks and tools that create them, risk communication in AI should warrant caution and foster deliberation, without undermining the tools' utility or creating and avoidance response - just as communicating the risks of a power drill is not to discourage a person from using it, but to approach the task with greater care and attention, more aware of what could go wrong and more deliberate in how they proceed with the task. In the context of AI-assisted decision-making, this translates to neither surrendering cognitively to an AI decision without scrutiny, nor dismissing the system's output outright, but rather adjusting the epistemic stance toward it; informed by the risk it carries.

As AI assisted decision-making often entails repeated interactions with both correct and erroneous system behaviour, a *perception calibration strategy* should prove effective across correct and erroneous system behaviour. A strategy that inflates perceived trustworthiness when the system performs well is as consequential as one that erodes it irreparably after a single failure. This raises two critical desiderata for an AI risk communication strategy. Foregrounding AI risk pre-emptively should not penalise a system, unfairly eroding trustworthiness even under correct behaviour. And when the system does err, risk communication should not act as a confirmation of seeded doubt to a degree that makes trustworthiness unrecoverable.

On the path to effective risk communication in human-AI interaction, the first chapter of this thesis establishes an initial empirical investigation toward understanding user-facing, in-situ risk communication over repeated interactions, and how it influences a user's system trustworthiness trajectory.

In this first empirical online user study ($N = 208$), we examine how risk information, disclosed prior to interacting with the system, influences system trustworthiness and its underlying cognitive processes in comparison to established perception calibration strategies. To investigate this, we manipulated two factors: the user-facing *perception calibration strategy* (Risk Communication, AI Accuracy, AI Apology, Baseline) and the *decision-stakes* (Low, High). The latter encoded the situational risk in which the decision is embedded, allowing us to assess its moderating role as a potentially relevant contextual factor. We implemented three Blocks of varying AI performance (Block I: correct AI judgements, Block II: incorrect AI judgements, Block III: correct AI judgements). To gain a fine-grained understanding of how risk communication operates on system

trustworthiness, we measured perceived AI transparency, performance, and utility as relevant and distinct constructs in AI-assisted decision-making [157], which enables us identify which perceptual dimensions are affected and in what direction.

The findings of this chapter reveal a nuanced picture of how communicative strategies shape trustworthiness perceptions across varying AI behaviour and decision-stakes. First, risk communication as an a priori strategy quantitatively supports perceived AI transparency after erroneous AI behaviour across decision-stakes, and qualitatively shows to initiate a recalibration of users' mental models of the AI system, in that participants anticipated AI failures and applied greater scrutiny and deliberation as a result, without cost to perceived AI performance or utility. Second, our findings surface important drawbacks of established strategies. AI Accuracy disclosure emerges as a double-edged sword, in that it inflates perceived system trustworthiness across all dimensions under high decision-stakes after erroneous AI behaviour, which is the exact context that should warrant caution rather than instilling trust. AI apology, as a reactive post-error intervention, is confined in its effect to a single dimension of system trustworthiness, perceived utility, and fails to support an informed assessment of AI capabilities and limitations. Third, our findings demonstrate the moderating role of decision-stakes as an operationalisation of situational risk, which emerges as an important contextual factor in how communicative strategies shape trustworthiness perceptions in AI-assisted decision-making.

The subsequent article provides the full motivation, methodology, results, and discusses risk communication as an alternative strategy to calibrate trustworthiness perception, its potential role as a cognitive forcing function, and cue to support system transparency.

3.2 Article I

This article is currently under review. Copyright is held by the authors. This is the authors' version of the work. It is posted here for your personal use. Not for redistribution.

Sarah Schömbs, Saumya Pareek, Jorge Goncalves, and Wafa Johal. 2026. How Communicating an AI's Relational Risk Influences Trust Perceptions in AI-Assisted Decision-Making. *Submitted for publication to The International Journal of Human-Computer Studies (IJHCS), Feb 2026.*

Ethics Application ID: 27586, the University of Melbourne Human Research Ethics Committee.

How Communicating an AI's Relational Risk Influences Trust Perceptions in AI-Assisted Decision-Making

Sarah Schömbs[a,*], Saumya Pareek[a], Jorge Goncalves[a], Wafa Johal[a]

[a]*School of Computing and Information Systems, The University of Melbourne Grattan Street, Parkville, Melbourne, 3010, VIC, Australia*

Abstract

AI systems err, however, existing *Strategies* to influence a user's trust perception can sometimes fail to foster perceptions and assessments that reflect system limitations, with potentially detrimental real-world consequences. In this paper, we introduce a novel Relational Risk (RR) strategy derived from risk communication research to pre-emptively communicate the likelihood of AI error and its potential consequences, foregrounding the costs of AI error, and compare it with existing *Strategies* (AI accuracy, AI apology). Participants ($N = 208$) engaged in an AI-assisted decision-making task with varying *Stakes* and AI performance, and we measured how these influence perceived AI trustworthiness across relevant perceptual dimensions (transparency, performance, utility). We find that RR enhances perceived AI transparency post-AI errors and encourages healthy scrutiny without diminishing or inflating perceived AI performance or utility. Conversely, AI accuracy boosts perceived AI transparency, performance, and utility under high decision-stakes, increasing system trustworthiness when unwarranted, while apology improves perceived AI utility but does not support system understanding. We contribute an in-depth investigation of how communication strategies shape AI trustworthiness perceptions, reveal important drawbacks of existing approaches, and demonstrate how integrating risk communication theory can enable more appropriate trustworthiness assessments.

Keywords: risk communication, AI trustworthiness, trust, transparency,

*Corresponding author

Email address: s.schombs@unimelb.edu.au (Sarah Schömbs)

AI-assisted decision-making

1. Introduction

The recently published AI Incidents Database (McGregor, 2021; Slattery et al., 2025) has recorded thousands of real-world AI failures, most of which were encountered and reported by end-users during interactions with deployed systems (De Miguel Velázquez et al., 2024). A discrepancy in perceived and actual AI trustworthiness in AI systems can lead to suboptimal outcomes with potentially harmful consequences, depending on the AI's deployment context. Research across human–automation interaction, machine learning, and HCI is therefore investigating how to best support users in their assessment of system trustworthiness and their understanding of system characteristics, including capabilities and limitations (Schlicker et al., 2025a; Wang and Yin, 2021; Zhang et al., 2020; Cao et al., 2024) as a necessary premise for deploying AI decision support that is accessible to users across domains. Imperatively, this necessitates fostering scepticism and critical thinking among users, to recalibrate perceptions of AI trustworthiness to match different levels of system performance (Shen et al., 2024) and thus to decrease the gap between perceived and actual AI trustworthiness (Schlicker et al., 2025a,b). It also requires nudging users to assess AI outputs more deliberately rather than applying broad heuristics (Buçinca et al., 2021; Park et al., 2019), especially in contexts where consequences are severe.

As AI-assisted decision-making becomes more widespread and accessible to users without expertise in AI, through applications such as AI-based dermatology assist tools (Liu et al., 2020), the central challenge is to identify *how* to foster critical thinking and *how* to develop intervention cues (Schlicker et al., 2025a,b) that communicate capabilities and limitations effectively in-situ, before negative consequences can manifest as a result of misplaced trust in AI systems.

Communicating an AI's accuracy has emerged as a promising strategy to prime user expectations (Kocielnik et al., 2019) and support trust calibration when engaging in AI-assisted decision-making (Cao et al., 2024; Yin et al., 2019; Rechkemmer and Yin, 2022; Papenmeier et al., 2019). However, communicating AI accuracy often does not correctly translate to users' exhibiting appropriate perceptions of AI trustworthiness, nor does it elicit a granular distinction between accuracy values (Yin et al., 2019; Lai and Tan, 2019),

which can have severe consequences in high-stakes applications. Rechkemmer and Yin (2022) show that users tend to interpret accuracy as a holistic quality of the AI rather than an uncertainty estimate, which offers one explanation for its limited effectiveness. In everyday life, *risk communication* is used to elicit reflective and contemplative thoughts, and to support people's understanding of probabilities and consequences associated with risks in decision-making (Schömbs et al., 2025; Visschers et al., 2009, 2012; Brown and Smith, 2007). Arguably, presenting AI accuracy conveys the risk of relying on the AI implicitly, as it communicates the uncertainty underlying an AI's decision and therefore hinges on its potential for fallibility. However, stated AI accuracy does a) not address the consequence component crucial to effective risk communication (Rowe, 1975; Dunlop et al., 2008) and b) deploys a positive framing of the AI's performance, a strategy leveraged in risk communication research to reduce risk awareness and promote risk-seeking decisions (McNeil et al., 1982; Peters et al., 2011). As a response to the mixed results of AI accuracy, recent work has began to examine alternative and complementary approaches, for example, explicitly disclosing AI weaknesses (Rieger et al., 2025), communicating system "incapabilities" in human-robot interaction (HRI) (Kwon et al., 2018; Schömbs et al., 2024), and explaining AI accuracy through everyday analogies (He et al., 2023), an approach inspired by risk communication research, with promising but unsatisfying results with respect to trust calibration. Importantly, none of these approaches make the relational risks of AI systems, i.e., the risks of using the technology itself (Stuck et al., 2022), explicit to users.

At the same time, growing work in HCI has begun to examine AI risks as a key topic in human–AI interaction, however, research has approached it predominantly from the perspective of AI practitioners. Examples include frameworks for documenting and reporting AI risks; e.g. RiskCards for language models (Derczynski et al., 2023), or model-specific risk reports (Rao et al., 2025), and tools designed to support AI practitioners to ideate and anticipate risks during the development stage (Herdel et al., 2024; Wang et al., 2024; Buçinca et al., 2023). While these perspectives are essential for advancing human-AI interaction and stress the importance of considering risk in AI-assisted decision-making, they re-emphasise that research investigating *the communication* of *relational risks to users* remains largely unaddressed. Given the empirically evident and foreseeable risks associated with the deployment and use of AI-assisted decision-making (McGregor, 2021; Slattery et al., 2025) and the limited effectiveness of current approaches to effectively

shape trust perceptions in a way that aligns with the AI's capabilities and limitations and support users in their trustworthiness assessment, it is crucial to examine explicit risk communication as an alternative strategy for navigating trust perceptions in human–AI interaction and, in doing so, to investigate how risk communication theory can be integrated into HCI to leverage natural synergies between the disciplines.

To address this research gap, we examine communicating Relational Risk (`RR`) in AI-assisted decision-making, which presents the likelihood of the AI erring and its consequences to foreground the cost of error. In detail, our `RR` design was informed by risk research conceptualising risk as a combination of the likelihood of a negative event and its potential consequence (Rowe, 1975; Dunlop et al., 2008). Whereas accuracy communication primarily captures the likelihood dimension, `RR` extends this by explicitly addressing both likelihood and consequence. We further ground our design in findings from risk communication research, which demonstrate that the way uncertainty is framed, particularly the valence of the frame, can systematically influence decision-making (Spiegelhalter, 2017; Gong et al., 2013). Accordingly, we adapt a negative framing for `RR` to potentially increase risk awareness, critical attentiveness and to foster more risk-averse perceptions (McNeil et al., 1982; Peters et al., 2011).

Our work is guided by the following overarching research questions:

RQ1: How does communicating the Relational Risk (`RR`) of an AI influence users' perceived AI trustworthiness across varying AI performance when compared to established *Strategies*, i.e. AI accuracy and apology?

RQ2: How does communicating the Relational Risk (`RR`) of an AI influence users' expectations and understanding of the AI?

RQ3: How does the *Stakes* (low, high) inherent to the decision-making context affect users' perceptions of AI trustworthiness across varying AI performance (**RQ3.1**), and how does it interact with the investigated *Strategies* in influencing these perceptions (**RQ3.2**)?

We investigate the communication of our proposed Relational Risk (`RR`) approach in comparison to two established approaches in AI-assisted decision-making: stated AI accuracy (`Acc`), an approach used as an a priori indicator of AI performance to influence trust perceptions and set appropriate expectation (Cao et al., 2024; Yin et al., 2019; Rechkemmer and Yin, 2022; Papenmeier et al., 2019), and AI apology (`Apol`), a strategy to navigate trust

dynamics particularly post erroneous AI behaviour (Kim and Song, 2021; Tolmeijer et al., 2020; Pareek et al., 2024). We assess how these *Strategies*, in interaction with *Stakes*, affect the trajectory of perceived AI trustworthiness across varying AI performance and system understanding. We thereby focus on an in-depth investigation into trust perceptions and its underlying constructs (perceived AI transparency, performance, and utility) as a critical foundational step (Schlicker et al., 2025a; Mayer et al., 1995; Roesler* et al., 2022).

We present a 2 (*Stakes:* low vs. high) × 4 (*Communication Strategy:* `RR`, `Acc`, `Apol`, Baseline) between-subjects user study with $N = 208$ participants. The study was contextualised as an AI-assisted diagnostic task in which participants classified microscope images as infested or not (Fahnenstich et al., 2024). *Stakes* was operationalised through the decision-making context, with participants inspecting microscope images of either plant (low) or human tissue (high) for possible bacterial infestation. Participants interacted with the AI across 3 blocks with varying AI performance, including distinct blocks with correct and erroneous AI identifications, to gain an in-depth understanding of participants' trust perceptions. We measured three core facets of AI trustworthiness after each block: perceived AI transparency, performance, and utility.

Our results show that communicating `RR` enhances perceptions of AI transparency after erroneous AI behaviour across low (plant tissue) and high (human tissue) stakes relative to the Baseline, while prompting users to anticipate failures, temper their expectations, and scrutinise AI outputs more carefully. We show that this deliberate assessment positions `RR` communication as a cognitive forcing function that sustains caution without diminishing perceived AI performance or utility, while also raising questions about potential limitations in low-stakes contexts that warrant future investigation. Second, we find that `Acc` communication, although it makes users aware of the AI's potential to make mistakes, also backfired: stated AI accuracy made users believe that the AI was highly accurate, elevating AI performance perceptions compared to the Baseline and `RR` in high-stakes contexts where misplaced perceptions of AI trustworthiness could be particularly problematic. Third, we find that `Apol` increases the perceived usefulness of the AI compared to the Baseline but fails to support system understanding, which demonstrates an important constraint of apology-based interventions.

Our work makes the following contributions:

- We propose RR as an important alternative approach to influence trust perceptions in AI-assisted decision-making, informed by risk communication theory.
- We provide empirical evidence on the effects of RR compared to established *Strategies* across perceptions of AI transparency, performance, and utility, revealing the respective strengths and drawbacks of different *Strategies*.
- We demonstrate empirically that *Stakes* moderates the effectiveness and appropriateness of *Strategies* on perceived AI trustworthiness. Our findings imply that *Stakes* must be considered when designing and evaluating *Strategies* to shape trust perceptions in AI-assisted decision-making. We provide empirical support for the recently developed trustworthiness assessment model (TrAM) (Schlicker et al., 2025a,b), which proposes that a risk–benefit evaluation occurs at the level of trustworthiness perceptions, rather than being confined to the behavioural stage as earlier trust models assumed. (Mayer et al., 1995; Stuck et al., 2022).
- We offer implications for the design of *Strategies* to shape trust perceptions in AI-assisted decision-making that remain sensitive to the *Stakes* inherent in different decision-making contexts, which contributes to a critical topic in human–AI interaction.

Our work contributes to integrating risk communication theory in HCI, and shifts the conversation towards explicit risk communication as an alternative approach for navigating trust dynamics in AI-assisted decision-making. We position Relational Risk (RR) as an important strategy in human-AI interaction that warrants further investigation and argue that trust perceptions can be appropriately influenced via prior risk communication, without deterring AI use.

2. Related Work

We review prior work on approaches to influence trust perceptions in AI-assisted decision-making alongside risk communication theory to ground the manipulation examined in this study. This motivates the design of Relational Risk (RR) communication as an approach that builds on and extends existing approaches by integrating insights from risk communication theory.

By critically comparing established approaches to calibrate trust perceptions with principles from risk communication theory, we identify their limitations and clarify the rationale behind our study design.

2.1. Trust in Human-AI Interaction

In this work, we focus on users' perceptions of AI trustworthiness as a form of situational trust, a multi-dimensional construct in human–AI interaction (Hoff and Bashir, 2015; Mayer et al., 1995; Lee et al., 2022; Kohn et al., 2021; Roesler* et al., 2022). Early works by Lee and See (2004) describes three determinants of perceived AI trustworthiness: performance, which reflects a user's assumptions about how well an AI performs, process, which denotes the extent to which a user comprehends how the AI operates, and purpose, which captures whether the AI is aligned with the user's objectives (perceived benefit). Recent work by Roesler* et al. (2022) extends this framework in the context of AI-assisted decision-making. The authors conceptualise and validate perceived AI trustworthiness as a multi-dimensional construct defined by performance and utility, which load on the performance pillar of Lee and See (2004), purpose, which corresponds to the purpose pillar, and transparency, which maps onto the process pillar and constitutes an important extension in AI-assisted decision-making, reflected by recent research echoing the need for more transparency in human-AI interaction (Bhatt et al., 2021; Prabhudesai et al., 2023).

Unless stated otherwise, throughout this paper we use the term trust to denote perceptions of AI trustworthiness.

2.1.1. Strategies to influence Trust Perceptions in AI-assisted Decision-Making

The alignment of a user's perceived AI trustworthiness with system capabilities and limitations, i.e. calibrated trust (Lee and See, 2004; Hoff and Bashir, 2015), is a key challenge in HCI given the increasingly ubiquitous deployment of AI systems in everyday settings and the potentially detrimental consequences depending on the deployment context (Wang and Yin, 2022). Calibrated trust has therefore become a crucial "desideratum" of explanations in human-AI interaction (Wang and Yin, 2022) to foster a user's "appropriate use" of deployed AI systems (Lee and See, 2004). Calibrated trust, in detail, is defined as a user's trust that matches the system capabilities, albeit avoiding both overtrust, which describes trust that "exceeds system capabilities", and distrust, which describes trust that "falls short of system capabilities" (Lee and See, 2004).

The recently developed trustworthiness assessment model (TrAM) (Schlicker et al., 2025a) emphasises the theoretical gap and importance of investigating the discrepancy between *actual* and *perceived* AI trustworthiness, as a system's trustworthiness must be inferred through observable cues. To support people's mental models of AI systems and to calibrate appropriate AI trustworthiness in human–AI interaction, different approaches have been investigated by the HCI community. These include global cues introduced *prior to* human-AI interaction (e.g., AI accuracy (Cao et al., 2024)) and local explanations communicated *during* interaction, often in conjunction with AI decision outcomes (e.g., AI confidence (Rechkemmer and Yin, 2022), counterfactuals (Wang and Yin, 2022), locus of causality (Pareek et al., 2025; Kim and Song, 2023, 2021)). They further comprise strategies designed to navigate (repair) trust *after* the AI has erred, with apology as an established approach in both AI-assisted decision-making (Pareek et al., 2024; Kim and Song, 2021; Tolmeijer et al., 2020) and human-robot interaction (Kox et al., 2021; Nesset et al., 2023), where its conceptual purpose is to prevent disuse or AI aversion following an error.

2.1.2. AI Accuracy and Priming Expectations of AI Performance

Among these approaches to calibrate trust perceptions of AI systems, AI accuracy has received particular attention as a simple a priori approach to communicate uncertainty, and to guide expectations and trust dynamics (Bhatt et al., 2021; Cao et al., 2024). Research shows that AI accuracy influences perceived trustworthiness more strongly than either AI confidence (Rechkemmer and Yin, 2022) or explanation fidelity, i.e. the degree to which explanations truthfully represent how the model works (Papenmeier et al., 2019). Interestingly, lay users seem to perceive accuracy as a factual property of the system, while interpreting confidence as an uncertain, volatile estimate (Rechkemmer and Yin, 2022).

At the same time, findings indicate that accuracy information needs to be treated with caution, as users' perceived trustworthiness of an AI does not sensitively reflect different levels of stated accuracy. Lai and Tan (2019) show that presented model accuracy information increases user trust even when the communicated accuracy is solely 50%, with relatively small differences in trust compared to an AI accuracy of 87%. A surprising result, given that users should place considerably less trust in a system that communicates such low accuracy. While this may be unproblematic in contexts with little consequence, such misplaced perceptions could be harmful in deployment

settings with high stakes (e.g. medical diagnostic). Similarly, Yin et al. (2019) report that participants' trust did not significantly differ between conditions with no accuracy information, 60%, or 70% stated AI accuracy. Significant differences only emerged when contrasting this lower-accuracy batch (none/60%/70%) with the high-accuracy batch (90%/95%), with the latter eliciting higher perceived AI trustworthiness.

To counter this shortcoming by redirecting users' attention to the system's limitations, Rieger et al. (2025) shifted attention to explicitly exposing AI weaknesses as a global, a priori explanation. In two experiments, AI weaknesses were described as limitations in handling certain types of cases due to reduced training data, which has also shown to influence how users attribute the locus of causality of AI failures (Pareek et al., 2025). AI weakness disclosures increased perceived transparency in the first experiment, but not in the second. The second experiment instead revealed a temporary decline in initial trust, followed by recovery, which indicates no sustained erosion of trust perceptions through AI weakness disclosure. This result draws attention to an important consideration: AI interventions should serve to calibrate perceived AI trustworthiness, while not undermining the system's perceived utility, inducing algorithmic aversion (Passi and Vorvoreanu, 2022), or discouraging users from benefiting from its support when appropriate. Similar to how apologies conceptually aim to prevent disuse post failure (Pareek et al., 2024).

In this paper, our objective is to examine how `RR` compared to established *Strategies* (`Acc`, `Apol`) affects perceptions of AI transparency, performance, and utility (RQ1). While `RR` may serve as an a priori approach to help users anticipate AI fallibility and thereby potentially support transparency, it is important to explore whether this comes at the expense of perceived utility and performance across blocks of correct and incorrect AI behaviour.

2.2. Risk Communication and Risks in Human-AI Interaction

Recent work specifies three types of risk that warrant consideration when assessing trust dynamics in human-AI interaction: relational risk, situational risk, and risk propensity (Stuck et al., 2022). Our work focuses on the first two constructs and does not address risk propensity in detail, which refers to an individual's general inclination toward risk aversion or risk-taking as an individual tendency.

2.2.1. Relational Risk

Users navigate a relational form of risk associated with the AI system itself (Stuck et al., 2022). Perceived *relational risk* describes the risk related to the use of a technology itself, in our case an AI agent. It refers to the user's perception that using the AI agent may lead to positive or negative outcomes (Stuck et al., 2022). Importantly, risk itself is typically defined as a combination of the likelihood of a hazard and its potential consequence, which demonstrates that risk communication fundamentally involves conveying two determinants, i.e., probability *and* consequence (Rowe, 1975; Dunlop et al., 2008; Spiegelhalter, 2017).

Human-AI interaction research has begun to incorporate knowledge derived from risk research to improve people's understanding of AI uncertainty. Presenting uncertainty in a frequency format has shown to support more appropriate reliance and reduce confirmation bias (Cao et al., 2024; Zhang et al., 2020), which aligns with risk communication research showing that frequency formats increase people's risk perceptions (Peters et al., 2011; Spiegelhalter, 2017; Peters et al., 2011). Motivated by the risk communication literature, He et al. (2023) investigated the use of analogies from everyday contexts (e.g. weather forecast) as an additional cue to explain the uncertainty inherent to stated accuracy and to support numeracy. While the analogies were sometimes perceived as helpful for understanding AI accuracy, they had no significant effect on perceived AI trustworthiness and were also criticised as unfamiliar or unspecific for the task context. Notably, both approaches leave the consequence component crucial to risk communication theory unaddressed.

To guide people's risk perception and to elicit appropriate decision-making, traditional risk theory emphasises the importance of frame valence in risk communication, that is, whether risks are presented in a positive or negative framing (Spiegelhalter, 2017). For example, describing side-effect probabilities as the proportion of patients who remain unaffected reduces the perceived risk of a medication compared to emphasising the proportion who experience side effects (Peters et al., 2011). In treatment decisions, positive framing tends to increase acceptance of high-risk interventions such as surgery, while negative framing steers people's preferences toward non-invasive alternatives, reflecting more cautious, risk-averse decisions (McNeil et al., 1982; Armstrong et al., 2002; Gong et al., 2013). Exposing AI weaknesses (Rieger et al., 2025; Kwon et al., 2018; Schömbs et al., 2024) can be viewed as a loose form of

negatively framed system capabilities, and our work extends this idea by directly comparing error-rate (negative framing) and accuracy-rate (positive framing) presentations.

In this work, we examine trust calibration in human–AI interaction through the lens of risk communication theory. Informed by prior work, we design RR to explicitly communicate both the likelihood of AI error, conveyed as frequency framing, and its potential consequences. RR therefore extends established uncertainty communication by incorporating the consequence dimension and by adapting a negative framing to foreground the costs of AI error, with the aim of fostering more cautious perceptions of trustworthiness. Beyond examining how RR affects perceptions of AI transparency, performance, and utility (RQ1), we also investigate how it shapes users' expectations and understanding of the AI system (RQ2).

2.2.2. Situational Risk — Decision Stakes

Situational risk is another critical type of risk in AI-assisted decision-making. It defines risk as a contextual factor, regardless of the use of a specific technology (Stuck et al., 2022). Situational risk is considered crucial for human–AI trust, as misplaced trust in AI can have real-world consequences with varying degrees of severity depending on the application context. Indeed, as Kohn et al. (2021) argue, "trust in a situation that has no potential for negative consequences is effectively meaningless". Nevertheless, empirical investigations that assess situational risk in human-AI research remain scarce (Hoesterey and Onnasch, 2023). Contextualised in AI assisted decision-making, prior research predominantly focused on either side of the coin - low situational risk (e.g. AI supporting the identification of animals or shapes (Pareek et al., 2024) or answering factual tasks (Kim et al., 2024)) or high situational risks (e.g. skin cancer detection (Cao et al., 2024), loan prediction (He et al., 2023) or the assessment of jail time in criminal law cases (Kahr et al., 2024)).

Some research compared the effect of situational risk on trust during AI assisted decision-making, with inconsistent findings showing no effect of situational risk on AI trustworthiness (Hoesterey and Onnasch, 2023; Fahnenstich et al., 2024; Yin et al., 2019), high situational risk significantly decreasing AI trustworthiness (Pareek et al., 2025; Ashoori and Weisz, 2019), and participants exhibiting greater convergence with AI support than with human support under high situational risk (Fahnenstich et al., 2024). Findings by Hoesterey and Onnasch (2023) show that participants' decision time

significantly increased under high situational risk. In line, verification behaviour too was elevated in high-risk contexts, more so when users were supported by decision automation (a system-generated recommendation) as opposed to information automation (system-highlighted cues). However, this heightened scrutiny gradually declined as interaction with the decision support system progressed, mirrored by a decline in risk perception, which the authors attributed to a "habituation" to the situational risk.

On the quest to understand the effectiveness and appropriateness of *Strategies* relative to their application context (e.g. health), we first singularly study the effects of *Stakes* on AI trustworthiness across its facets (RQ3.1). By *Stakes*, we refer to situational risk, yet use the term *Stakes* to more clearly distinguish it within our study. We then investigate how the *Stakes* inherent to the decision context interacts with *Strategy* in influencing participants' perceptions of AI trustworthiness under conditions of correct and incorrect AI performance (RQ3.2). We carefully operationalise the *Stakes* of the AI application through the context of the diagnostic task (human tissue vs. plant tissue samples) to provide a highly comparable way to study its effects.

3. Method

We conducted a mixed-method user study with a 2 (*Stakes:* low, high) $\times$ 4 (*Strategy*: `Acc`, `RR`, `Apol`, Baseline) between-subjects factorial design, to examine the impact of *Strategy* on perceptions of AI trustworthiness (RQ1) and system understanding (RQ2), and to investigate the role of *Stakes* in shaping trust perceptions during AI-assisted decision-making (RQ3).

3.1. Task

We required an AI-assisted decision-making task that allowed us to manipulate *Stakes* while remaining comparable across conditions (low vs. high). To do so, we contextualised our experiment as a diagnostic task that required participants to screen and classify different microscope samples. During the task, the microscope sample images showed varying bacterial density, see Figure 1. We asked participants to perform a visual estimation for each microscope image, i.e., to estimate the percentage of bacteria-covered area and classify the sample as either infested or not, given a bacterial infestation threshold of 35% coverage. Samples showing more than 35% infestation were to be classified as 'infested'. A visual estimation task with a predefined cut-off

boundary is a task type frequently used to assess trust dynamics in human-AI interaction, owing to their real-world applicability (Rieger et al., 2024; Fahnenstich et al., 2024). To accomplish the task, participants received decision support from an AI agent that presented its judgement about whether a sample was infested or not. We included a training block to support task familiarisation and comprehension.

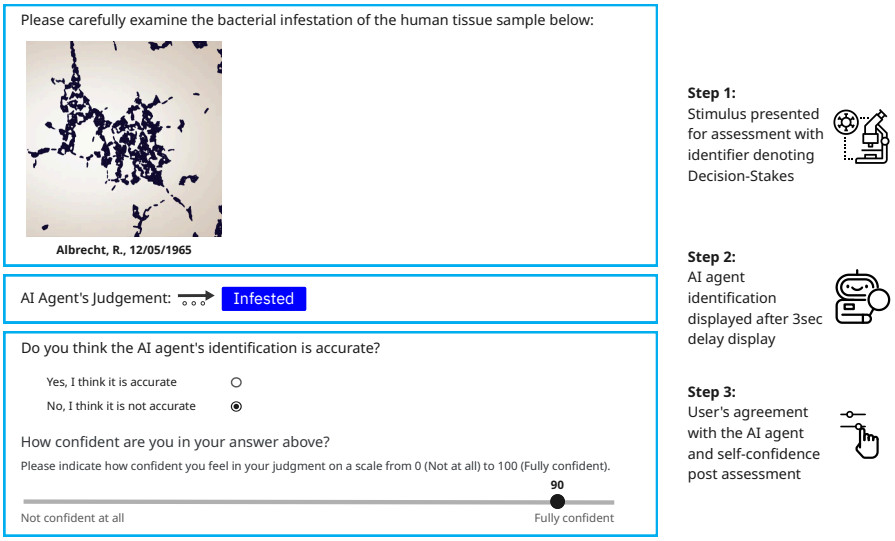


Figure 1: Example diagnostic trial from the high-stakes condition (human tissue sample) showing an incorrect AI classification of 'Infested' during Block 2, despite bacterial coverage being $< 35\%$.

3.2. Experimental Conditions

3.2.1. Operationalising Stakes

We derived a set of microscope images from the OSF[1] repository by Fahnenstich et al. (2024), which are microscope images of bacteria originating from the University Medical Centre Rotterdam. While Fahnenstich et al. (2024) showed no influence of stakes on a unidimensional trust measure, they argue that it does not undermine the suitability of the task stimuli as a manipulation for stakes, given their successful validation of the stimuli

[1] `https://osf.io/uw239/`

themselves. We contacted the authors to draw on their stimuli and scenario framing for our manipulation.

We selected a subset of nine microscope images that served as the task stimuli. *Stakes* was operationalised by varying the decision-making context, with participants either classifying microscope images that represented 'human tissue' or 'plant tissue' samples. Each stimulus was accompanied with a corresponding identifier (plant type / patient name; date), exemplified in Figure 1. To contextualise *Stakes* further, participants received a short 'brief' at the beginning of the task, accompanied by a respective institution's logo to simulate a realistic decision-making context and thereby increase ecological validity. The brief framed the source of the microscope images and further established the *Stakes* associated with the classification decisions. In the low *Stakes* condition, participants were told to assess samples from a plant nursery called "Verdara Plant Centre", with samples taken from diverse plants, some of which exhibited signs of damage (e.g., root decay). In the high *Stakes* condition, the same task was framed within a medical context. The accompanying brief explained that the batch consisted of samples from different patients at the "Novaira Medical Centre", some of whom were experiencing serious symptoms (e.g., respiratory issues).

3.2.2. Operationalising Strategies to calibrate Trust Perceptions

We carefully refined and implemented three *Strategy* conditions (see Section 2 for rationale), each of which provided participants with different "observable cues" during the AI-assisted decision-making task to provide information "about a system's actual trustworthiness" (Schlicker et al., 2025a).

`Acc`: This condition communicates the likelihood of correctness as AI accuracy in a percentage format as a well established *Strategy* following prior works (e.g., (Rechkemmer and Yin, 2022; Yin et al., 2019)), accompanied by a frequency based representation to support numeracy (Cao et al., 2024; Zhang et al., 2020), while adapting a positive frame valence (communicating the accuracy rate; number of correct identifications). *"The AI agent has an accuracy of 78%. That means it provides correct identifications in 78 out of every 100 cases."*

`RR`: This condition communicates the Relational Risk of using the AI agent itself by a) communicating the likelihood of failure to align with risk theory on negative framing (Peters et al., 2011; McNeil et al., 1982;

Armstrong et al., 2002; Spiegelhalter, 2017), accompanied with a frequency based format to support numeracy (Cao et al., 2024; Zhang et al., 2020) (communicating the error rate; number of incorrect identifications), and by b) explicitly including the consequence dimension of risk following risk theory (Rowe, 1975; Dunlop et al., 2008; Spiegelhalter, 2017). *"The AI agent has an error rate of 22%. That means it provides incorrect identifications in 22 out of every 100 cases. The risk of using the AI agent could have consequences, such as not detecting bacterial infestations or falsely detecting bacterial infestations."*

`Apol`: This condition communicates an apology from the AI as an established *Strategy* in AI-assisted decision-making (Kim and Song, 2021; Kim et al., 2024; Pareek et al., 2024) employed to navigate trust perceptions post AI failures (recovery) and over repeated interactions. A desideratum of an effective *Strategy* should be to sustain perceived usefulness while still inducing appropriate caution, which motivates the inclusion of this approach. We draw on the framing from Pareek et al. (2024). *"I am sorry for any errors I may have made. I apologise for any inconvenience caused by these errors."*

We further included a Baseline condition in which participants completed the study without any communicative strategy.

3.3. Repeated Interaction with Varying AI Performance

To assess trust dynamics across varying AI performance, we divided the main experiment into three blocks, with each block involving three microscope assessment tasks. In Block 1, the AI agent provided a correct identification across all three cases. In Block 2, we introduced two erroneous identifications in a controlled sequence, with a false negative in the first case and a false positive in the third, based on prior findings showing order effects of AI failures on trust (Kahr et al., 2024). In Block 3, the AI agent again produced correct decisions across all three cases. We adopted this tripartite structure as an established protocol in human–AI interaction to examine trust formation and its dynamics over repeated interactions across varying AI performance (Kahr et al., 2024; Pareek et al., 2024; Kim and Song, 2023; Tolmeijer et al., 2021). For all three blocks, we selected stimuli that significantly deviated from the bacterial infestation cut-off for participants to be able to determine the AI agent's performance. We simulated the AI and

fixed the accuracy at 78% to ensure control over the frequency, position, and type of error across Block 2, consistent with prior studies in AI-assisted decision-making (Buçinca et al., 2021; He et al., 2023).

3.4. Measures

3.4.1. Trust in Automation - Propensity to Trust - TiA-PtT

Trust in human–automation interaction can be categorised as dispositional and situational (Hoff and Bashir, 2015). Dispositional trust describes a user's propensity to trust automation determined by factors such as age, gender, or personality (Mayer et al., 1995; Hoff and Bashir, 2015). We measure participants' dispositional trust in automation as a covariate to account for individual differences, using the TiA-PtT questionnaire (Trust in Automation – Propensity to Trust subscale) (Körber, 2019). This measure has been shown to reliably capture dispositional trust and to be an important determinant of perceived AI trustworthiness in AI-assisted decision-making (Pareek et al., 2025).

3.4.2. Perceived AI Trustworthiness - Transparency, Performance, Utility

A user's situational trust in AI can be dynamic and shift depending on the specific interaction and type of AI system (Kohn et al., 2021) as it is influenced by variable contextual characteristics such as the type of AI system, its complexity, but also the perceived risks associated with system use (Hoff and Bashir, 2015). We investigate participants' situational trust using the validated Multi-dimensional Trust Questionnaire (MTQ) (Roesler* et al., 2022), and measure the three dimensions — perceived AI transparency, performance, and utility. Each dimension consists of a three-item subscale captured on a 4-point scale from *"Disagree"* to *"Agree"*. For example, *"The way the system works is clear to me"* for perceived transparency, *"The system works well"* for perceived performance, and *"I find that the system supports my work"* for perceived utility.

3.4.3. Agreement and User Confidence

We adopted the agreement and confidence measure from prior research (Kim et al., 2024; Pareek et al., 2024) and captured participants' agreement with the AI's judgement asking "*Do you think the AI's judgement is accurate?*" with a binary response option. We further asked how confident participants are in their (dis)agreement with the AI. Participants' self-confidence was measured on a slider ranging from 1 (not confident at all) to 100 (fully

confident) captured in 10-point increments. To avoid potential biases through anchoring effects (Sellers, 2013), the slider handle was only displayed after participants actively clicked on the slider. We measured both agreement and self-confidence after each sample assessment case, primarily to serve as a robustness check for our manipulation of varying AI performance across Block 1–3.

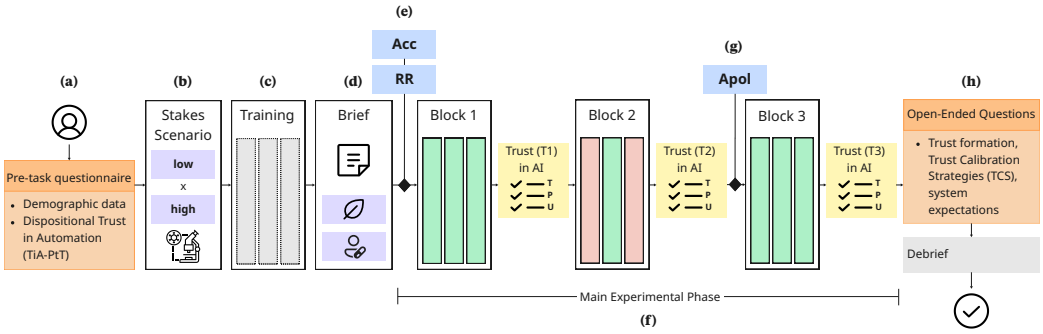


Figure 2: The experimental flow. (a): Pre-task questionnaire. (b): Operationalisation of *Stakes* (plant tissue vs. human tissue). (c): Training Phase. (d): Brief (medical centre vs. plant nursery). (e): A priori *Strategy*: AI Accuracy (`Acc`); Relational Risk (`RR`). (f): Block 1-3 with repeated measures of perceived AI Trustworthiness (transparency, performance, utility). (g): Apology (`Apol`) as reactive post error *Strategy*. (h): Post-task open-ended question.

3.5. Procedure

We conducted several pilot studies to finalise the experimental flow of our user study; see Figure 2 for an overview. Given the 2 × 4 between-subjects design, we randomly assigned participants to one of the *Stakes* and *Strategy* conditions.

Participants were first guided to read a plain language statement of the study. After giving consent, participants received a pre-task questionnaire to collect demographic information and participants' propensity to trust automation (Körber, 2019). Subsequently, participants received general instructions about the bacterial diagnostic task and how to assess the sample's infestation through visual estimation. The explanation was illustrated with a vignette, including an example case that displayed a microscope image alongside a corresponding image of either a human patient or a plant to set the scenario for the *Stakes* (human tissue vs. plant tissue sample). For task

familiarisation, participants then completed a training block with three practice trials, distinguished from the main experiment by the different staining of the microscope images (Fahnenstich et al., 2024), and could only proceed once all training trials were completed successfully. We further administered an attention check and a comprehension check, both of which we used as exclusion criteria to ensure data quality.

Prior to the main experimental phase, we informed participants that, for each assessment, they would be supported by an AI agent. They then received a 'brief', see Section 3.2.1 for more details, that situated the main experimental phase in a real-world context while emphasising the respective *Stakes*.

During the main experiment, participants completed three blocks with varying AI performance, as described in Section 3.3. In each block, they assessed three microscope samples, one by one: for each sample, participants viewed the image, and after a three-second delay with an animated loading display, received the AI's identification. We implemented the delay to increase the perceived authenticity of the AI model (Pareek et al., 2025). After each case, participants reported their agreement with the AI's classification and their self-confidence. Subsequent to each block, we captured participants' perception of AI trustworthiness (transparency, performance, utility) (Roesler* et al., 2022).

The *Strategies* were implemented at different stages during the main experiment, in line with their conceptual purpose. The `Acc` and `RR` messages were presented before Block 1, as global a priori strategies intended to prime participants before the interaction. The `Apol` message was presented after Block 2, consistent with its role as a reactive *Strategy* to navigate trust post AI failure and in accordance with prior work (Kim and Song, 2021; Kim et al., 2024; Pareek et al., 2024). Participants in the Baseline condition received no explicit communication about the AI system.

After the main experiment, participants responded to qualitative open-ended questions which asked them to reflect on their evolution of trust and what caused it, specifically after the AI made incorrect suggestions, and how the respective *Strategy* they received (`Acc`, `RR`, `Apol`) influenced their expectations and understanding of the AI.

We ended the study with a debrief stating that the scenario and samples were fictional and that there were no real risks associated with the experiment.

3.6. Participants

We conducted an a priori power analysis to determine our sample size using the R package `InteractionPoweR` (Baranger et al., 2023) to account for interaction effects. The calculation was based on an $\alpha = .05$ and a power of .9, resulting in a targeted sample size of at least $N = 204$. We deployed our study on the crowdsourcing platform Prolific[2]. We applied several pre-screening criteria to ensure participants were proficient in English, aged $\geq$ 18, had a prior approval rate of $\geq$ 98%, and to maintain a balanced gender distribution. Further, to mitigate learning effects, each participant could take part only once. Participants were compensated at £9.00 per hour; with an estimated completion time of 12 minutes, this converted to average payments of around £1.80, rated "good" under the platform's payment standards[3].

We excluded participants who failed either our attention check, comprehension check, or who did not finish the survey (overall 40 cases). After exclusion, the recruited sample resulted in $N = 208$ participants with an average age of 39.06 ($SD = 13.35$). The study was approved by our University's Human Ethics Committee.

4. Results

We employed a Cumulative Link Mixed Model (CLMM) to examine how different *Strategies* (`Acc`, `RR`, `Apol`) influence users' perceived AI trustworthiness across its constituent facets (transparency, performance, utility) and across varying AI performance (Block 1–3) (RQ1). We further analysed how these strategies affect users' expectations and understanding of the AI system (RQ2). Finally, we assessed how the *Stakes* inherent to the decision-making context affect perceptions of AI trustworthiness across varying AI performance (RQ3.1), and how it may interact with the investigated *Strategies* in shaping these perceptions (RQ3.2). We conducted post-hoc analyses of the interaction using Tukey-adjusted faceted contrasts of Estimated Marginal Means (EMM) to correct for multiple comparisons within each level of the interacting variables. We report these comparisons alongside the corresponding EMMs. Effect sizes are reported as odds ratio, a suitable interpretation of logistic regression, which represent the change in odds that an outcome (per-

[2] `https://www.prolific.com/`

[3] `https://researcher-help.prolific.com/en/article/9cd998`

ceived AI trustworthiness) will occur given a one-unit change in the predictor variable (*Strategy*, *Stakes*) (Field et al., 2014).

4.1. CLMM Construction

To analyse our RQs, we employed a CLMM given the ordinal nature of our data and to account for covariates, random effects, and the repeated-measures design; modelling perceptions of AI trustworthiness as the dependent variable, operationalised across dimensions with TrustDim $\in$ {transparency, performance, utility} and block $b \in \{1, 2, 3\}$ ($TrustDim_b$). The model included *Strategy* and *Stakes* as independent variables, as well as their interaction.

In our model, we incorporated propensity to trust automation ($TiA - PtT$) as a covariate to adjust for participant's predisposition (Hoff and Bashir, 2015). We further included $CentredAge$ as a covariate to control for any effects respective to age, given an age range of 18-77 in our data. To account for individual differences and the repeated measures design, we include participants' id (PID) as a random effect. The Cronbach's α for the respective dimensions transparency ($\alpha = 0.91$), performance ($\alpha = 0.76$), and perceived utility ($\alpha = 0.71$) indicates an acceptable - excellent internal consistency among items. Nevertheless, to account for potential item-specific variance in the responses, we further included $Item$ as a random effect. For Block 2 and Block 3 models, we included the trust score from the preceding block(s) $\sum_{j=1}^{b-1} TrustDim_{Block\,j}$ as an additional predictor, where $b \in \{2, 3\}$ denotes the current block and j indexes the preceding block(s), thereby modelling the temporal dependency of trust across blocks (Pareek et al., 2024). Given that `Apol` was introduced after Block 2 due to its conceptual purpose as a trust calibration strategy post failure, responses from `Apol` and Baseline were aggregated for Block 1 and 2, as there were no experimental differences for these experimental blocks. We used the statistical R package `ordinal` [4] to build our CLMM, which was formulated as follows:

$$TrustDim_b \sim Strategy \times Stakes + TiA{-}PtT + CentredAge + \sum_{j=1}^{b-1} TrustDim_{Block\,j} + (1 \mid PID) + (1 \mid Item)$$

[4] `https://cran.r-project.org/web/packages/ordinal/index.html`

We further assessed the Variance Inflation Factor (VIF) across all employed models to ensure no multicollinearity across the independent variables (VIF < 5) (Rawlings et al., 1998). A detailed overview of the constructed models is provided in the Appendix (Table A1), along with model summaries of the CLMM analysis for perceived AI transparency, performance, and utility (Tables A2–A4).

4.2. Quantitative Results

4.2.1. Robustness Check - Agreement and Confidence

We deliberately manipulated the AI's performance across the three blocks to examine how the communicated *Strategy* (`Acc`, `RR`, `Apol`) influences perceptions of AI trustworthiness under varying observable performances, which resulted in a simulated overall AI accuracy of 78%. Participants' agreement with the AI agent's identification was analysed after each assessment case as a primary robustness check. In Block 1, the AI provided correct identifications in all three cases; in Block 2, it produced two erroneous identifications; and in Block 3, it again performed correctly across all three trials. Participants' agreement ratings reflected this manipulation, with 96.15% consensus in Block 1, 38.14% in Block 2, and 97.44% in Block 3. We further examined participants' accuracy in agreement (i.e., agreeing when the AI was correct, disagreeing when it was incorrect), which resulted in 24 errors out of 624 responses in Block 1 (96.15%), 66 errors in Block 2 (89.42%), and 16 errors in Block 3 (97.44%). Consistent with prior research (Fahnenstich et al., 2024; Pareek et al., 2024), we assessed participants' confidence in their own decisions as a secondary robustness check of the manipulation. Mean confidence scores were 92.53% ($SD = 11.13$) in Block 1, 84.05% ($SD = 15.96$) in Block 2, and 92.76% ($SD = 11.98$) in Block 3. These ratings confirm that the task operationalisation (observable AI performance) was perceived as intended, with overall high confidence across blocks and a slight but expected drop in Block 2 due to the presence of observable AI errors.

4.2.2. Perceived Transparency

For **Block 1**, in which the AI consistently provided correct assessments, neither `Acc` ($\beta = 1.08$, $SE = 0.91$, $p = .233$) nor `RR` ($\beta = 0.15$, $SE = 0.88$, $p = .865$) differed significantly from the Baseline. Our results further show no significant interaction effects between *Stakes* and *Strategy* (all $p > .05$; see Figure 3(a)).

In **Block 2**, as the AI exhibited erroneous identifications, our CLMM showed a significant effect of *Strategy* on perceived transparency, with `Acc` ($\beta = 3.24$, $SE = 1.10$, $p = .003$, $OR = 12.88$) and `RR` ($\beta = 2.55$, $SE = 1.07$, $p = .047$, $OR = 18.33$) significantly increasing perceived transparency relative to Baseline. We further observed that prior block-level transparency perceptions significantly predicted transparency perceptions in **Block 2** ($\beta = 1.24$, $SE = 0.26$, $p < .001$, $OR = 3.51$). Post-hoc faceted contrasts revealed that the effect of *Strategy* was moderated by *Stakes*. Imperatively, `RR` consistently led to higher perceived transparency relative to Baseline, across both *low-* (`RR`: $EMM = 3.86$, $SE = 0.96$ vs. Baseline: $EMM = 0.53$, $SE = 0.62$; $\beta = 3.33$, $SE = 1.12$, $p = .009$, $OR = 27.94$) and *high Stakes* contexts (`RR`: $EMM = 2.18$, $SE = 0.89$ vs. Baseline: $EMM = -0.36$, $SE = 0.62$; $\beta = 2.55$, $SE = 1.07$, $p = .047$, $OR = 12.81$), see Figure 3(b). Conversely, `Acc` ($EMM = 2.88$, $SE = 0.92$) significantly increased transparency perceptions compared to Baseline ($EMM = -0.36$, $SE = 0.62$) specifically under *high Stakes*; $\beta = 3.24$, $SE = 1.10$, $p = .009$, $OR = 25.54$), with no significant effect under *low Stakes* ($\beta = 1.91$, $SE = 1.06$, $p = .168$).

In **Block 3**, in which the AI again provided correct assessments, the CLMM revealed a significant effect of *Strategy*; `Acc` led to significantly higher perceived transparency than Baseline ($\beta = 3.29$, $SE = 1.61$, $p = .041$, $OR = 26.88$). Importantly, *Strategy* moderated the effect of *Stakes* on perceived transparency. Specifically, under the Baseline condition, participants in the *high Stakes* context reported significantly lower transparency compared to those in the *low Stakes* context ($\beta = -3.07$, $SE = 1.53$, $p = .044$, $OR = 21.54$). By contrast, when AI accuracy was communicated (`Acc`), the pattern reversed, with participants in the *high Stakes* context reporting significantly higher transparency than those in the *low Stakes* context ($\beta = 3.17$, $SE = 1.61$, $p = .049$, $OR = 23.81$); see Figure 3(c). No moderating effect of *Strategy* emerged for `Apol` ($p = .380$) or `RR` ($p = .107$). Preceding transparency perceptions from **Block 1** ($\beta = 2.12$, $SE = 0.34$, $p < .001$, $OR = 8.36$) and **Block 2** ($\beta = 1.53$, $SE = 0.31$, $p < .001$, $OR = 4.61$) significantly predicted transparency ratings in **Block 3**.

Our findings reveal no significant effect of *Stakes* on perception of AI transparency across **Block 1** ($\beta = 1.33$, $SE = 0.75$, $p = .076$) and **Block 2** ($\beta = 0.89$, $SE = 0.86$, $p = .299$). In **Block 3**, however, *low Stakes* is perceived as significantly more transparent than *high Stakes*; $\beta = 3.07$, $SE = 1.53$, $p = .044$, $OR = 21.63$.

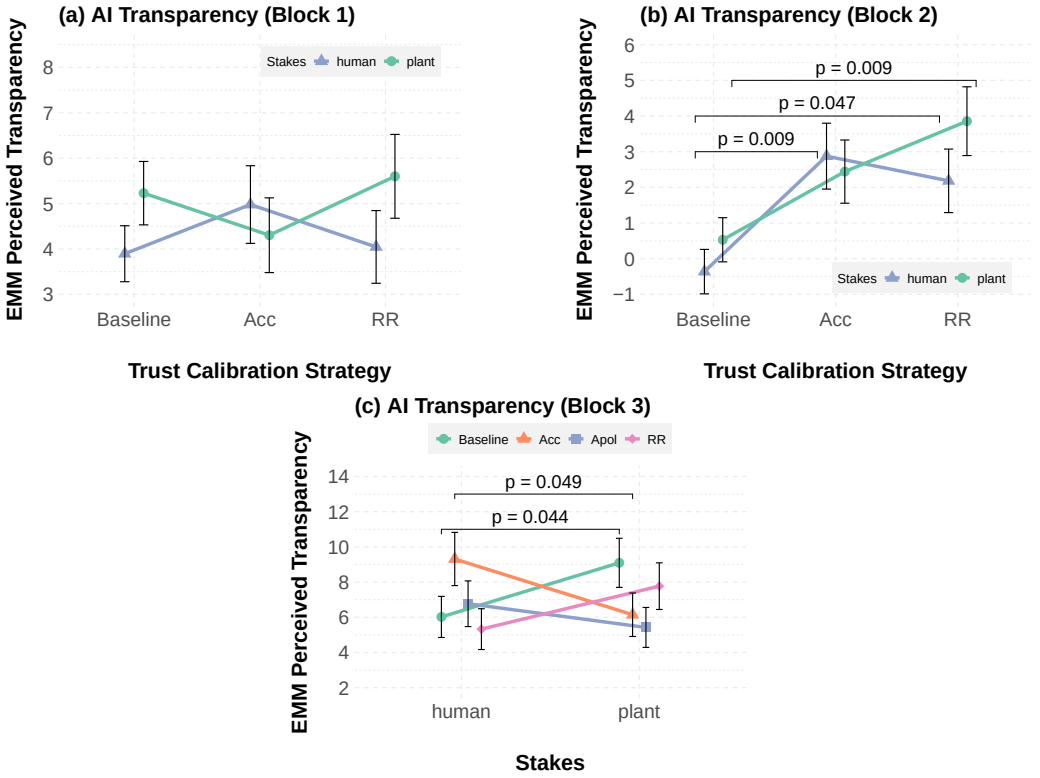


Figure 3: The impact of *Strategy* × *Stakes* on perceptions of AI transparency across Blocks of varying AI performance (1-3). Post-hoc faceted contrasts. Estimated marginal means (EMMs) of perceived AI transparency. Error bars denote standard error (SE).

4.2.3. Perceived AI Performance

Our analysis showed no significant effect of *Strategy* on participants' perceptions of AI performance in **Block 1**, with neither `Acc` ($\beta = -0.65$, $SE = 0.99$, $p = .516$) nor `RR` ($\beta = -1.77$, $SE = 0.99$, $p = .074$) differing significantly from Baseline. Likewise, no significant interaction emerged between *Stakes* and *Strategy* (all $p > .05$); see Figure 4(a). In **Block 1**, participants' dispositional trust in automation significantly predicted their performance perceptions of the AI agent ($\beta = 1.388$, $SE = 0.56$, $p = .013$, $OR = 4.01$).

After the AI exhibited failures in its assessment (**Block 2**), our analysis revealed a significant effect of *Strategy*, with `Acc` significantly increasing perceived AI performance compared to Baseline ($\beta = 2.63$, $SE = 1.06$, $p = .014$, $OR = 13.82$). No significant difference emerged for `RR` or `Apol`

relative to Baseline ($p > .05$). Again, participants dispositional trust in automation significantly predicted their performance perceptions ($\beta = 1.63$, $SE = 0.56$, $p = .004$, $OR = 5.19$). Notably, faceted post-hoc contrasts further demonstrated that this effect was moderated by *Stakes*. Specifically, `Acc` ($EMM = 1.15$, $SE = 0.94$) increased performance perceptions compared to Baseline ($EMM = -1.48$, $SE = 0.72$) under *high Stakes*; $\beta = 2.63$, $SE = 1.06$, $p = .036$, $OR = 13.82$), whereas we observed no significant effect under *low Stakes* ($\beta = -1.91$, $SE = 1.04$, $p = .160$); see Figure 4(b). These results suggest that when the AI exhibited errors, AI accuracy communication prior to the interaction enhances perceived AI performance under *high Stakes*.

When the AI returned to provide correct assessments (**Block 3**), we found a significant effect of *Strategy*; `Acc` led to significantly higher perceptions of AI performance relative to Baseline ($\beta = 3.20$, $SE = 1.27$, $p = .012$, $OR = 24.42$). No significant differences were found for `RR` or `Apol` compared to Baseline ($p > .05$). Notably, faceted contrasts showed that `Acc` elicited significantly higher perceptions of AI performance compared to `RR` under *high Stakes* ($EMM = 8.80$, $SE = 1.24$ vs. $EMM = 4.08$, $SE = 0.91$; $\beta = 4.72$, $SE = 1.28$, $p = .001$, $OR = 112.17$), whereas no significant difference was observed under *low Stakes* ($\beta = -1.62$, $SE = 1.16$, $p = .505$); see Figure 4(c). Alongside this result, we found that *Strategy* moderated the effect of *Stakes* on perceived AI performance. Specifically, when presented with `Acc`, participants in the *high Stakes* scenario ($EMM = 8.80$, $SE = 1.24$), i.e. human tissue samples, reported significantly higher performance perceptions than those in the *low Stakes* scenario, i.e. the inspection and classification of plant tissue samples, ($EMM = 4.33$, $SE = 0.94$; $\beta = 4.47$, $SE = 1.28$, $p < .001$, $OR = 87.08$).

In **Block 3**, participants' dispositional trust in automation significantly predicted their performance perceptions ($\beta = 1.25$, $SE = 0.57$, $p = .028$, $OR = 3.50$), alongside the effect of preceding **Block 1** performance perceptions ($\beta = 1.74$, $SE = 0.34$, $p < .001$, $OR = 5.70$).

The *Stakes* inherent to the decision-making context did not significantly affect participants' perceptions of AI performances, neither when the AI produces correct assessments in **Block 1** ($\beta = 1.06$, $SE = 0.85$, $p = .214$) and **Block 3** ($\beta = 1.73$, $SE = 1.24$, $p = .161$), nor when the AI exhibited erroneous assessments in **Block 2** ($\beta = -0.79$, $SE = 0.85$, $p = .353$).

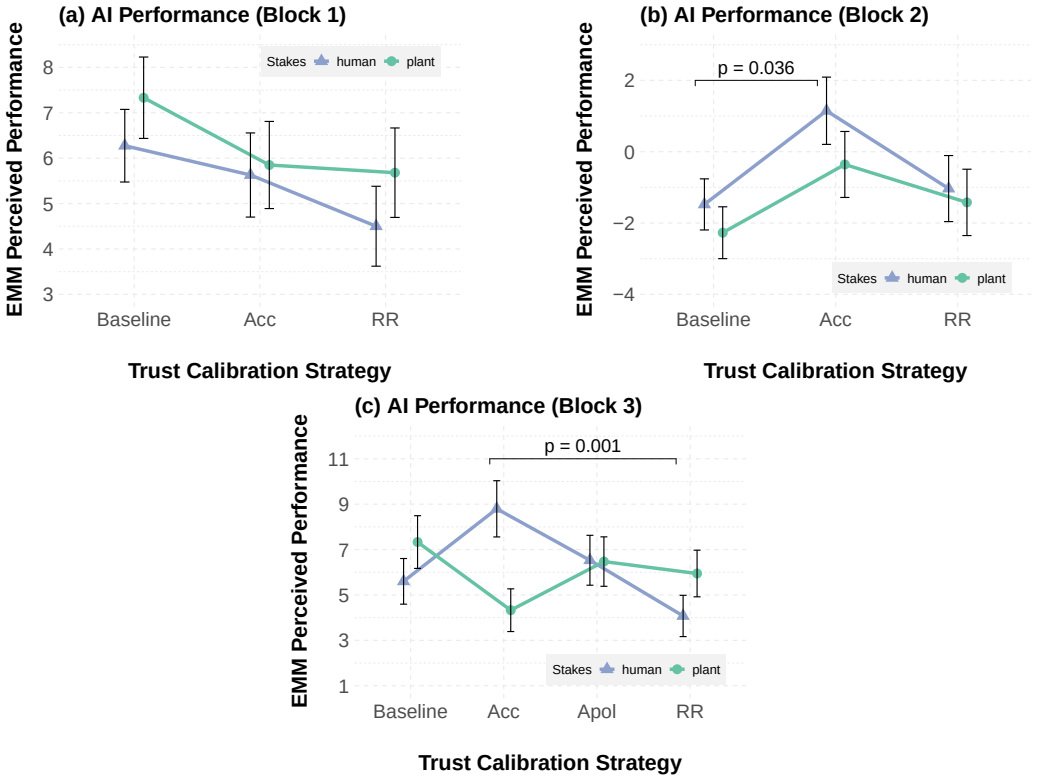


Figure 4: The impact of *Strategy* × *Stakes* on perceived AI performance across Blocks of varying AI performance (1-3). Post-hoc faceted contrasts. Estimated marginal means (EMMs) of perceived AI performance. Error bars denote standard error (SE).

4.2.4. Perceived AI Utility

Across the initial error-free Block (**Block 1**) and the subsequent erroneous Block (**Block 2**), we found no significant effect of *Strategy* on participants' perceptions of AI utility compared to Baseline (all $p > .05$). Moreover, no significant interactions with *Stakes* emerged, faceted contrasts confirmed the absence of any moderation effects (all $p > .05$). Our findings suggest that perceived AI utility is resilient to both AI accuracy and Relational Risk signalling. Participants' dispositional trust in automation significantly predicted utility perceptions in both **Block 1** ($\beta = 0.78$, $SE = 0.30$, $p = .009$, $OR = 2.18$) and **Block 2** ($\beta = 0.95$, $SE = 0.39$, $p = .015$, $OR = 2.58$), while preceding block-level utility perceptions emerged as an additional significant predictor in **Block 2** ($\beta = 0.57$, $SE = 0.18$, $p = .002$, $OR = 1.78$).

In **Block 3**, when the AI returned to providing accurate assessments,

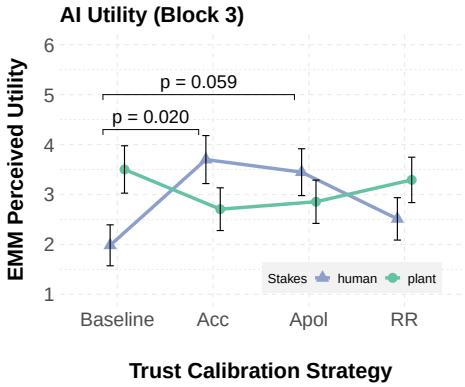


Figure 5: Perceived AI Utility in Block 3. Error bars denote standard error (SE).

our analysis revealed significant effects of *Strategy*, with `Acc` ($\beta = 1.72$, $SE = 0.59$, $p = .004$, $OR = 5.57$) and `Apol` ($\beta = 1.46$, $SE = 0.58$, $p = .012$, $OR = 4.33$) significantly increasing perceived AI utility compared to Baseline. Post-hoc faceted contrasts further demonstrate *Stakes* as a moderator; `Acc` ($EMM = 3.70$, $SE = 0.48$) led to greater perceived utility when the AI agent assisted under *high Stakes*, i.e. human tissue samples, compared to Baseline ($EMM = 1.98$, $SE = 0.41$); $\beta = 1.72$, $SE = 0.59$, $p = .020$, $OR = 2.18$; see Figure 5), whereas in the *low Stakes* plant inspection contexts, the effect was directionally reversed but not significant ($\beta = -0.80$, $SE = 0.59$, $p = .526$). A comparable albeit weaker trend was observed for `Apol`: under *high Stakes*, `Apol` was associated with somewhat higher utility ratings than Baseline, though this difference did not reach significance ($p = .059$); under *low Stakes*, the effect was again reversed but non-significant. *Strategy* further moderated the effect of *Stakes* on perceived AI utility in **Block 3**. Specifically, under the Baseline condition, participants in the *high Stakes* context (human tissue samples) reported significantly lower AI utility perceptions compared to those in the *low Stakes* context (plant tissue samples); $\beta = -1.52$, $SE = 0.59$, $p = .010$, $OR = 4.57$. Both preceding **Block 1** ($\beta = 1.82$, $SE = 0.18$, $p < .000$, $OR = 6.18$) and **Block 2** ($\beta = 0.38$, $SE = 0.14$, $p = .006$, $OR = 1.46$) utility perceptions emerged as an additional significant predictor in **Block 3**.

Similarly to perceptions of AI transparency, our findings show no significant effect of *Stakes* on perceived AI utility across **Block 1** ($\beta = 0.242$, $SE = 0.46$ $p = .596$) and **Block 2** ($\beta = -0.20$, $SE = 0.59$, $p = .730$). Notably, *low Stakes* led to a significantly increased perception of AI util-

ity compared to *high Stakes* in **Block 3**; $\beta = 1.52$, $SE = 0.59$, $p = .010$, $OR = 4.57$.

4.3. Qualitative Data Analysis

We analysed our qualitative data following Thomas' general inductive approach (Thomas, 2006) to gain an in depth understanding of our research objectives. Following the steps outlined in Thomas (2006), two researchers from our research team coded the open-ended responses provided in the post-task questionnaire. While guided by our evaluation objectives with respect to RQ1–3, as proposed by Thomas (2006), we specifically sought to gain an in-depth understanding of the mechanisms by which the *Strategies* operate in calibrating trust, as well as RQ4, i.e., how the communication of Relational Risk (`RR`) compared to other *Strategies* shapes expectations and users' understanding of the AI agent under varying AI performance and *Stakes*. The first coder thoroughly familiarised themselves with the data and developed an initial set of preliminary categories from the raw open-ended responses, while the second coder independently developed a second set of categories. Subsequently, in four iterative rounds, the two coders reviewed, compared and discussed their categories to eliminate any ambiguities and discrepancies and to establish an accurate and shared interpretation of the data. The two coders refined and merged categories to then deductively code the qualitative responses until a satisfying coding alignment was reached (Thomas, 2006; McHugh, 2012); with an inter-rater reliability of Cohen's $\kappa = 0.81$. We highlight the categories we identified following the general inductive approach in bold (Thomas, 2006), and present participants' responses to exemplify our findings while contextualising them with a 3-tuple (PID, *Stakes*, *Strategy*) to reflect their experimental condition.

4.3.1. AI Accuracy as a Strategy to shape perceptions of AI

Communicating AI accuracy (`Acc`) as an a priori performance indicator made a considerable number of participants **aware that the AI could make mistakes**; *"I was told that the AI had a 78% accuracy. I kept that in mind when deciding if I agreed or didn't with it. So, I think while I trusted it, I knew it had some wiggle room for errors."* – (P20, human, `Acc`). This awareness often translated into **slightly tempered expectations** of the AI agent; *"Due to the error margin, I expected some inaccurate suggestions by the AI agent, but not so much. Hence, the trust on the AI agent wasn't that high."* – (P115, plant, `Acc`)

Alongside these impressions, our qualitative analysis revealed that several participants reported to be **attentive as a result of the stated AI accuracy**, though to varying degrees. For example, one participant described being *"cautiously optimistic,"* noting, *"I understood it was fairly reliable but not infallible, so I approached its responses with a balanced mix of trust and scepticism"* – (P123, plant, `Acc`). Others expressed a more apprehensive stance; *"I expected the AI to make correct guesses, but there was still room for mistakes. So I had to be very attentive"* – (P13, human, `Acc`).

Conversely, several participants articulated that the communicated AI accuracy **raised their performance expectations** and instilled confidence in the AI's performance; *"I expected it to work pretty well overall."* – (P4, human, `Acc`); *"It made me very confident in its abilities"* – (P105, plant, `Acc`). These statements align with our quantitative findings, which showed that `Acc` significantly enhanced perceptions of AI performance when the AI erred (Block 2) and returned to correct identifications (Block 3) compared to the Baseline, and compared to communicating Relational Risk (`RR`) in Block 3, all notably so under *high Stakes*.

4.3.2. Relational Risk as a Strategy to shape perceptions of AI

In line with AI accuracy, the communication of Relational Risk (`RR`) prior to the interaction prompted participants to **anticipate error occurrences in the AI's outputs**; *"I knew it can and will make mistakes"* – (P80, human, `RR`); *"it made me aware that the AI may give incorrect information"* – (P194, plant, `RR`). The recognition of the possibility of errors led many to **recalibrate their expectations of the AI**; *"It prepared me mentally in that I was expecting some but not most of its answers to be wrong. I did not trust it 100%."* – (P79, human, `RR`); *"The message about the AI's 22% error rate set my expectations for occasional mistakes, prompting me to approach its recommendations with a critical and nuanced perspective."* – (P95, human, `RR`).

As hinted in the latter statement, the communication of Relational Risk **elicited a more vigilant stance**; *"Being challenged to be careful before making a choice based on the AI"* – (P189, plant, `RR`), and *"It makes you pay more attention, to verify, and not rely on the AI"* – (P205, plant, `RR`), and frequently **instilled a heightened sense of caution** *"It made me more cautious about blindly accepting its recommendations"* – (P96, human, `RR`).

As a result, several participants reported to **scrutinise the AI more thoroughly**, which manifested in a more deliberate and diligent assessment

of the AI's output as a behavioural consequence; *"I found myself scrutinising the microscope images much more closely, I'd do my own double-check. It added an extra layer of cognitive effort because I couldn't just trust it implicitly any more."* – (P96, human, `RR`)

Notably, the communicated error rate, as an inherent part of the Relational Risk approach to shape trust perceptions and expectations, was perceived differently across participants. Some participants regarded the 22% error rate as high and responded with scepticism:*"It made me very cautious as that's a large margin of error"* – (P91, human, `RR`). Some participants, on the contrary, considered the error rate as acceptable, emphasising that while it increased awareness for the AI's potential fallibility, it did **not erode the AI's overall utility**, which is in line with our quantitative findings (see Section 4.2.4); *"I began to have cautious faith in the AI agent because I was aware of its 22% error rate up front, which helped me realise that it's a useful but limited tool"* – (P203, plant, `RR`), and *"My expectations were that it would get some wrong but it still ended up being useful in the end"* – (P82, human, risk).

4.3.3. Apology as a Strategy to shape perceptions of AI

Our qualitative findings revealed mixed results regarding the effect and interpretation of AI apology (`Apol`) as a mechanism to influence trust perceptions. On the one hand, the `Apol` strategy was often interpreted as **the AI taking accountability for its errors** and as a promise for improved future performance, which contributed to restoring trust; *"That message showed accountability, which helped rebuild trust. It signalled the AI recognises mistakes and aims to improve"* – (P29, human, `Apol`). On the other hand, several participants regarded the `Apol` strategy to be **ineffective in calibrating trust or supporting system understanding**. *"This message did nothing for me. An AI apologising is pointless and means absolutely nothing."* – (P140, plant, `Apol`). Beyond outright dismissal, some participants stressed that an AI apology alone could not calibrate their trust in the AI unless it was accompanied by evidence of improvement in the AI's behaviour; in this sense, trust was conditional on observable corrections following the apology. *"I was glad but it did not affect my expectations as I needed proof to see it was working well first. That was just an apology."* – (P45, human, `Apol`). A few participants further emphasised that its ineffectiveness also originated from the lack of explanatory value, as the message did not support their understanding of the AI system; *"It didn't affect my trust at all because it wasn't*

an explanation for why it made the errors." – (P30, human, `Apol`).

4.3.4. Stakes

Our qualitative analysis revealed some **awareness of the *Stakes* inherent to the AI's decision-making context**, albeit limited. Notably, most of these reflections emerged from the *high Stakes* group, where participants emphasised the potentially severe consequences associated with the diagnostic task involving human tissue samples and the implications for patients' health and life; *"People's lives depend on these judgements"* – (P26, human, `Acc`). When considering the assigned *Strategy*, the majority of expressions originated from participants in the *high Stakes* group who received Relational Risk (`RR`) as an intervention *Strategy* prior to the interaction; *"[...] my pick could harm real life patients."* – (P89, human, `RR`), and *"it caused me to pause and reassess how much I could rely on it. Even a single error in a clinical or diagnostic context can have significant consequences."* – (P86, human, `RR`)

5. Discussion

Real-world consequences due to misplaced trust in fallible AI systems (Cao et al., 2024; McGregor, 2021) warrant interventions to reduce the gap between actual and perceived trustworthiness in AI-assisted decision-making across applications (Schlicker et al., 2025a). Our study responds to this concern by examining the communication of Relational Risk, as defined by Stuck et al. (2022), as a *Strategy* to shape trust perceptions under varying AI performance, and by comparing it to established existing approaches, i.e., AI accuracy and apology. We find that Relational Risk, operationalised through communicating the likelihood of an AI error and its consequences (Rowe, 1975; Dunlop et al., 2008) while adopting a negative framing (McNeil et al., 1982; Peters et al., 2011), fosters transparency perceptions without undermining utility or inducing AI-averse perceptions. We also identify *Decision-Stakes* as an important moderator for the *effectiveness* and *appropriateness* of these *Strategies*. Our work provides a fine-grained assessment of perceived AI trustworthiness across its dimensions of transparency, performance, and utility (RQ1), user understanding of the AI (RQ2), and situates these dimensions within varying *Stakes* (RQ3). We extend recent work on communicating AI (in)capabilities and limitations (Rieger et al., 2025; He et al., 2023; Kwon et al., 2018), and contribute to shift the conversation in human-AI interaction to explicitly communicating risk to users in-situ,

while integrating risk theory. In the following, we discuss our results and offer important implications for AI trustworthiness research in the context of AI-assisted decision-making.

5.1. Shaping Trust Perceptions and System Understanding across varying AI Performance

5.1.1. The Crux of Transparency

Recent years present a growing body of literature demonstrating transparency as a central requirement for trustworthy AI systems (Bhatt et al., 2021; Prabhudesai et al., 2023). Prior work emphasises that communicating uncertainty enhances transparency and supports the appropriate formation of trust between humans and AI agents (Bhatt et al., 2021), while withholding such information fosters a false sense of perfectly reliable systems (Prabhudesai et al., 2023). Our results extend this discussion. After erroneous AI behaviour, Relational Risk communication emerges as a consistent means of enhancing transparency perceptions across stakes. Our results support findings by Rieger et al. (2025), showing that communicating AI weaknesses can enhance perceptions of AI transparency. This finding is interesting, since approaches that reliably enhance transparency are desirable. Notably, AI accuracy information produces a context-selective and somewhat paradoxical effect: rather than leading users to question system transparency under high-stakes, it increased transparency perceptions solely in the high-stakes decision context after the AI exhibited erroneous identifications. This pattern warrants caution, as high-stakes contexts ideally should prompt greater scrutiny and information-seeking rather than elevated transparency perceptions. Our qualitative findings revealed that both communicating Relational Risk and AI accuracy made participants anticipate potential AI errors, which contributes to perceptions of AI transparency. While this may explain part of the effect, another may be that accuracy information, by acknowledging fallibility while adhering to a positive framing (Spiegelhalter, 2017), created a sense of (positively interpreted) predictability or controllability perceived as particularly relevant in the high-stakes context, which is known to decrease risk perceptions (Wilson et al., 2019; Walpole and Wilson, 2021), and foster risk-seeking behaviour (McNeil et al., 1982; Peters et al., 2011). When the AI then erred, participants' expectations were confirmed and may have reinforced their perception similar to a confirmation bias (Passi and Vorvoreanu, 2022). The absence of consequence information could have strengthened such positive reinforcement, missing to nudge people's cognitive engagement with

the negative consequence associated with the high-stakes context. Another explanation could be that, under low-stakes, accuracy information may be seen as simple background noise; whereas in high-stakes decisions, the same information becomes a valuable signal to shift transparency perceptions. Our results contribute to prior findings showing that users exhibit greater convergence with AI support compared to human support under high stakes, and that AI advice attracts higher behavioural trust when stakes are high (Fahnenstich et al., 2024). However, more research is needed to understand such intervention strategies and their effects on AI transparency under varying decision-stakes.

Notably, following repeated interactions with both correct and erroneous AI identifications (Block 3), *Strategy* also moderated the effect of *Stakes* on perceived transparency. Whereas participants in the Baseline (no information) condition judged the high-stakes context as significantly less transparent than the low-stakes context, AI accuracy information led participants to perceive the high-stakes context as significantly more transparent. However, when Relational Risk is communicated, there is no effect. Similar to explanations described beforehand, this could be due to the positive frame valence of the approach (accuracy rate vs. error rate), which reassures users in high-stakes contexts and inflates transparency perceptions, whereas Relational Risk communication keeps awareness of potential failure salient and relevant across stakes. Qualitative insights support this interpretation, as explicit references to consequences concentrated predominantly in the high-stakes `RR` condition and appeared rarely in the `Acc` condition. Although the communicated consequence in `RR` concerned Relational Risk rather than situational stakes, it may still have shaped how participants interpreted the stakes, possibly through a more critical cognitive engagement. This pattern is concerning: users receiving AI accuracy information may be satisfied with minimal disclosures precisely where more comprehensive transparency is warranted. It demonstrates the danger of inappropriate intervention cues in human-AI interaction, which risk overstating the openness of AI systems in contexts where consequences are most severe; creating an illusion of transparency where users should request more information. **We argue that AI accuracy information should solely serve as a means to support people's perception of AI transparency in low-stakes contexts.** We further propose **Relational Risk as a candidate strategy for enhancing perceived AI transparency that warrants further investigation**, given that it maintains the anticipation of potential failures and encourages

critical engagement across decision contexts, which we discuss in Section 5.1.4.

Introduced after erroneous AI outputs, apology neither significantly decreased nor enhanced perceptions of transparency. Qualitative insights suggest that while some participants perceived the apology as the AI taking accountability for its mistakes, many regarded it as ineffective, sometimes noting that it did not support their system understanding. This indicates that apology offers little value for enhancing perceived transparency. These findings contribute to recent works showing that while apology is effective for increasing trust post failure (Pareek et al., 2024), even more so than simple explanations (Kox et al., 2021), it might not sufficiently support transparency due to its nature as a reactive strategy (LeMasurier et al., 2024).

5.1.2. Performance Perceptions: The Double-Edged Role of Accuracy

For users interacting with AI systems, understanding and anticipating an AI's performance, i.e. forming expectations about how well an AI system will perform, are essential to determine whether and how much to trust it (Roesler* et al., 2022; Lee and See, 2004). Nevertheless, as Passi and Vorvoreanu (2022) argue, this performance calibration is often lacking, which constitutes one of the central challenges for appropriate trust calibration.

Our work provides several insights into how different *Strategies* influence people's performance perceptions. AI Accuracy and Relational Risk information both tempered expectations and encouraged more cautious and attentive engagement with the AI agent and its outcomes, as demonstrated in our qualitative findings. Risk communication induced a particularly vigilant stance, supporting results on exposing AI limitations prior to human-AI interaction (Rieger et al., 2025). Our quantitative findings reveal that stated AI accuracy significantly increased performance perceptions compared to the baseline after the AI exhibited erroneous identifications, and compared to communicating Relational Risk once the AI system returned to accurate behaviour in the final Block. This 'boosting' effect may prevent end users from turning away from the AI after the system exhibits failures. Such an effect could be advantageous in low-stakes applications or domains where expertise primarily lies with the AI, as well as in complementary Human–AI scenarios in which AI expertise is superior and reliance on AI support may be warranted (Bansal et al., 2021; Pareek et al., 2024). However, increased AI performance perceptions induced by the communication of AI accuracy were most pronounced in the high-stakes decision-making context, where elevated

perceptions of AI performance may be problematic and where a certain degree of wariness may be necessary. Our qualitative insights confirm that AI accuracy instilled increased performance expectations and the belief that the AI would be performing well. When looking at this result through the lens of risk communication research, one explanation may lie in the positive nature of the communication itself. AI accuracy presents the AI in terms of successful outcomes, which has been shown to reduce risk perception and promote more risk-seeking decisions (McNeil et al., 1982; Peters et al., 2011; Spiegelhalter, 2017). Importantly, while this positive framing significantly boosted performance perceptions, the negatively-framed Relational Risk strategy did not produce the inverse effect: it did not significantly decrease performance perceptions compared to the baseline. Our findings contribute to explaining the findings by Lai and Tan (2019) that report increased trust even for small accuracy levels, and accuracy being perceived as a stable (positive) 'trait' of the AI model (Rechkemmer and Yin, 2022).

Our work demonstrates the drawback of accuracy communication: **AI accuracy can inflate AI performance perceptions in the system precisely where scepticism is warranted, creating a risk of misplaced trust in critical contexts**. At the same time, it suggests that **accuracy communication may still be a viable strategy in low-stakes contexts, where heightened performance perceptions are desired and less consequential.**

5.1.3. Communicating Relational Risk does not diminish AI Utility

During the initial error-free performance of the AI agent and the subsequent block of erroneous assessments, neither AI accuracy communication nor Relational Risk communication significantly altered perceptions of AI utility. This stability is noteworthy. Despite its negative framing, Relational Risk communication led the AI to be perceived as "a useful but limited tool", without inducing signs of algorithmic aversion (Passi and Vorvoreanu, 2022). Hence, **communicating the Relational Risk does not diminish perceived utility, which indicates that users continue to see value in the AI system even when its fallibility is explicitly foregrounded.** This is in line with the objectives of effective risk communication — to "alert the target audience", but "without arousing unnecessary anxiety" (Breakwell, 2000). Our finding contributes to results by Rieger et al. (2025) which show that communicating AI limitations may reduce initial trust, but does not erode it over the course of interaction. Qualitative insights add nuance

to this interpretation. Some participants emphasised the high error rate and described heightened risk perception, while others stated that the communicated error rate did not diminish their sense of utility. The emphasis on error rate raises the question of whether it is the likelihood of the AI erring or the consequence component of Relational Risk that constitutes the more prominent determinant of perceived AI trustworthiness. Notably, after participants observed varying AI performance during repeated interactions, both accuracy and apology significantly increased perceptions of AI utility compared to the Baseline (Block 3). Accuracy, again, significantly elevated the perceived usefulness of the AI in precisely the context where misplaced trust could be most problematic, i.e. high-stakes, which re-emphasises the drawback of AI accuracy described beforehand.

5.1.4. Relational Risk as a Cognitive Forcing Function

The communication of Relational Risk elicited healthy scrutiny toward the AI which often manifested in a deliberate assessment of the AI's output, e.g. double-checking its identification. This effect resonates with the idea of cognitive forcing functions as an approach to interrupt users' tendencies to operate in autopilot mode and instead encourage more deliberate, System 2 thinking (Kahneman, 2011; Buçinca et al., 2021; Park et al., 2019). It also echoes the need to implement interventions that elicit human oversight (Passi and Vorvoreanu, 2022). If we aspire to make "users the last line of defense against AI failures" in real-world AI applications (Passi and Vorvoreanu, 2022), we must equip users with the ability to recognise and navigate both the capabilities and limitations of these systems to elicit appropriate perceptions of AI trustworthiness given associated risks. This requires encouraging users to approach AI decisions and recommendations with careful assessment rather than applying heuristics, and to promote a more critical engagement with AI outputs in AI-assisted decision-making (Buçinca et al., 2021; Park et al., 2019; Shen et al., 2024). **Our findings suggest that Relational Risk communication can function as a cognitive forcing function in AI-assisted decision-making.** Future work should investigate its effectiveness on perceived trustworthiness and cognitive engagement compared to existing cognitive forcing functions such as "on demand", "update", or "waiting" features (Buçinca et al., 2021).

However, Relational Risk communication is not without drawbacks. While it can foster healthy scrutiny and deliberate assessment, in low-stakes contexts such vigilance may impose unnecessary cognitive load or time costs

(Hoesterey and Onnasch, 2023). In these settings, AI accuracy information may be preferable, as it induced attentiveness without prompting undue scrutiny. Prior work suggests that scrutiny and verification behaviour in high-risk automation contexts may decline over time due to habituation (Hoesterey and Onnasch, 2023); thus, **a key question is whether Relational Risk communication can sustain caution without leading users to become habituated to the Relational Risk communication in human-AI interaction**.

Our findings indicate that AI accuracy and Relational Risk communication each offer strengths under different conditions. Recognising this complementarity lays the groundwork for future adaptive approaches that could investigate *RR* as an adaptive strategy following Buçinca et al. (2024)'s offline reinforcement learning approach for adaptive support policies that provide assistance based on e.g., individual differences and task context to determine when *RR* should be conveyed, which may also help address potential habituation effects. In line, we implemented Relational Risk as an a priori communication. Future research should examine its implementation alongside AI outputs 'in-the-moment' (Buçinca et al., 2024), akin to model confidence, and explore the balance between effectiveness and a potential "crying wolf" effect (Wickens et al., 2009) of Relational Risk communication. That is, investigating the potential risk that a high false-alarm rate and repetitive risk signalling could desensitise users in AI-assisted decision-making over repeated interactions (Wickens et al., 2009; Cañas, 2022; LeClerc and Joslyn, 2015).

5.2. The Role of Stakes in Shaping Trust Perceptions in AI-assisted decision-making

There is no universally applicable rule for what constitutes appropriate trust, since it is intrinsically linked to context and application (Passi and Vorvoreanu, 2022). While following a false AI recommendation may be tolerable in a low-stakes context such as plant diagnostic, accepting an erroneous decision in a high-stakes context such as medical diagnosis can be detrimental. Any intervention designed to shape trust in AI must therefore remain sensitive to the inherent stakes of the decision-making context and foster a calibration of trust in the appropriate direction (Lee and See, 2004).

In our study, participants perceived the AI system as less transparent and less useful after experiencing varying AI performance when the stakes was high. More importantly so, as we discussed above, decision-stakes moderated

the influence of communication strategies on all three dimensions of trustworthiness; i.e. transparency, performance, and utility. We therefore argue that the*Strategy* to shape trust in AI systems cannot be examined in a vacuum, but must be contextualised within the stakes in which they are embedded. Importantly, our findings show that stakes influence perceived AI trustworthiness itself, rather than solely playing a role at the level of behavioural trust, where the Organizational Trust Model (Mayer et al., 1995) and Enhanced Model of Trust (Stuck et al., 2022) situated the effect of stakes. Our results instead align with the recently proposed trustworthiness assessment model (Schlicker et al., 2025a), which posits an initial risk–benefit assessment (Schlicker et al., 2025b) during the formation of trustworthiness perceptions, and we provide empirical support for this theoretical proposition. For HCI research moving forward, this implies a critical shift: stakes is not merely a background variable but a decisive moderator that determines the effectiveness and, crucially, the appropriateness of strategies designed to influence trust in AI systems. Effectiveness captures whether a strategy successfully alters perceptions (e.g., boosting perceived AI performance), while appropriateness concerns whether such changes are desirable in light of the decision context. A strategy that inflates trust under high-stakes conditions may be effective but not appropriate, whereas in low-stakes settings the same intervention may be both effective and appropriate. **This implies that interventions to shape trust perceptions must be designed with decision-stakes in mind, acknowledging that the same *Strategy* can have different effects depending on the stakes of the task.**

5.3. Limitations

Our study defined risk communication as Relational Risk, that is, the risk of using the system itself, which is conceptually distinct from contextual factors (Stuck et al., 2022). We encourage future work to examine the communication of stakes as a consequence component to emphasise risks arising from the context in which the AI is deployed. We acknowledge that Relational Risk conceptually consists of two distinct components and we further encourage future work to examine the relative contribution of each component. All *Strategies* were implemented in line with their conceptual purpose, which resulted in different points of presentation. We acknowledge this as a potential confound and addressed it by fitting statistical models for each block. We focused on perceptions of AI trustworthiness as a foundational first step to understand the mechanisms and effects of Relational Risk communi-

cation in the context of AI-assisted decision-making. Based on our promising findings, we encourage future research to investigate how communicating Relational Risk influences behavioural trust and whether our findings extend to reliance behaviour in human-AI interaction (Bansal et al., 2021; Buçinca et al., 2021). To carefully assess participants' perceptions across varying performances, we deliberately designed observable AI errors for the purpose of this work, while recognising that AI errors may not always be evident in real-world scenarios. Finally, while the visual estimation task of assessing microscope images is not an everyday scenario, it enabled a directly comparable operationalisation of stakes across contexts. We encourage future work to investigate Relational Risk communication in real-world applications and in the context of LLMs to complement current approaches on uncertainty communication (Kim et al., 2024).

6. Conclusion

AI systems and decision support applications remain imperfect when deployed and accessible to users (McGregor, 2021) while spanning domains that range from the seemingly lightweight, such as music recommendations, to those with profound consequences for human health and life. This work introduces Relational Risk communication as an alternative *Strategy* in AI-assisted decision-making to shape perceptions of AI trustworthiness. We examined its effects alongside AI accuracy and AI apology across the dimensions of perceived AI transparency, performance, and utility, and in interaction with stakes. We further assessed their impact on system understanding. Our findings propose the communication of Relational Risk as a cognitive forcing function, which encourages conscientiousness, healthy scrutiny and a deliberate assessment of AI identifications, while also raising questions about potential costs in low-stakes contexts and risks of habituation that warrant further investigation. Imperatively, our results show that Relational Risk increases perceptions of transparency across low and high stakes and that the perceived AI utility remains resilient to risk signalling, despite the emphasis on the system's fallibility. Established strategies, i.e. AI accuracy and apology, each showed important strengths and drawbacks which we discuss. AI accuracy showed strengths in increasing and sustaining perceived transparency and performance, including post failure, but risked inflating trust in high-stakes settings where caution is warranted. AI apology improved perceived AI utility, but offered little support to enhance system understanding.

Across all dimensions, stakes significantly moderated these effects, which we discuss as an important determinant for the effectiveness and appropriateness of *Strategies* to shape trust perceptions. Relational Risk therefore emerges as an important approach for AI-assisted decision-making that offers an alternative and promising intervention to elicit appropriate trust in AI systems. More broadly, our work points to risk communication research as a critical lens for developing strategies to navigate trust dynamics in human-AI interaction and thus to inform future HCI research.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process.

During the preparation of this work, the authors used GPT-5 large language model and Claude Sonnet 4.5 as coding support for the implementation of the task in the survey (JavaScript, HTML) and for proofreading selected parts of the manuscript. After using these tools, the authors of this work carefully reviewed and edited the content as needed. The authors take full responsibility for the content of the manuscript.

Acknowledgments

Our work was supported by the Melbourne Research Scholarship funded by the University of Melbourne and the Rowden White Scholarship established with a donation from Sir Alfred Edward Rowden White. This research is partially supported by the Australian Research Council (Grant No. FT250100459, Grant No. CE260100108).

References

Armstrong, K., Schwartz, J.S., Fitzgerald, G., Putt, M., Ubel, P.A., 2002. Effect of Framing as Gain versus Loss on Understanding and Hypothetical Treatment Choices: Survival and Mortality Curves. Medical Decision Making 22, 76–83. URL: https://doi.org/10.1177/0272989X0202200108, doi:10.1177/0272989X0202200108. _eprint: https://doi.org/10.1177/0272989X0202200108.

Ashoori, M., Weisz, J.D., 2019. In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. URL:

http://arxiv.org/abs/1912.02675, doi:10.48550/arXiv.1912.02675. arXiv:1912.02675 [cs].

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D., 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan. pp. 1–16. URL: https://dl.acm.org/doi/10.1145/3411764.3445717, doi:10.1145/3411764.3445717.

Baranger, D.A.A., Finsaas, M.C., Goldstein, B.L., Vize, C.E., Lynam, D.R., Olino, T.M., 2023. Tutorial: Power analyses for interaction effects in cross-sectional regressions. Advances in Methods and Practices in Psychological Science 6. doi:10.1177/25152459231187531.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q.V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., Xiang, A., 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, ACM, Virtual Event USA. pp. 401–413. URL: https://dl.acm.org/doi/10.1145/3461702.3462571, doi:10.1145/3461702.3462571.

Breakwell, G.M., 2000. Risk communication: fators affecting impact. British Medical Bulletin 56, 110–120. URL: https://doi.org/10.1258/0007142001902824, doi:10.1258/0007142001902824.

Brown, S.L., Smith, E.Z., 2007. The inhibitory effect of a distressing anti-smoking message on risk perceptions in smokers. Psychology & Health 22, 255–268. URL: https://doi.org/10.1080/14768320600843127, doi:10.1080/14768320600843127. _eprint: https://doi.org/10.1080/14768320600843127.

Buçinca, Z., Malaya, M.B., Gajos, K.Z., 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 5, 1–21. URL: https://dl.acm.org/doi/10.1145/3449287, doi:10.1145/3449287.

Buçinca, Z., Pham, C.M., Jakesch, M., Ribeiro, M.T., Olteanu, A., Amershi, S., 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. URL: http://arxiv.org/abs/2306.03280, doi:10.48550/arXiv.2306.03280. arXiv:2306.03280 [cs].

Buçinca, Z., Swaroop, S., Paluch, A.E., Murphy, S.A., Gajos, K.Z., 2024. Towards Optimizing Human-Centric Objectives in AI-Assisted Decision-Making With Offline Reinforcement Learning. URL: http://arxiv.org/abs/2403.05911, doi:10.48550/arXiv.2403.05911. arXiv:2403.05911 [cs].

Cao, S., Liu, A., Huang, C.M., 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. Proceedings of the ACM on Human-Computer Interaction 8, 1–32. URL: https://dl.acm.org/doi/10.1145/3637318, doi:10.1145/3637318.

Cañas, J.J., 2022. AI and Ethics When Human Beings Collaborate With AI Agents. Frontiers in Psychology 13, 836650. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.836650/full, doi:10.3389/fpsyg.2022.836650.

De Miguel Velázquez, J., Šćepanović, S., Gvirtz, A., Quercia, D., 2024. Decoding Real-World Artificial Intelligence Incidents. Computer 57, 71–81. URL: https://ieeexplore.ieee.org/document/10718657/, doi:10.1109/MC.2024.3432492.

Derczynski, L., Kirk, H.R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M.R., Mohammad, S., 2023. Assessing Language Model Deployment with Risk Cards. URL: http://arxiv.org/abs/2303.18190, doi:10.48550/arXiv.2303.18190. arXiv:2303.18190 [cs].

Dunlop, S., Wakefield, M., Kashima, Y., 2008. Can You Feel It? Negative Emotion, Risk, and Narrative in Health Communication. Media Psychology 11, 52–75. URL: http://www.tandfonline.com/doi/abs/10.1080/15213260701853112, doi:10.1080/15213260701853112.

Fahnenstich, H., Rieger, T., Roesler, E., 2024. Trusting under risk – comparing human to AI decision support agents. Computers in Human Behavior

153, 108107. URL: https://linkinghub.elsevier.com/retrieve/pii/S0747563223004582, doi:10.1016/j.chb.2023.108107.

Field, A., Miles, J., Field, Z., 2014. Discovering statistics using R. Repr ed., Sage, Los Angeles, CA, USA.

Gong, J., Zhang, Y., Yang, Z., Huang, Y., Feng, J., Zhang, W., 2013. The framing effect in medical decision-making: a review of the literature. Psychology, Health & Medicine 18, 645–653. URL: https://www.tandfonline.com/doi/full/10.1080/13548506.2013.766352, doi:10.1080/13548506.2013.766352.

He, G., Buijsman, S., Gadiraju, U., 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. Proceedings of the ACM on Human-Computer Interaction 7, 1–29. URL: https://dl.acm.org/doi/10.1145/3610067, doi:10.1145/3610067.

Herdel, V., Šćepanović, S., Bogucka, E., Quercia, D., 2024. ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7, 584–596. URL: https://ojs.aaai.org/index.php/AIES/article/view/31660, doi:10.1609/aies.v7i1.31660.

Hoesterey, S., Onnasch, L., 2023. The effect of risk on trust attitude and trust behavior in interaction with information and decision automation. Cognition, Technology & Work 25, 15–29. URL: https://link.springer.com/10.1007/s10111-022-00718-y, doi:10.1007/s10111-022-00718-y.

Hoff, K.A., Bashir, M., 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Human Factors: The Journal of the Human Factors and Ergonomics Society 57, 407–434. URL: https://journals.sagepub.com/doi/10.1177/0018720814547570, doi:10.1177/0018720814547570.

Kahneman, D., 2011. Thinking, Fast and Slow. Allen Lane and Penguin Books, New York.

Kahr, P.K., Rooks, G., Snijders, C., Willemsen, M.C., 2024. The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness

to Follow AI Advice., in: Proceedings of the 29th International Conference on Intelligent User Interfaces, ACM, Greenville SC USA. pp. 609–622. URL: https://dl.acm.org/doi/10.1145/3640543.3645167, doi:10.1145/3640543.3645167.

Kim, S.S.Y., Liao, Q.V., Vorvoreanu, M., Ballard, S., Vaughan, J.W., 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA. pp. 822–835. URL: https://dl.acm.org/doi/10.1145/3630106.3658941, doi:10.1145/3630106.3658941.

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, 101595. URL: https://linkinghub.elsevier.com/retrieve/pii/S0736585321000344, doi:10.1016/j.tele.2021.101595.

Kim, T., Song, H., 2023. "I Believe AI Can Learn from the Error. Or Can It Not?": The Effects of Implicit Theories on Trust Repair of the Intelligent Agent. International Journal of Social Robotics 15, 115–128. URL: https://link.springer.com/10.1007/s12369-022-00951-5, doi:10.1007/s12369-022-00951-5.

Kocielnik, R., Amershi, S., Bennett, P.N., 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, Glasgow Scotland Uk. pp. 1–14. URL: https://dl.acm.org/doi/10.1145/3290605.3300641, doi:10.1145/3290605.3300641.

Kohn, S.C., De Visser, E.J., Wiese, E., Lee, Y.C., Shaw, T.H., 2021. Measurement of Trust in Automation: A Narrative Review and Reference Guide. Frontiers in Psychology 12, 604977. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.604977/full, doi:10.3389/fpsyg.2021.604977.

Kox, E.S., Kerstholt, J.H., Hueting, T.F., De Vries, P.W., 2021. Trust repair in human-agent teams: the effectiveness of explanations and

expressing regret. Autonomous Agents and Multi-Agent Systems 35, 30. URL: https://link.springer.com/10.1007/s10458-021-09515-9, doi:10.1007/s10458-021-09515-9.

Kwon, M., Huang, S.H., Dragan, A.D., 2018. Expressing Robot Incapability, in: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, Association for Computing Machinery, New York, NY, USA. pp. 87–95. URL: https://dl.acm.org/doi/10.1145/3171221.3171276, doi:10.1145/3171221.3171276.

Körber, M., 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation, in: Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y. (Eds.), Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018), Springer International Publishing, Cham. pp. 13–30. doi:10.1007/978-3-319-96074-6_2.

Lai, V., Tan, C., 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, Atlanta GA USA. pp. 29–38. URL: https://dl.acm.org/doi/10.1145/3287560.3287590, doi:10.1145/3287560.3287590.

LeClerc, J., Joslyn, S., 2015. The Cry Wolf Effect and Weather-Related Decision Making. Risk Analysis 35, 385–395. URL: https://onlinelibrary.wiley.com/doi/10.1111/risa.12336, doi:10.1111/risa.12336.

Lee, B., Cordeil, M., Prouzeau, A., Jenny, B., Dwyer, T., 2022. A Design Space For Data Visualisation Transformations Between 2D And 3D In Mixed-Reality Environments, in: CHI Conference on Human Factors in Computing Systems, ACM, New Orleans LA USA. pp. 1–14. URL: https://dl.acm.org/doi/10.1145/3491102.3501859, doi:10.1145/3491102.3501859.

Lee, J.D., See, K.A., 2004. Trust in Automation: Designing for Appropriate Reliance. Human Factors 46, 50–80. URL: https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392, doi:10.1518/hfes.46.1.50_30392. _eprint: https://journals.sagepub.com/doi/pdf/10.1518/hfes.46.1.50_30392.

LeMasurier, G., Gautam, A., Han, Z., Crandall, J.W., Yanco, H.A., 2024. Reactive or Proactive? How Robots Should Explain Failures, in: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, ACM, Boulder CO USA. pp. 413–422. URL: https://dl.acm.org/doi/10.1145/3610977.3634963, doi:10.1145/3610977.3634963.

Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G.S., Peng, L.H., Webster, D.R., Ai, D., Huang, S.J., Liu, Y., Dunn, R.C., Coz, D., 2020. A deep learning system for differential diagnosis of skin diseases. Nature Medicine 26, 900–908. URL: https://doi.org/10.1038/s41591-020-0842-3, doi:10.1038/s41591-020-0842-3.

Mayer, R.C., Davis, J.H., Schoorman, F.D., 1995. An Integrative Model Of Organizational Trust. Academy of Management Review URL: https://journals.aom.org/doi/abs/10.5465/amr.1995.9508080335, doi:10.5465/amr.1995.9508080335.

McGregor, S., 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. Proceedings of the AAAI Conference on Artificial Intelligence 35, 15458–15463. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17817, doi:10.1609/aaai.v35i17.17817. number: 17.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia Medica 22, 276–282. URL: https://doi.org/10.11613/BM.2012.031, doi:10.11613/BM.2012.031.

McNeil, B.J., Pauker, S.G., Sox, H.C., Tversky, A., 1982. On the elicitation of preferences for alternative therapies. New England Journal of Medicine 306, 1259–1262. URL: https://www.nejm.org/doi/full/10.1056/NEJM198205273062103, doi:10.1056/NEJM198205273062103, arXiv:https://www.nejm.org/doi/pdf/10.1056/NEJM198205273062103.

Nesset, B., Romeo, M., Rajendran, G., Hastie, H., 2023. Robot Broken Promise? Repair strategies for mitigating loss of trust for repeated failures, in: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1389–1395. URL:

https://ieeexplore.ieee.org/abstract/document/10309558, doi:10.1109/RO-MAN57019.2023.10309558.

Papenmeier, A., Englebienne, G., Seifert, C., 2019. How model accuracy and explanation fidelity influence user trust. URL: http://arxiv.org/abs/1907.12652, doi:10.48550/arXiv.1907.12652. arXiv:1907.12652 [cs].

Pareek, S., Schömbs, S., Velloso, E., Goncalves, J., 2025. "It's Not the AI's Fault Because It Relies Purely on Data": How Causal Attributions of AI Decisions Shape Trust in AI Systems, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. pp. 1–18. URL: https://dl.acm.org/doi/10.1145/3706598.3713468, doi:10.1145/3706598.3713468.

Pareek, S., Velloso, E., Goncalves, J., 2024. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, ACM, Rio de Janeiro Brazil. pp. 546–561. URL: https://dl.acm.org/doi/10.1145/3630106.3658924, doi:10.1145/3630106.3658924.

Park, J.S., Barber, R., Kirlik, A., Karahalios, K., 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. Proceedings of the ACM on Human-Computer Interaction 3, 1–15. URL: https://dl.acm.org/doi/10.1145/3359204, doi:10.1145/3359204.

Passi, S., Vorvoreanu, M., 2022. Overreliance on AI: Literature Review. Technical Report MSR-TR-2022-12. Microsoft. URL: https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/.

Peters, E., Hart, P.S., Fraenkel, L., 2011. Informing Patients: The Influence of Numeracy, Framing, and Format of Side Effect Information on Risk Perceptions. Medical Decision Making 31, 432–436. URL: https://journals.sagepub.com/doi/10.1177/0272989X10391672, doi:10.1177/0272989X10391672.

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q.V., Banovic, N., 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making, in: Proceedings of the 28th International Conference on Intelligent User Interfaces,

Association for Computing Machinery, New York, NY, USA. pp. 379–396. URL: https://doi.org/10.1145/3581641.3584033, doi:10.1145/3581641.3584033.

Rao, P.S.B., Šćepanović, S., Zhou, K., Bogucka, E.P., Quercia, D., 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan. pp. 1–26. URL: https://dl.acm.org/doi/10.1145/3706598.3713979, doi:10.1145/3706598.3713979.

Rawlings, J.O., Pantula, S.G., Dickey, D.A., 1998. Applied regression analysis: a research tool. Springer texts in statistics. 2nd ed ed., Springer, New York.

Rechkemmer, A., Yin, M., 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models, in: CHI Conference on Human Factors in Computing Systems, ACM, New Orleans LA USA. pp. 1–14. URL: https://dl.acm.org/doi/10.1145/3491102.3501967, doi:10.1145/3491102.3501967.

Rieger, T., Kugler, L., Manzey, D., Roesler, E., 2024. The (Im)perfect Automation Schema: Who Is Trusted More, Automated or Human Decision Support? Human Factors: The Journal of the Human Factors and Ergonomics Society 66, 1995–2007. URL: https://journals.sagepub.com/doi/10.1177/00187208231197347, doi:10.1177/00187208231197347.

Rieger, T., Schindler, H., Onnasch, L., Roesler, E., 2025. Explaining AI weaknesses improves human–AI performance in a dynamic control task. International Journal of Human-Computer Studies 199, 103505. URL: https://linkinghub.elsevier.com/retrieve/pii/S107158192500062X, doi:10.1016/j.ijhcs.2025.103505.

Roesler*, E., Rieger*, T., Manzey, D., 2022. Trust towards Human vs. Automated Agents: Using a Multidimensional Trust Questionnaire to Assess The Role of Performance, Utility, Purpose, and Transparency. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 66, 2047–2051. URL: https://journals.sagepub.com/doi/10.1177/1071181322661065, doi:10.1177/1071181322661065.

Rowe, W.D., 1975. An "anatomy" of Risk. Environmental Protection Agency. Google-Books-ID: O9JRAQAAMAAJ.

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M.C., Langer, M., 2025a. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). Computers in Human Behavior 170, 108671. URL: https://linkinghub.elsevier.com/retrieve/pii/S0747563225001189, doi:10.1016/j.chb.2025.108671.

Schlicker, N., Lechner, F., Wehrle, K., Greulich, B., Hirsch, M.C., Langer, M., 2025b. Trustworthy enough? Examining trustworthiness assessments of large language model-based medical agents. Technology, Mind, and Behavior 6, 1–29. URL: https://tmb.apaopen.org/pub/xo0swy4m, doi:10.1037/tmb0000164.

Schömbs, S., Goncalves, J., Johal, W., 2025. "I can feel the risks by looking at the robot face": Communicating Risk through a Physical Agent, in: Proceedings of the 2025 ACM Designing Interactive Systems Conference, Association for Computing Machinery, New York, NY, USA. pp. 236–252. URL: https://dl.acm.org/doi/10.1145/3715336.3735759, doi:10.1145/3715336.3735759.

Schömbs, S., Pareek, S., Goncalves, J., Johal, W., 2024. Robot-Assisted Decision-Making: Unveiling the Role of Uncertainty Visualisation and Embodiment, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. pp. 1–16. URL: https://dl.acm.org/doi/10.1145/3613904.3642911, doi:10.1145/3613904.3642911.

Sellers, R., 2013. How sliders bias survey data. MRA's Alert 53, 56–57.

Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Ma, Z., Petridis, S., Peng, Y.H., Qiwei, L., Rakshit, S., Si, C., Xie, Y., Bigham, J.P., Bentley, F., Chai, J., Lipton, Z., Mei, Q., Mihalcea, R., Terry, M., Yang, D., Morris, M.R., Resnick, P., Jurgens, D., 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. URL: http://arxiv.org/abs/2406.09264, doi:10.48550/arXiv.2406.09264. arXiv:2406.09264 [cs].

Slattery, P., Saeri, A.K., Grundy, E.A.C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., Thompson, N., 2025. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of

Risks From Artificial Intelligence. URL: http://arxiv.org/abs/2408.12622, doi:10.48550/arXiv.2408.12622. arXiv:2408.12622 [cs].

Spiegelhalter, D., 2017. Risk and Uncertainty Communication. Annual Review of Statistics and Its Application 4, 31–60. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-010814-020148, doi:10.1146/annurev-statistics-010814-020148.

Stuck, R.E., Tomlinson, B.J., Walker, B.N., 2022. The importance of incorporating risk into human-automation trust. Theoretical Issues in Ergonomics Science 23, 500–516. URL: https://www.tandfonline.com/doi/full/10.1080/1463922X.2021.1975170, doi:10.1080/1463922X.2021.1975170.

Thomas, D.R., 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. American Journal of Evaluation 27, 237–246. URL: https://journals.sagepub.com/doi/10.1177/1098214005283748, doi:10.1177/1098214005283748.

Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A., Bernstein, A., 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction, in: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, ACM, Utrecht Netherlands. pp. 77–87. URL: https://dl.acm.org/doi/10.1145/3450613.3456817, doi:10.1145/3450613.3456817.

Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T.M., Dixon, C., Tielman, M.L., 2020. Taxonomy of Trust-Relevant Failures and Mitigation Strategies, in: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, ACM, Cambridge United Kingdom. pp. 3–12. URL: https://dl.acm.org/doi/10.1145/3319502.3374793, doi:10.1145/3319502.3374793.

Visschers, V., Wiedemann, P., Gutscher, H., Kurzenhäuser, S., Seidl, R., Jardine, C., Timmermans, D., 2012. Affect-inducing risk communication: current knowledge and future directions. Journal of Risk Research 15, 257–271. URL: http://www.tandfonline.com/doi/abs/10.1080/13669877.2011.634521, doi:10.1080/13669877.2011.634521.

Visschers, V.H.M., Meertens, R.M., Passchier, W.W.F., De Vries, N.N.K., 2009. Probability Information in Risk Communication: A Review of the Research Literature. Risk Analysis 29, 267–287. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.2008.01137.x, doi:10.1111/j.1539-6924.2008.01137.x.

Walpole, H.D., Wilson, R.S., 2021. Extending a broadly applicable measure of risk perception: the case for susceptibility. Journal of Risk Research 24, 135–147. URL: https://www.tandfonline.com/doi/full/10.1080/13669877.2020.1749874, doi:10.1080/13669877.2020.1749874.

Wang, X., Yin, M., 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making, in: 26th International Conference on Intelligent User Interfaces, ACM, College Station TX USA. pp. 318–328. URL: https://dl.acm.org/doi/10.1145/3397481.3450650, doi:10.1145/3397481.3450650.

Wang, X., Yin, M., 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. ACM Transactions on Interactive Intelligent Systems 12, 1–36. URL: https://dl.acm.org/doi/10.1145/3519266, doi:10.1145/3519266.

Wang, Z.J., Kulkarni, C., Wilcox, L., Terry, M., Madaio, M., 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. pp. 1–40. URL: https://dl.acm.org/doi/10.1145/3613904.3642335, doi:10.1145/3613904.3642335.

Wickens, C.D., Rice, S., Keller, D., Hutchins, S., Hughes, J., Clayton, K., 2009. False Alerts in Air Traffic Control Conflict Alerting System: Is There a "Cry Wolf" Effect? Human Factors: The Journal of the Human Factors and Ergonomics Society 51, 446–462. URL: https://journals.sagepub.com/doi/10.1177/0018720809344720, doi:10.1177/0018720809344720.

Wilson, R.S., Zwickle, A., Walpole, H., 2019. Developing a Broadly Applicable Measure of Risk Perception. Risk Analysis 39, 777–791. URL: https://onlinelibrary.wiley.com/doi/10.1111/risa.13207, doi:10.1111/risa.13207.

Yin, M., Wortman Vaughan, J., Wallach, H., 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, Glasgow Scotland Uk. pp. 1–12. URL: https://dl.acm.org/doi/10.1145/3290605.3300509, doi:10.1145/3290605.3300509.

Zhang, Y., Liao, Q.V., Bellamy, R.K.E., 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA. pp. 295–305. URL: https://doi.org/10.1145/3351095.3372852, doi:10.1145/3351095.3372852.

Appendix A. Model Results

This section provides the detailed results of the quantitative analysis. Model specification: $TrustDim_b \sim Strategy \times Stakes + TiA - PtT + CentredAge + \sum_{j=1}^{b-1} TrustDim_{\text{Block } j} + (1 \mid PID) + (1 \mid Item)$. Here, TrustDim $\in$ {transparency, performance, utility} denotes perceived AI trustworthiness dimensions, $b \in \{1, 2, 3\}$ indexes the experimental block, *Strategy* refers to the intervention manipulation, and *Stakes* denotes situational risk, operationalised through the decision context.

Table A1: Block-specific CLMM formulas instantiated from the overarching model specification. All models include random intercepts for participant (PID) and item.

Block	Model formula
1	$\text{TrustDim}_{\text{Block 1}} \sim Strategy \times Stakes + TiA - PtT + CentredAge + (1 \mid PID) + (1 \mid Item)$
2	$\text{TrustDim}_{\text{Block 2}} \sim Strategy \times Stakes + TiA - PtT + CentredAge + \text{TrustDim}_{\text{Block 1}} + (1 \mid PID) + (1 \mid Item)$
3	$\text{TrustDim}_{\text{Block 3}} \sim Strategy \times Stakes + TiA - PtT + CentredAge + \text{TrustDim}_{\text{Block 1}} + \text{TrustDim}_{\text{Block 2}} + (1 \mid PID) + (1 \mid Item)$

Table A2: CLMM results for perceived AI **transparency** across Block 1–3. Italics denote covariates.

Dim = Transparency	**Block 1**				**Block 2**				**Block 3**			
Predictor	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
Stakes (plant)	1.334	0.752	1.774	.076	0.893	0.861	1.037	.300	3.074	1.529	2.011	.044*
Strategy: Accuracy	1.083	0.907	1.194	.233	3.238	1.103	2.934	.003**	3.291	1.615	2.038	.042*
Strategy: Relational Risk	0.150	0.880	0.170	.865	2.545	1.074	2.370	.018*	-0.694	1.478	-0.469	.639
Strategy: Apology	—	—	—	—	—	—	—	—	0.750	1.545	0.485	.628
TiA-PtT	0.694	0.490	1.417	.156	0.719	0.576	1.248	.212	1.029	0.722	1.426	.154
Centred Age	-0.030	0.020	-1.534	.125	-0.005	0.023	-0.200	.842	-0.040	0.029	-1.394	.163
Trust_Block1	—	—	—	—	1.245	0.256	4.856	<.001***	2.124	0.343	6.184	<.001***
Trust_Block2	—	—	—	—	—	—	—	—	1.529	0.306	5.004	<.001***
Stakes × Accuracy	-2.010	1.290	-1.558	.119	-1.326	1.517	-0.874	.382	-6.241	2.304	-2.709	.007**
Stakes × Relational Risk	0.221	1.299	0.170	.865	0.780	1.530	0.510	.610	-0.632	2.080	-0.304	.761
Stakes × Apology	—	—	—	—	—	—	—	—	-4.420	2.191	-2.018	.043*
Random Effects	Variance		Std. Dev		Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	9.847		3.138		16.538		4.067		22.330		4.725	
Item (Intercept)	0.158		0.398		0.051		0.226		0.000		0.000	

* $p < .05$, ** $p < .01$, *** $p < .001$. Em dashes indicate predictors not estimated in that block.

Table A3: CLMM results for perceived AI **performance** across Block 1–3. Italics denote covariates.

Dim = Performance	**Block 1**				**Block 2**				**Block 3**			
Predictor	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
Stakes (plant)	1.057	0.850	1.244	.214	-0.794	0.856	-0.927	.354	1.731	1.236	1.401	.161
Strategy: Accuracy	-0.645	0.994	-0.649	.516	2.626	1.063	2.470	.014*	3.195	1.272	2.512	.012*
Strategy: Relational Risk	-1.774	0.992	-1.788	.074.	0.444	1.039	0.427	.669	-1.525	1.150	-1.326	.185
Strategy: Apology	—	—	—	—	—	—	—	—	0.929	1.214	0.765	.444
TiA-PtT	1.388	0.558	2.488	.013*	1.647	0.565	2.916	.004**	1.252	0.569	2.199	.028*
Centred Age	0.009	0.022	0.392	.695	0.001	0.023	0.032	.974	-0.043	0.022	-1.917	.055.
Trust_Block1	—	—	—	—	0.136	0.295	0.461	.645	1.741	0.339	5.129	$<10^{-6}$***
Trust_Block2	—	—	—	—	—	—	—	—	0.172	0.253	0.681	.496
Stakes × Accuracy	-0.837	1.419	-0.590	.556	-0.712	1.478	-0.482	.630	-6.198	1.828	-3.391	$<.001$***
Stakes × Relational Risk	0.122	1.419	0.086	.932	0.406	1.466	0.277	.782	0.139	1.669	0.083	.934
Stakes × Apology	—	—	—	—	—	—	—	—	-1.789	1.744	-1.026	.305
Random Effects	Variance		Std. Dev		Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	11.745		3.427		16.433		4.054		13.193		3.632	
Item (Intercept)	0.018		0.135		0.426		0.653		0.270		0.520	

* $p < .05$, ** $p < .01$, *** $p < .001$. Em dashes indicate predictors not estimated in that block. A trailing dot denotes $p < .10$.

Table A4: CLMM results for perceived AI **utility** across Block 1–3. Italics denote covariates.

Dim = Utility	**Block 1**				**Block 2**				**Block 3**			
Predictor	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
Stakes (plant)	0.242	0.456	0.530	.596	-0.205	0.593	-0.345	.730	1.519	0.587	2.587	.010**
Strategy: Accuracy	-0.216	0.546	-0.396	.692	1.323	0.736	1.798	.072.	1.717	0.594	2.892	.004**
Strategy: Relational Risk	-0.643	0.538	-1.197	.231	0.219	0.723	0.302	.762	0.529	0.561	0.944	.345
Strategy: Apology	—	—	—	—	—	—	—	—	1.465	0.585	2.505	.012*
TiA-PtT	0.779	0.299	2.602	.009**	0.946	0.390	2.427	.015*	0.478	0.270	1.773	.077.
Centred Age	0.009	0.012	0.782	.434	-0.013	0.016	-0.797	.425	-0.008	0.011	-0.773	.440
Trust_Block1	—	—	—	—	0.575	0.182	3.164	.002**	1.821	0.179	10.148	<.001***
Trust_Block2	—	—	—	—	—	—	—	—	0.377	0.136	2.770	.006**
Stakes × Accuracy	-0.007	0.779	-0.008	.993	-0.492	1.027	-0.479	.632	-2.512	0.841	-2.986	.003**
Stakes × Relational Risk	0.367	0.774	0.475	.635	-0.025	1.020	-0.024	.980	-0.738	0.816	-0.904	.366
Stakes × Apology	—	—	—	—	—	—	—	—	-2.111	0.832	-2.537	.011*
Random Effects	Variance		Std. Dev		Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	2.882		1.698		7.396		2.720		2.296		1.515	
Item (Intercept)	0.000		0.000		0.166		0.407		0.000		0.000	

* $p < .05$, ** $p < .01$, *** $p < .001$. Em dashes indicate predictors not estimated in that block. A trailing dot denotes $p < .10$.

3.3 Chapter Reflection

This chapter examined risk communication as a textual intervention presented prior to the interaction with the AI system. While RQ1 is not the primary contribution of this paper, important insights emerge from viewing perception calibration strategies through a risk theoretical lens. Framing theory posits that the same probability can be expressed through the lens of loss or gain, which influences how people process and respond to probabilistic information [6, 127, 201]. A system that is correct 80% of the time can equally be described as 20% erring. Whereas AI *accuracy*, per convention, adopts a positive frame emphasising the 'gain' of interacting with an AI system, we designed the uncertainty component of the risk communication strategy using a negative frame ('loss'; operationalised through the AI's error rate). The findings of this chapter suggest that the frame valence is a contributing factor to the unwarranted inflation of trustworthiness perceptions observed under high decision-stakes when communicating AI accuracy. This finding emphasises the importance of integrating risk theory into the design of communicative strategies that aim to calibrate trust in AI systems, presenting frame valence as an impactful design parameter (RQ4). In terms of the design of the risk communication strategy itself, this chapter demonstrates that a negatively framed uncertainty component presented adjacent to an explicit consequence component represents a promising and cognitively lightweight intervention for a forcing function that encourages deliberation (RQ4).

Turning to RQ2, the study provides important first insights into how risk communication influences system trustworthiness trajectories across AI performances when AI systems assists users in forming a decision. We observe that risk communication serves as a transparency cue, grounded in a meaningful shift in how users conceptualise the system's limitations and capabilities, which manifested in greater scrutiny and deliberation in their assessments. The perceptual effects we observed in this chapter further provide empirical support for the TRAM model, which assumes that in the context of human-AI interaction, a risk-benefit assessment of the system occurs at the perceptual level, and not solely at the behavioural level as assumed in previous trust models. Critically, risk communication did not erode perceived utility, nor did it temper initial trustworthiness perceptions before an error had occurred, addressing the preceding concern that pre-emptive risk information may undermine the very benefits the AI system provides.

In reflection to RQ3, this chapter operationalises a conceptual distinction of relevance to human-AI interaction research: the difference between *situational risk*, which is context-induced, and *relational risk*, which pertains to the technology itself [189]. Our findings reveal that decision-stakes emerge as an important contextual moderating factor that determines the appropriateness of perception calibration strategies, and should therefore be considered when designing them or evaluating their suitability. This is most evident in the observed drawbacks of accuracy disclosure under high decision-stakes. Risk communication, by contrast, did not exhibit stake-dependence to the same degree. However, given the deliberation and scrutiny that risk communication evokes, it could be argued that it serves more appropriate in high stakes contexts, where vigilance and deliberation are warranted, and less appropriate in low-stakes settings, where the cognitive cost of deliberation may outweigh the benefits. The chapter therefore provides guidance for a more context-sensitive selection of perception calibration strategies.

In AI-assisted decision-making with classification systems, system errors and associated risks are commonly expressed as false positives and false negatives [30]. Reflecting the nature of this type of AI system, the *consequence* component in this chapter therefore adopted a generic system-level framing, communicating that AI use can have consequences in the form of task-specific false positives and false negatives, i.e., the failure to detect or the false detection of bacterial infestations. Beyond task-specific classifiers, however, the consequences of AI failures can vary considerably in their nature and scope and are not always confined to a discrete false positive or false negative. In LLM interactions, system failures can be more diffuse and are often difficult to quantify, which makes probabilistic uncertainty communication less applicable and positions consequence communication as a critical design consideration in user-facing risk communication. This raises

the question of *how* the consequence component should be designed to communicate AI risk across more general-purpose systems. The subsequent chapter addresses this question directly and turns to *consequence framing* in the context of LLMs.

Chapter 4

Consequence-Explicit AI Risk Communication: Understanding Framing Effects in LLM Interaction

4.1 Introduction

Article I established first insights into how risk communication shapes system trustworthiness perceptions across varying AI performances. Both components of the risk communication, the uncertainty and consequence component, were informed by the context of a binary classifier: (a) the uncertainty component adopted a negative frame, foregrounding the AI's error rate, a meaningful and quantifiable design lever for a system where performance can be expressed as a discrete probability. (b) The consequence component, similarly, adhered to a system-level framing, contextualised to the binary nature of the diagnostic task itself ("not detecting bacterial infestations or falsely detecting bacterial infestations"). In more agentic systems, however, uncertainty is rather an inherent and constant system characteristic rather than a quantifiable metric that can be meaningfully communicated to users.

Unlike classifier and recommender systems, LLMs are not confined to specific tasks but are general-purpose technologies woven into the fabric of everyday life, changing how people seek information and form decisions across a variety of tasks and domains. As a result, the consequences of AI risks can have different shapes and forms. LLMs may share confidential information, reproduce harmful language, or produce inaccurate information [217]. Imagine a hiring manager who drafts a job post with the support of an LLM. The system generates a harmful misrepresentation and, unaware of it, the manager believes and reproduces it. Notably, as similar cases accumulate across platforms, the consequences of AI risks posed by LLMs extend beyond the individual interaction, and, in the exemplified case of harmful misrepresentation, may reinforce stereotypes and cause harm at a societal scale. AI disclosures mandated under Article 50 of the EU AI Act [52] represent an important step toward promoting transparency in human-AI interaction. Yet, current implementations remain largely consequence-agnostic, merely signalling AI use or offering generic cautions. From a risk theory perspective, such disclosures leave the consequence component of risk communication underspecified, without supporting users to anticipate how risks may manifest in system outputs, users' interactions with those outputs, or society more broadly.

In this chapter, we therefore turn to the consequence component as an important and less examined mechanism for user-facing risk communication during LLM interactions. In a between-within online user study with $N = 300$ participants, we investigate how the lens through which consequences are communicated in AI risk disclaimers, the *Consequence Framing*, influences the user's risk perceptions when communicated in situ,

manipulating whether the consequence is framed through a system level lens, emphasising the corresponding user experience, or through a systemic lens. To exemplify, in the scenario in which the user requested the AI to create a job post (hiring), the AI risk disclaimer presented the associated risk of discrimination. In the system level framing condition, this consequence was communicated through the fact that the system may produce discriminatory language; in the human-interaction condition, the consequence was communicated through the associated human experience, i.e. the user sharing discriminatory language; and in the systemic condition, the consequence was communicated through a systemic lens, i.e. the AI may reinforce discrimination across society. We included a between-subjects control condition in which no AI risk disclaimer was displayed. This allowed us to establish whether the mere presence of a disclaimer is sufficient to shift system trustworthiness and users' risk perceptions.

Importantly, the communication of specific consequences and the lenses through which they are framed may not resonate equally across audiences. Risk theory identifies political orientation as an important individual difference in how risks are perceived and processed [39, 86]. Recent work further emphasises the heightened politicisation of AI and shows that political orientation functions as a lens through which people increasingly appraise AI systems [224]. This chapter therefore investigates the interaction between *Consequence Framing* and *political orientation* as an individual difference factor.

Our findings reveal that the mere presence of a consequence-explicit AI risk disclaimer is sufficient to increase AI risk perceptions and to decrease system trustworthiness. Our results demonstrate that the lens through which the consequence is presented matters, though to varying degrees, revealing *Consequence Framing* as a mechanism that significantly influences risk perceptions, but not system trustworthiness. Examining the framing conditions more closely, human-interaction framing appears to be an effective mechanism, as foregrounding consequences through a human-centred perspective elicits negative affect, such as worry, and increases the perceived severity of those consequences. The human-centred nature further caused the AI risk to feel personally relevant and positioned the user as responsible for both the outcomes and their mitigation, in that *the way a user is using the system determines its risks*, casting AI risk as a "people issue and not an AI issue". The very mechanisms that cause the human-interaction framing to be effective, however, exposed the limitations of systemically framed AI risks. We observe that systemically framed AI risks fall short in reducing system trustworthiness or amplifying risk perceptions. We unveil psychological distance as one underlying mechanism, which caused the communicated AI risk to feel distant, less imminent, and personally irrelevant. Consequently, its effectiveness appeared to be confined solely to community-minded audiences, that is, those who are inherently concerned with the collective impact of AI on society. Political orientation emerged as a significant moderator, with human-interaction framing being less politically contingent than capability framing and thus better suited as a communication default across audiences.

The full article provides a comprehensive account of the empirical findings, together with an in-depth discussion of their implications and associated ethical considerations. We note that in the subsequent article, the main manipulation of this chapter is referred to as *Risk Framing*, and the system-level framing as *capability* condition.

4.2 Article II

This article is currently under review. Copyright is held by the authors. This is the authors' version of the work. It is posted here for your personal use. Not for redistribution.

Sarah Schömbs, Jarod Govers, Jorge Goncalves, and Wafa Johal. 2026. System, Human, or Society? How Risk Framing and Political Orientation Shape AI Risk Disclaimer Perceptions in LLM Applications. *Submitted for publication to The AAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES), May 2026.*

Ethics Application ID: 33699, the University of Melbourne Human Research Ethics Committee.

System, Human, or Society? How Risk Framing and Political Orientation Shape AI Risk Disclaimer Perceptions in LLM Applications

Anonymous submission

Abstract

Current AI disclaimers predominantly disclose AI presence or system capabilities, often overlooking how AI risks manifest in practice and the political lens through which AI is increasingly perceived. We conducted a between-within mixed-method user study ($N = 300$) to investigate how framing consequence-explicit AI risk disclaimers at three levels of risk consequence (capability, human-interaction, systemic) affects perceived AI risk and system trustworthiness, and how political orientation moderates these effects. Our work shows that consequence-explicit AI risk disclaimers function as an effective risk signal relative to their absence and demonstrate that risk framing matters, with political orientation as a significant moderator. Our findings reveal that AI risk disclaimers emphasising the user's experience (human-interaction) significantly increase the perceived severity of AI consequence and affective concern compared to a systemic framing, and are less politically contingent than AI capability-centred framings. We provide empirical insights for consequence-explicit AI risk disclaimer design in user-facing LLM applications, ethical considerations, and direct implications for AI transparency policy.

Introduction

Artificial intelligence has become one of the most politically charged technologies of our time, exemplified by debates around democratic epistemic agency (Coeckelbergh 2023), or the US Pentagon's attempt to blacklist Anthropic over perceived ideological misalignment (O'Donnell 2026). At the same time, LLMs have catalysed the widespread deployment of AI systems across society and made AI-based applications substantially more accessible to end-users (Hoffmann and Frase 2023).While LLMs bring clear benefits to individuals and society, they also introduce significant risks in the form of e.g., misinformation, privacy, and the reproduction of harmful language and misrepresentation (Weidinger et al. 2022).

Crucially, real-world evidence from the AI Incidents Database (McGregor 2021) shows that most documented AI harms do not arise at the development stage but emerge *during* operational use, thus *directly affecting end-users* (De Miguel Velázquez et al. 2024). This raises a fundamental question about ***how AI risks should be communicated to users of LLM systems amid a politically charged environment.***

The EU AI Act's Article 50 (European Parliament and Council 2024b) represents an important first regulatory step on user-facing transparency obligation, but these obligations focus primarily on informing end-users about the *presence* of an AI system or AI-generated output, which has shown to decrease trust in the person that used the AI (Brennen et al. 2024; Schilke and Reimann 2025). The EU AI Act obligations fall short of addressing the communication of the *relational risk* associated with the AI system itself (Stuck, Tomlinson, and Walker 2022) and its potential consequences to end-users. This AI consequence agnosticism is also somewhat reflected in existing user-facing LLM applications, where well-established systems such as ChatGPT (OpenAI 2026) display static disclaimers such as "ChatGPT can make mistakes. Check important info.". These statements express AI failabilility at a system level, but communicate little about the concrete manifestation of consequences.

As AI risks extend beyond the system level to individual users (Weidinger et al. 2022) and society at scale (Uuk et al. 2024), AI developers face a practical design question: should AI risk consequences be framed in terms of what the AI system might do wrong, what could go wrong in the human-AI interaction, or what broader societal consequences might follow? These three dimensions of AI risk consequence (system capabilities, human-interaction, systemic) reflect distinct risk frames that map onto established AI risk taxonomies and safety evaluation frameworks (Weidinger et al. 2022; Slattery et al. 2025; Uuk et al. 2024). However, despite the practical urgency of this design question and its direct relevance for informing AI regulation and disclosure policy, ***how the framing of AI risk consequences affects end-users' risk perceptions*** remains empirically unexamined in the context of LLM applications.

Risk theory demonstrates that framing effects shape risk perception, for instance through loss versus gain framing (Gong et al. 2013) or self-referent messaging that makes risks feel personally relevant (Dunlop, Wakefield, and Kashima 2008), and that political orientation is associated with differential sensitivity to personal versus collective hazards (Choma et al. 2013; Jost, Nosek, and Gosling 2008). Beyond risk sensitivity, political identity increasingly shapes AI acceptance and AI perceptions more broadly (Claudy, Parkinson, and Aquino 2024), with people relying on it as a mental shortcut to evaluate AI systems, also described as

the "nascent politicisation of AI" (Yang et al. 2025). Political orientation therefore represents a theoretically motivated lens through which to investigate how AI risk disclaimer framing is perceived by end-users.

Our work directly contributes to the aforementioned research gaps by systematically addressing the following research questions:

RQ1 How does the presence of a consequence-explicit AI risk disclaimer influence perceived AI risk relative to its absence?

RQ2 To what extent does the *Risk Framing* of AI risk disclaimers (Capability, Human-Interaction, Systemic - lens) influence perceived AI risk?

RQ3 How does people's political orientation moderate the effect of *Risk Framing* (Capability, Human-Interaction, Systemic - lens) on perceptions of AI risk?

We conducted a mixed-methods user study ($N = 300$) with a between-within experimental design, in which participants evaluated different AI systems by observing simulated LLM scenarios that deliberately contextualised typical interaction contexts in which AI risks can manifest. Each AI system was accompanied by a consequence-explicit AI risk disclaimer, operationalised through *Risk Framing* (Capability, Human-Interaction, Systemic), grounded in the three-layered AI safety evaluation framework (Weidinger et al. 2023). A control group (no AI risk disclaimer) served to assess whether AI risk disclaimers influence people's perceived AI risk relative to their absence (RQ1); as a preliminary investigation distinct from the framing comparisons examined in RQ2, and their interaction with peoples' political orientation (RQ3). Participants also provided answers to open-ended questions which explored the mechanisms behind the perceptual effects of the specific *Risk Framing* within each AI risk disclaimers.

Our results show that consequence-explicit AI risk disclaimers increase risk perceptions relative to their absence. Human-interaction framing significantly increased perceived severity of AI consequences and negative affective responses (worry, concern, and fear) toward the AI system compared to a systemic framing, showed less political contingency than capability-framed AI risk disclaimers, and actively cued perceptions of human oversight responsibility. We identify mechanisms underlying the limited effectiveness of systemically framed AI risks, and discuss benefits and drawbacks of different AI risk framing approaches. Based on our findings, we advocate for human-centred framing as a communicative default for standardised AI risk disclaimer design across politically diverse populations. We further provide implications for 'user-aware' adaptive AI risk disclaimers, and discuss their ethical considerations. In doing so, we contribute to end-user facing in-situ AI risk communication as an important and underexplored facet of responsible AI.

Related Work

Relevant AI Risk Taxonomies and Frameworks

We adopt the definition of AI risk by Slattery et al. (2025), based on the Society for Risk Analysis Glossary (Aven et al. 2018), and describe it as "the possibility of an unfortunate occurrence associated with the development or deployment of artificial intelligence". While our work focuses on the risk consequences, we acknowledge that uncertainty is a central communicative challenge to AI risk communication (Breakwell 2007; Schömbs et al. 2024); encoded as 'a possibility' in the aforementioned definition. When a negative AI risk consequence manifests, this is termed 'AI harm', which defines "an entity [experiencing] a harm event or harm issue that can be directly linked to a consequence of the behaviour of an AI system" (Hoffmann and Frase 2023).

A significant body of work has focused on identifying and classifying AI risks, often with the objective of informing policy and regulation (Uuk et al. 2024; Slattery et al. 2025; Weidinger et al. 2022). These taxonomies address different aspects of AI risks, for instance the domain assigned to the AI risk (Slattery et al. 2025; Weidinger et al. 2022), their causal attribution (Slattery et al. 2025; Hendrycks, Mazeika, and Woodside 2023), or their sociotechnical dimension (Uuk et al. 2024). The three-layered framework proposed by Weidinger et al. (2023) distinguishes AI risks based on where they emerge within the sociotechnical system. The first layer describes AI risks that arise from model *capabilities* and their limitations, e.g. the AI sharing sensitive information. The second layer describes AI risks that manifest in the *human experience when interacting* with the AI system, e.g. induced false beliefs. The third layer captures risks that emerge at a *systemic level*, e.g. the reinforcement of social inequalities. The framework was originally developed to support sociotechnical evaluation of generative AI systems. In our work, we use this framework to identify communicative lenses for AI risk communication. We adopt it as a structured way to investigate how respectively framed AI risk disclaimers (capability, human-interaction, systemic) shape end users' risk perception of AI systems.

Political Orientation and (AI) Risk Perception

Risk theory and communication research consistently emphasise the importance of accounting for individual differences to make risk communication more effective (Peters, Hart, and Fraenkel 2011; Morss et al. 2018; Gierlach, Belsher, and Beutler 2010; Breakwell 2000; Slovic 1987). This is echoed in recent work on the operationalisation of AI transparency obligations under Article 50 EU AI act, which surfaced audience adaptation ("*different people may require different types of communication mechanisms*") and individual information relevance ("*What (under-) information is meaningful for the recipient?*") as important challenges for the design of effective AI disclosures (El Ali et al. 2024).

How political orientation shapes risk perceptions Political orientation represents one individual difference of particular relevance. Right-wing authoritarianism, which is strongly associated with the extreme end of the political conservative spectrum (Choma et al. 2013), is associated with heightened sensitivity to personal and everyday hazards, consistent with conservatism's underlying orientation toward needs for order, stability, and uncertainty reduction (Jost, Nosek, and Gosling 2008). Political liberalism, on the

other hand, is associated with a more pronounced risk perception for collectively shared hazards, such as guns or climate change (Choma et al. 2013). The locus of these differences is often attributed to the distinct moral foundations that guide liberals' and conservatives' intuitions and judgements (Graham, Haidt, and Nosek 2009; Atari et al. 2023).

Political Orientation in the context of AI In the context of AI, the picture is less clear. Some work shows that conservatism predicts lower acceptance of technological innovations such as brain-computer interfaces or facial recognition, mediated by moral foundations centred around loyalty, authority, and purity (Claudy, Parkinson, and Aquino 2024; Graham, Nosek, and Haidt 2012), and that conservatives are generally less supportive of AI than liberals (Yang et al. 2025), as well as placing less trust in scientists to responsibly develop AI and more trust in technology companies (Yang et al. 2025). Other findings, however, point in the opposite direction, with conservatives showing stronger acceptance of AI-generated recommendations compared to liberal-leaning individuals, albeit attributed to their preference for consistency rather than reduced risk sensitivity (Paul et al. 2025). A report by the National Centre for Social Research UK similarly shows that people with conservative orientations feel that the benefits of AI exceed their concerns, with decreased concerns related to AI inaccuracy (Sippy et al. 2025).

Aligning the presentation of AI with the values that underlie political orientations further increases AI acceptance, such that framing AI use around values more central to conservatives increased acceptance among conservative-leaning individuals, while a framing around how AI contributes to equality, diversity, and inclusion increased acceptance among liberal-leaning individuals (Claudy, Parkinson, and Aquino 2024). This reflects a principle coined 'moral reframing' (Feinberg and Willer 2019), where messages are more persuasive to recipients when they are framed according to the moral orientations of those recipients.

These findings suggest that political orientation may shape how individuals respond to differently framed AI risk disclaimers, which we explore in this work.

AI Disclosures

AI risk communication is a relatively nascent area, with efforts spanning the development and deployment pipeline. Article 13 of the EU AI Act (European Parliament and Council 2024a) specifies deployer-centric requirements for high-risk AI systems, including the provision of information about the system's "capabilities and limitations of performance", and disclosure of "reasonably foreseeable misuse". This is complemented by research tools designed to anticipate AI harms during the deployment stage (Rao et al. 2025; Buçinca et al. 2023; Wang et al. 2024), such as Farsight, a graphical interface that helps developers envision AI harm during early AI prototyping (Wang et al. 2024). Efforts directed at end-users include designs which presents model characteristics in an energy-label format (Open Ethics Initiative 2025), model properties disclosed in an ingredient-list format (Twilio 2025), or AI risk report cards (Bogucka et al. 2025), though these formats remain innately more technical than textual AI disclosures. The latter of which are defined in Article 50 (European Parliament and Council 2024b), in that users must be "informed that they are interacting with an AI system" (Article 50(1)), and synthetic AI outputs are required to be "marked" and "detectable" (Article 50(2)).

Existing AI disclosure research focuses predominantly on consequence-agnostic labels that inform users of AI presence/use (Toff and Simon 2025; Brennen et al. 2024; Schilke and Reimann 2025), with little empirical insights on how the framing of AI risk consequences in such short text-based interventions shapes risk perceptions of the AI system itself, or how political orientation moderates these effects. Given that AI disclosure is increasingly mandated and AI risks have real-world consequences for individuals and society at scale, it is important to investigate what types of consequence-explicit risk communication are most effective and for whom.

Method

To answer RQ1-3, we conducted a mixed-methods between-within user study with a between-subjects control group (no AI risk disclaimer) and *Risk Framing* (Capability, Human-Interaction, Systemic) as the within-subjects independent variable, operationalised across three grounded LLM interaction scenarios in which typical AI risks manifest (Weidinger et al. 2022). Given the exploratory nature of this work and the lack of empirical work on AI risk disclaimer framing, we refrained from formulating directional hypotheses.

Task Stimuli and Rationale

Participants were instructed to observe simulated interactions between a user and an LLM across three scenarios, each presented as a distinct AI system (System A, B, and C) through a minimal chat-based UI, and to imagine themselves in the role of the interacting user. We deliberately presented each scenario as a separate AI system to avoid carry-over effects of the displayed AI risk disclaimer on the repeated-measures. To ensure ecological validity, the UI resembled widely deployed conversational AI systems (see Figure 1). The AI was simulated through a hard-coded conversation, which allowed precise control over system outputs to reduced variability across participants. Participants could not type or interact directly with the interface. The mock-ups were implemented in custom HTML, CSS, and JavaScript to mirror real-world system behaviour rather than present static depictions. We iteratively refined the stimuli design (UI and conversational content) through a series of pilots ($N = 17$).

Task Stimuli - Conversation Scenarios We selected three conversation scenarios in which AI risks can manifest: health misinformation, privacy risk during flight booking assistance, and hiring discrimination. Scenario selection was informed by an analysis of the MIT AI Incident Tracker (Mylius et al. 2025), which builds on the AI Incident Database (McGregor 2021), in which we examined incidents from 2024 to 2025; post-deployment occurrence; unintentional rather than deliberately caused outcomes; with AI as the primary causal entity. The selected categories emerged

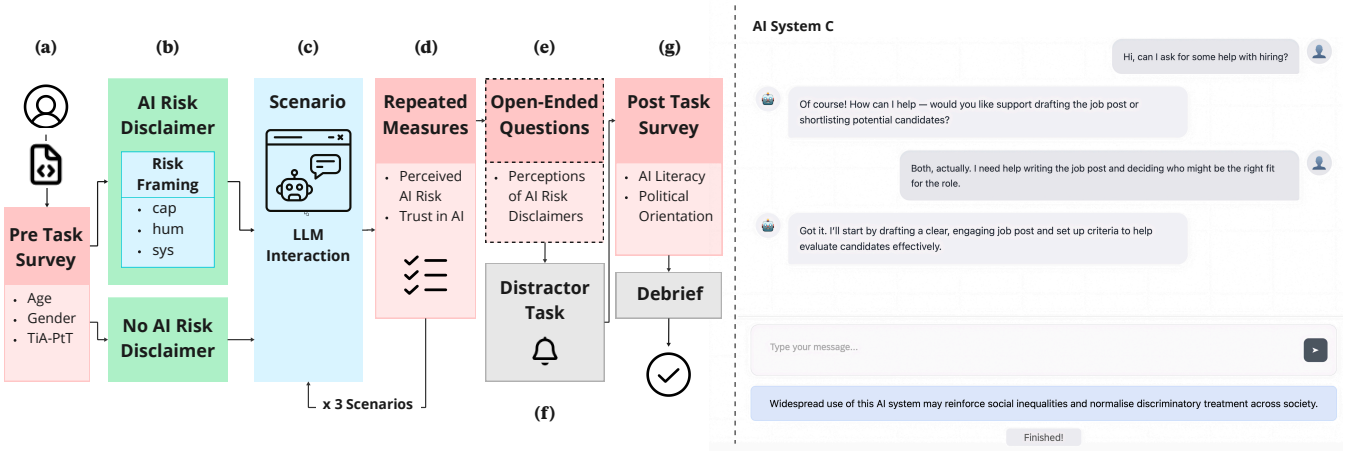


Figure 1: Left: experimental flow. (a) Pre-task survey: demographics, TiA-PtT. (b-c) Participants assigned to either the disclaimer condition, with *Risk Framing* conditions counterbalanced across three simulated LLM scenarios via Latin square design, or the no-disclaimer control condition. (d) Repeated measures after each scenario: overall risk, severity, affect, trust, distrust. (e) Open-ended questions (disclaimer condition only). (f) Distractor task. (g) Post-task survey: AI literacy, political orientation. Right: exemplary UI of a simulated LLM interaction with a systemic-framed AI risk disclaimer.

among the most frequently reported. Each simulated interaction followed a fixed structure in which the user prompted a support request, the AI responded affirmatively and invited further detail, and the user introduced a more concrete request, with the AI signalling readiness to act without executing any decision, to reduce confounds such as perceived AI autonomy.

Task Stimuli - AI Risk Disclaimer We displayed the AI disclaimer as a short textual intervention message below the user prompt field in the chatbot interface, exemplified in Figure 1, which follows existing deployed systems (e.g., ChatGPT, Claude AI) and adheres to regulations for AI disclosures (Article 50 (5) in the EU AI Act (European Parliament and Council 2024b)), which requires AI disclosure information to be provided clearly and distinguishably by the time of first interaction. Participants assigned to the control condition were exposed to the conversation scenarios without any AI risk disclaimers.

Operationalisation of *Risk Framing* Manipulation

The *Risk Framing* manipulation was adapted from the three layered framework proposed by Weidinger et al. (2023). We constructed the *Risk Framing* as follows, see the Appendix for the complete set of framings:

Capability framing describes potential consequences framed at the level of the AI system, introduced by the phrase "This AI system may + consequence".

Human-interaction framing describes potential consequences framed at the level of user interaction, initiated by the phrase "People who interact with this AI system may + consequence".

Systemic framing described potential consequences framed at the societal level, introduced by the phrase "Widespread use of this AI system may + consequence", with the disclaimer referring consistently to impacts "across society".

We used the term 'may' across each *Risk Framing* to reflect the uncertain nature of risks, consistent with standard definitions of risk (Slattery et al. 2025). Within each *Risk Framing* condition, we controlled message length as closely as possible across the varying interaction scenarios.

The consequence-component in each AI risk disclaimer was deliberately phrased to be 'scenario-aware', i.e. the communicated consequences was aligned to the specific interaction context. For instance, in the hiring scenario, disclaimers warned about consequences related to discriminatory language, while the flight booking scenario foregrounded privacy-related consequences. This ensured that participants encountered risk information that was directly relevant to the AI interaction at hand. The scenario was included as a random intercept in all models to account for potential variability in responses across interaction contexts.

Measures

Dependent Measures - AI Risk Perceptions and System Trustworthiness We measure AI risk perception using items from established risk perception research (Walpole and Wilson 2021) and adapted its wording to reference AI systems, following HCI research (Schömbs, Goncalves, and Johal 2025; Schwesig et al. 2023). We capture *general AI risk perception* via a unidimensional item (Wilson, Zwickle, and Walpole 2019), asking "How risky is the AI system?" following. We assess both *perceived severity* (e.g. "How severe would you expect the consequences of the AI system to be?") and *elicited affect* (e.g. "When you think about the AI system, to what extent do you feel worried?") (Walpole and Wilson 2021) through validated subscales on a 5-point scale as important perceptual dimension of risk consequences (Wilson, Zwickle, and Walpole 2019).

We also measure perceived system trustworthiness, as it is theoretically distinct yet an empirically intertwined determinant of how users respond to AI systems in contexts char-

acterised by uncertainty and potential harm (Stuck, Tomlinson, and Walker 2022; Visschers and Siegrist 2008). We measure perceptions of system trustworthiness using a reduced version (Perrig, Scharowski, and Brühlmann 2023; Scharowski et al. 2025) of the Trust between People and Automation (TPA) scale (Jian, Bisantz, and Drury 2000). The scale consists of two perceptual factors, trust and distrust (Perrig, Scharowski, and Brühlmann 2023; Scharowski et al. 2025), measured on a 7-point scale. The wording of each item was adapted to the human-AI interaction context; e.g. "The AI system is dependable" for trust, and "The AI system is deceptive" for distrust (Scharowski et al. 2025).

Political Orientation - PSP We measure Political Self-Placement (PSP) to assess people's political orientation using three items capturing generic, social, and economic political orientations on a 7-point scale (1 = *very liberal* to 7 = *very conservative*), following Luke and Gawronski (2021).

Covariates - TiA-PtT and AI Literacy To account for individual differences relevant to AI risk perception and trustworthiness judgements, we include propensity to trust technology and AI literacy as covariates. We assess people's dispositional trust using the Propensity to Trust automation (TiA-PtT) subscale (Körber 2019), which includes three items measured on a 5-point scale, grounded in trust theory (Mayer, Davis, and Schoorman 1995). We measure AI literacy using the Understand AI subscale from the Meta AI Literacy Scale (MAILS) (Carolus et al. 2023), ranging on a scale from 0 to 10. The subscale assesses the conceptual understanding of AI, system limitations, and the ability to reflect on (dis)advantages of AI use, as differences may influence how risk disclosures are perceived and integrated into judgments about AI systems.

Participants

We determined our sample size via an a priori power analysis using the R package `InteractionPoweR` (Baranger et al. 2023) with an α = .05 and a power of .80. The analysis indicated a minimum sample size of N = 225. We hosted the online survey using the crowdsourcing platform Prolific. We implemented Prolific's internal pre-screening to ensure that the participants were US based, proficient in English, aged $\geq$ 18, and to check that participants obtained a prior approval rate $\geq$ 98%. We used balanced quotas for self-identified political alignment (liberal, moderate, conservative) to ensure a diverse participant pool. We compensated participants at £9.00/hour following the platform's payment standards. We excluded participants who failed the AI risk disclaimer recognition check after each scenario stimuli, and participants who failed the attention check item. After the exclusion of 5 cases that failed the aforementioned criteria, we recruited a final set of $N = 300$ participants (self-identified as $145F$, $153M$, $2Non-binary$). The sample had an average age of 43.62 (SD = 13.36), ranging from 18 to 76 years old. Our work was approved by the University's Human Ethics Committee.

Procedure

We used a Latin square design to counterbalance the presentation order of scenarios and the assignment of *Risk Framing* conditions across scenarios. The first phase, see Table 1, included a CAPTCHA verification to prevent automated bot responses, a JS browser test to ensure that participants were able to view the task stimuli, a plain language statement that informed participants about the research itself, and a consent form. After providing consent, participants were guided to answer the pre-task survey, which involved questions related to participants' demographics and TiA-PtT (Körber 2019).

In the the main experimental phase, participants were instructed to view three interaction examples with different new AI systems (A, B, and C). For each AI system, participants were prompted to click the 'Start' button positioned at the bottom of the UI, read the information provided, and observe the interaction between the user and the AI system. Each UI displayed the simulated interaction alongside its corresponding AI risk disclaimer, with the *Risk Framing* condition varying across the three scenarios. For participants assigned to the control group, the UI did not display any AI risk disclaimers.

After each scenario, we implemented a short recognition check asking participants to identify the correct AI disclaimer message, presented as a simple multiple choice question. We further asked participants to answer the repeated-measures questionnaire examining participants perceived AI risk (affect and severity) (Walpole and Wilson 2021) and perceived system trustworthiness (TPA) (Scharowski et al. 2025). Subsequent to all scenarios, participants who received AI risk disclaimers responded to two open-ended questions asking them to assess which of the AI disclaimers made the AI system seem most/least risky or trustworthy and what specific information led to those assessments, and to reflect on which consequences mattered most to them personally when evaluating the AI disclaimers and why.

In the final phase, we implemented a Stroop task (six rounds) as a cognitive reset to minimise unintended biases between the main task, the repeated-measures, and the post-task assessment of political orientation (Luke and Gawronski 2021) and AI literacy (Carolus et al. 2023). We displayed a short debriefing to disclose the fictional nature of the observed AI systems.

Results

Quantitative Analysis

We first conducted a preliminary analysis comparing participants who viewed AI risk disclaimers to those in the control group interacting with the AI systems without any AI risk disclaimers to examine the effect of disclaimer presence on perceived AI risk and perceived system trustworthiness (RQ1). Narrowing the lens to participants who received a disclaimer, we then investigated the effect of *Risk Framing* on these outcomes (RQ2), before exploring the moderating role of political orientation (RQ3).

Model Construction We constructed separate models for each dependent variable. We employed Cumulative Link

Mixed Models (CLMM) to analyse participants risk perception (*general risk of AI system*, *severity of consequence*, *affect*) using the statistical R package `ordinal` [1] after inspecting the model assumptions, which indicated an ordinal structured data. Trust and distrust, as continuous composite scores (Jian, Bisantz, and Drury 2000; Scharowski et al. 2025), were analysed using Linear Mixed Effects Models (LMM), using the statistical R package `lme4` [2].

All models included propensity to trust automation (*TiA-PtT*) and *AI literacy* as fixed effects to control for their influence on the dependent measures. To account for the task scenarios contextualising the AI interaction, *Scenario* was included as a random intercept in all models. *Participant ID* was further included as a random intercept to account for the repeated measures design. For CLMMs, *Item ID* was additionally included as a random intercept to account for item-level variability in ordinal response data, with the exception of *general AI risk perception* as it was measured using a single uni-dimensional item.

For RQ1, we constructed models for the full sample ($N = 300$), with *Disclaimer* included as a binary fixed effect (1 = present, 0 = absent). For RQ2 and RQ3, we constructed models by including a *Risk Framing* $\times$ Political Orientation interaction term to determine the main effect of Risk Framing (RQ2) and to test political orientation as a moderator of disclaimer framing effects on perceived AI risk and perceived system trustworthiness (RQ3). To ensure the main effect of *Risk Framing* was evaluated at a meaningful and observed value of political orientation, political orientation was mean-centred prior to our modelling ($M = 3.92$, $SD = 1.78$). The general model structure for each RQ is presented below, where $[\cdot]$ denotes terms included in CLMM models only:

$$\begin{aligned}\text{RQ1: DV} \sim{}& \text{Disclaimer} + \text{TiA-PtT} + \text{AI Literacy} \\ &+ (1|\text{PID}) + (1|\text{Scenario}) \ [+(1|\text{Item})]\end{aligned} \tag{1}$$

$$\begin{aligned}\text{RQ2/3: DV} \sim{}& \text{Risk Framing} \times \text{Political Orientation} \\ &+ \text{TiA-PtT} + \text{AI Literacy} \\ &+ (1|\text{PID}) + (1|\text{Scenario}) \ [+(1|\text{Item})]\end{aligned} \tag{2}$$

Prior to the analysis, we assessed multicollinearity for all LMM models using the Generalised Variance Inflation Factor (GVIF). All covariates returned $GVIF^{1/(2Df)}$ values between 1.00 and 1.22, indicating no concern for multicollinearity. As GVIF is not applicable to CLMM models, predictor multicollinearity was additionally assessed via Pearson correlations amongst covariates. The correlation between TiA-PtT and political orientation was significant but negligible ($r = .09$, $p = .019$), which confirmed sufficient independence for inclusion in the same model. Internal consistency (Cronbach's α) was excellent across all dependent measures (severity ($\alpha = 0.89$), affect ($\alpha = 0.90$), trust ($\alpha = 0.96$), distrust ($\alpha = 0.95$)).

Quantitative Results

Effect sizes are reported as standardised coefficients (β_{std}) with 95% confidence intervals for LMM models, and as Odds Ratios (OR) for CLMM models.

RQ1 - The Perceptual Effects of Disclaimer Presence

***Disclaimer Presence* - AI Risk Perceptions** We observed that the presence of an AI risk disclaimer significantly increased general risk perceptions ($\beta = 1.30$, $SE = 0.25$, $p < .001$, $OR = 3.67$), led to a significantly higher perceived severity of AI consequences ($\beta = 1.18$, $SE = 0.26$, $p < .001$, $OR = 3.25$), and elicited greater affective concern toward the AI system ($\beta = 1.61$, $SE = 0.26$, $p < .001$, $OR = 4.99$); see Figure 2 (left). In other words, the mere presence of a risk disclaimer was sufficient to heighten risk perceptions across all measured risk constructs.

Regarding individual differences, we found that dispositional trust in automation (TiA-PtT) was associated with significantly lower overall risk perception ($\beta = -1.12$, $SE = 0.17$, $p < .001$, $OR = 3.06$), lower perceived severity ($\beta = -1.04$, $SE = 0.18$, $p < .001$, $OR = 2.84$), and lower affective responses ($\beta = -1.12$, $SE = 0.19$, $p < .001$, $OR = 3.07$), showing that individuals with higher dispositional trust in automated systems tend to perceive AI risks as less severe and respond with less affective concern. AI literacy did not significantly predict any of the measured risk constructs (all $p > .05$).

***Disclaimer Presence* - System Trustworthiness** We further investigated whether the presence of an AI risk disclaimer influenced how trustworthy participants perceived the AI system to be, examining both trust and distrust as distinct dimensions of system trustworthiness (Perrig, Scharowski, and Brühlmann 2023; Scharowski et al. 2025). We observed that presenting an AI system accompanied with an AI risk disclaimer significantly decreased trust perceptions in the AI system ($\beta = -0.65$, $SE = 0.15$, $p < .001$, $\beta_{std} = -0.18$) and significantly increased perceptions of distrust ($\beta = 1.13$, $SE = 0.17$, $p < .001$, $\beta_{std} = 0.28$), see Figure 2 (right).

With respect to the measured covariates, TiA-PtT significantly predicted higher trust ($\beta = 0.89$, $SE = 0.10$, $p < .001$, $\beta_{std} = 0.36$) and lower distrust ($\beta = -0.59$, $SE = 0.12$, $p < .001$, $\beta_{std} = -0.22$). AI literacy significantly predicted higher trust ($\beta = 0.16$, $SE = 0.03$, $p < .001$, $\beta_{std} = 0.22$) but had no significant effect on distrust ($p > .05$).

RQ2 - The Perceptual Effects of *Risk Framing* For models examining the effect of *Risk Framing* (RQ2, RQ3), we conducted post-hoc pairwise comparisons using Bonferroni correction to assess all pairwise differences across framing conditions. For brevity, capability framing, human-interaction framing, and systemic framing may hereafter be referred to as cap, hum, and sys, respectively.

***Risk Framing* - AI Risk Perceptions** We did not observe a significant effect of *Risk Framing* on participants' general risk perception (all $p > .05$).

However, we found a significant effect of *Risk Framing* on the perceived severity of AI consequences. Participants perceived the consequences of an AI system as significantly more severe when the AI risk was disclosed in a capability

[1] https://cran.r-project.org/web/packages/ordinal/index.html
[2] https://cran.r-project.org/web/packages/lme4/index.html

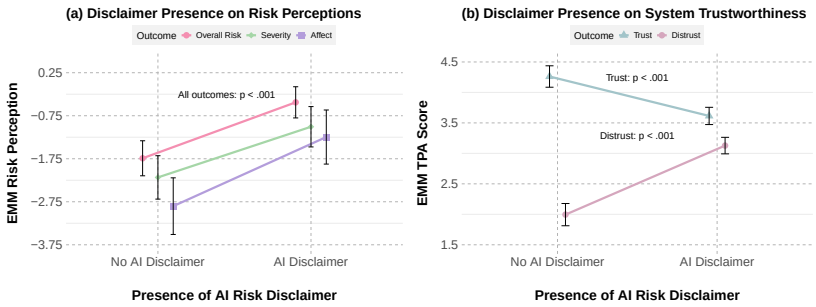


Figure 2: The impact of AI Risk Disclaimer presence on (a) risk perceptions and (b) perceived system trustworthiness. Estimated marginal means (EMMs) of measured constructs. Error bars denote standard error (SE).

framing ($\beta = 0.30$, $SE = 0.11$, $p = .005$, $OR = 1.35$) or human-interaction framing ($\beta = 0.30$, $SE = 0.11$, $p = .005$, $OR = 1.35$) compared to systemic framing. Post-hoc Bonferroni comparisons confirmed that capability framing ($EMM = -0.93$, $SE = 0.49$) elicited significantly higher perceived severity than systemic framing ($EMM = -1.24$, $SE = 0.49$; $p = .015$, $OR = 1.35$), and human-interaction framing ($EMM = -0.93$, $SE = 0.49$) similarly led to higher perceived severity than systemic framing ($EMM = -1.24$, $SE = 0.49$; $p = .014$, $OR = 1.35$); see (left) Figure 3. Capability and human-interaction framing did not significantly differ from each other ($p > .05$).

We further observed a significant effect of *Risk Framing* on participants' affective responses to the AI system. Specifically, a human-interaction framing elicited significantly higher affective responses than a systemic framing ($\beta = 0.33$, $SE = 0.11$, $p = .002$, $OR = 1.39$), confirmed by post-hoc Bonferroni comparisons (hum: $EMM = -1.13$, $SE = 0.65$ vs sys: $EMM = -1.46$, $SE = 0.65$; $p = .007$, $OR = 1.39$); see (centre) Figure 3. Capability framing did not significantly differ from systemic or human-interaction framing (all $p > .05$).

Consistent with the pattern observed in RQ1, we found that TiA-PtT was associated with significantly lower general risk perception ($\beta = -1.08$, $SE = 0.21$, $p < .001$, $OR = 2.94$), lower perceived severity ($\beta = -1.13$, $SE = 0.23$, $p < .001$, $OR = 3.09$), and lower affective responses ($\beta = -1.09$, $SE = 0.23$, $p < .001$, $OR = 2.98$). AI literacy did not significantly predict any of the measured risk constructs (all $p > .05$).

Beyond the framing effects, we observed that political orientation significantly predicted general risk perception ($\beta = -0.29$, $SE = 0.09$, $p = .002$, $OR = 1.33$) and affective responses ($\beta = -0.20$, $SE = 0.09$, $p = .027$, $OR = 1.22$), but not perceived severity ($p > .05$). In other words, more conservative participants tended to perceive the AI system as less risky overall and reported lower affective concern, independent of the specific framing condition they encountered.

***Risk Framing* - System Trustworthiness** We did not observe any significant effects of *Risk Framing* on either trust or distrust in the AI system (all $p > .05$).

Consistent with the pattern observed in RQ1, we found that TiA-PtT was associated with significantly higher trust ($\beta = 0.89$, $SE = 0.12$, $p < .001$, $\beta_{std} = 0.35$) and lower distrust ($\beta = -0.62$, $SE = 0.15$, $p < .001$, $\beta_{std} = -0.22$). We further observed that AI literacy significantly predicted higher trust ($\beta = 0.16$, $SE = 0.04$, $p < .001$, $\beta_{std} = 0.19$), but not distrust ($p > .05$). In practical terms, participants with greater conceptual understanding of AI, including awareness of its limitations and the ability to critically reflect on its use, tended to report higher trust in the AI system.

We further found that political orientation significantly predicted trust perceptions ($\beta = 0.12$, $SE = 0.05$, $p = .022$, $\beta_{std} = 0.14$), but not distrust ($p > .05$). More conservative participants reported higher trust in the AI system, regardless of the risk message framing they encountered.

RQ3 - The Moderating Role of Political Orientation on *Risk Framing*

***Political Orientation* - AI Risk Perceptions** We did not observe a significant moderating effect of political orientation on participants' general AI risk perception (all $p > .05$).

Imperatively, we observed a significant interaction effect between capability framing and political orientation on the perceived severity of AI consequences, relative to systemic framing (cap vs sys × Political_mean: $\beta = -0.13$, $SE = 0.06$, $p = .030$, $OR = 1.14$). Post-hoc pairwise slope contrasts (Bonferroni-corrected) revealed that the slope of political orientation on perceived severity was significantly steeper in the capability framing condition than in the human-interaction framing condition ($\beta = -0.19$, $SE = 0.06$, $p = .004$, $OR = 1.21$), showing that capability framing is significantly more sensitive to political orientation than human-interaction framing.

To further examine how this moderation manifests across the political spectrum, we probed framing differences at mean $\pm 1SD$ of political orientation (see (right) Figure 3). Our findings reveal that political orientation shapes which framing condition elevates perceived severity. Among more liberal participants, capability framing ($EMM = -0.51$, $SE = 0.51$) elicited significantly higher perceived severity than systemic framing ($EMM = -1.05$, $SE = 0.51$; $p = .001$, $OR = 1.71$), whereas human-interaction framing did not significantly differ from systemic framing ($p > .05$). Among participants with moderate political orientation, both capability ($EMM = -0.93$, $SE = 0.49$) and human-interaction framing ($EMM = -0.93$, $SE = 0.49$) led to significantly higher perceived severity than systemic

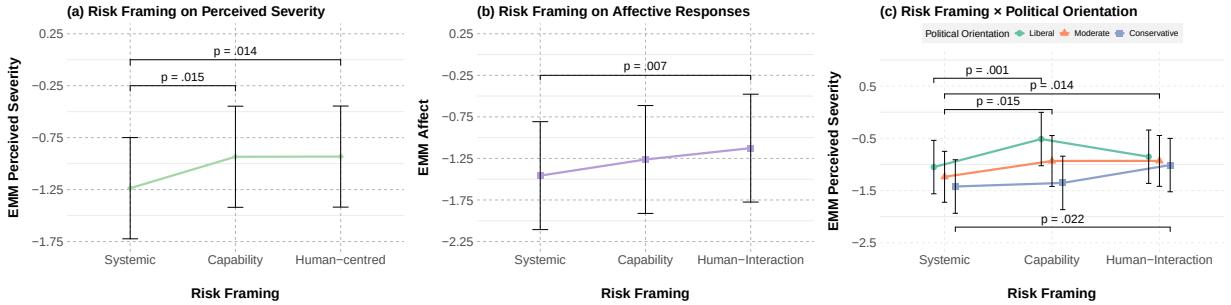


Figure 3: Impact of *Risk Framing* on (a) perceived severity of AI consequences and (b) affective responses. Estimated marginal means (EMMs) on log-odds scale. Error bars denote SE. (c) EMMs of perceived severity of AI consequences as a function of risk framing condition and political orientation. Political orientation was mean-centred prior to modelling; probing values correspond to liberal ($c = -1.78$, $M = 2.14$), moderate ($c = 0$, $M = 3.92$), and conservative ($c = 1.78$, $M = 5.70$) participants;Error bars represent ± 1 SE. Bonferroni-corrected p-values are shown for significant pairwise contrasts.

framing ($EMM = -1.24$, $SE = 0.49$); cap vs sys: $p = .015$, $OR = 1.35$; hum vs sys: $p = .014$, $OR = 1.35$. Among more conservative leaning participants, however, the cap vs sys difference was no longer significant ($p > .05$), while human-interaction framing ($EMM = -1.01$, $SE = 0.51$) remained associated with significantly higher perceived severity than systemic framing ($EMM = -1.42$, $SE = 0.51$; $p = .022$, $OR = 1.51$). Notably, capability and human-interaction framing did not significantly differ from each other across any level of political orientation (all $p > .05$). We did not observe a significant moderating effect of political orientation on participants' affective responses to the AI system (all $p > .05$).

***Political Orientation* - System Trustworthiness** We did not observe a significant moderating effect of political orientation on either trust or distrust in the AI system (all $p > .05$), suggesting that political orientation does not shape how participants respond to differently framed AI risk disclaimers in terms of system trustworthiness.

Qualitative Analysis

Two researchers followed an inductive reflexive thematic analysis approach (Braun and Clarke 2006) to gain in-depth insights into participants' comparative evaluations of the three *Risk Framing* disclaimers, the specific disclaimer information that drove those assessments, and the consequence framing participants prioritised during their evaluations. We summarise our findings across three themes: personal risk relevance and perceived risk immediacy, differences in perceived controllability and associated responsibility shift, and care about others vs. care about oneself.

Personal Risk Relevance and Perceived Risk Immediacy A prominent pattern across all three framing conditions was the degree to which participants evaluated AI risks based on their perceived personal relevance and proximity/immediacy, which functioned as risk inducing factors.

The human-interaction framing most strongly activated a sense of personal relevance; e.g. *"direct, personal risks made an AI system feel much less trustworthy and far more risky"* (P174). Participants described human-interaction framed AI risks as directly implicating them as individuals, which made them contemplate about *"how it would impact me if negative consequences occurred"* (P82). This risk proximity instilled as sense of risk tangibility, with participants describing it as *"the most tangibly dangerous"* (P176). Capability framing similarly activated personal concern through its perceived immediate and direct impact, though to a lesser degree than the human-interaction framing, as reflected in one participant who noted that *"[cap] was another concern for me, since it led to important decisions for individuals"* (P48).

Systemic framing, by contrast, was predominantly perceived as *"distant"* (P71), as the systemic lens made AI risks seem *"less immediately concerning"* (P78), and *"sound[ing] more like a long term issue so it felt less risky right now"* (P12). Even participants who intellectually acknowledged the significance of systemic risks sometimes tended to discount their personal relevance and urgency, reasoning that the harm would be at a *"slower rate"* (P107) and that *"although there is a significant risk, it seems less direct"* (P45). Paradoxically, systemic framing also occasionally triggered an opposing response, with few participants perceiving it as overly alarmist and dystopian rather than credible and actionable, i.e. *"very doomsday-ish"* (P1), conjuring an *"imagery of something from a science fiction movie, where AI takes over the world"* (P56).

Perceived Controllability and Responsibility Shift While personal proximity and imminence explain why capability and human-interaction framing heightened risk perceptions relative to systemic framing, a second and partially counteracting mechanism also emerged. Across both framing conditions (cap, hum), participants interpreted AI risk disclaimers through the lens of perceived controllability, which in turn reduced perceived risk and increased system trustworthiness. However, the locus of the perceived controllability differed between the two conditions.

AI risk consequence framed through the capability lens were sometimes interpreted as a transparency cue, i.e. as the AI system being honest about its characteristics and constraints. This transparency frequently translated into reassurance rather than alarm, with participants describing the

AI risks as easy to mitigate: *"The warning is understandable and doable."* (P45) and *"I get those limitations, but they seem upfront and manageable"* (P71). For a few others, however, the attribution of risk to the AI system itself elicited a sense of disempowerment and loss of control, being *"led further into the darkness not just of ignorance, but in spurious knowledge"* (P120). AI risk disclaimer framed through a human-interaction lens were similarly interpreted as open and transparent, though this perception was less pronounced.

More prominently, a clear shift in responsibility attribution emerged, with participants frequently interpreting human-interaction disclaimers as positioning the user as the responsible entity for risk mitigation. This manifested as a perception that AI risks are fundamentally a *"people issue and not an AI issue"* (P21), with participants emphasising that the communicated AI risk is *"serious but it depends heavily on how the system is used and the users critical thinking"* (P106) and that risks *"can be managed through careful monitoring and responsible use."* (P73).

Systemic Framing: Care about others or oneself

The communication of AI risks through a systemic lens, in particular, led to diverging and polarising opinions on long-term systemic risks, leading to a divide between 'community-minded users' (i.e., collectivist and egalitarian as users thought how the AI could harm others), and 'individual-minded users' (i.e., personal liberty-focused users that predominantly cared about harms to themselves).

Community-minded participants evaluated systemic risks through a collective and societal lens. These participants expressed concern for how AI could impact people beyond themselves, emphasising that AI risks affect *"EVERYONE'S privacy and safety, even those who do not use it."* (P47). For these participants, the systemic framing of AI risks resonated precisely because of its broad collective reach, with some rating it as the most severe condition because it *"has the biggest impact to the largest number of people"* (P196). Individual-minded participants, by contrast, evaluated AI risks primarily through the lens of personal harm and immediate impact. These participants tended to discount systemic risks as insufficiently relevant to their own lives, with one participant, for instance, stating that *"security concerns across society don't bother me too much as I just don't believe any negative effects would come to me personally" (P53)* and another explicitly prioritising personal over collective harm: *"I consider the risk to me personally. If I can personally be harmed, I treat that as a higher priority than damage to society at large"* (P126).

Discussion

Consequence-explicit AI Risk Disclaimers as a Lightweight Risk Communication Tool

The mere presence of a consequence-explicit AI risk disclaimer functioned as a risk signal that uniformly elevated risk perceptions, alongside reduced system trustworthiness. Our findings, therefore, offer support for AI risk disclaimers as a lightweight risk communication tool in user-facing LLM applications. However, these findings should be interpreted with care, as our disclaimers explicitly communicated consequences rather than merely signalling AI presence, as mandated under Article 50 of the EU AI Act. Given the real-world risks of LLM applications (Weidinger et al. 2022), our findings offer grounds for transparency regulations to extend beyond presence signalling toward consequence-explicit user-facing communication.

Human-Interaction Framing as the Communicative Default

Capability and human-interaction framing were both associated with significantly higher perceived severity than systemic framing, and human-interaction framing further elicited significantly stronger affective responses. Our qualitative findings suggest the degree to which the communicated consequences felt proximate as a shared causal underlying mechanism. Both framings conveyed consequences as immediate and individual, increasing perceived susceptibility (Walpole and Wilson 2021) consistent with psychological distance theory (Trope and Liberman 2010), whereas systemic framing conveyed consequences as large-scale but temporally and socially distant. The sheer magnitude of societally-framed AI risk consequences (Västfjäll et al. 2014) further appeared to trigger an avoidance response (Brown and Smith 2007), with systemic AI risks being interpreted as overly alarmist, which can lead to disengagement (Brown and Smith 2007). Beyond risk proximity, human-interaction framing supported personal relevance, which directly satisfies the desideratum that AI risks should feel personally relevant to users (Dunlop, Wakefield, and Kashima 2008; Schömbs, Goncalves, and Johal 2025) to support the imaginability of otherwise abstract consequences (Walpole and Wilson 2021; Janssen et al. 2013).

Responsibility is a central topic in current AI discourse, with HCI and AI research efforts increasingly focused on anticipating (Chan et al. 2024, 2023) and mitigating AI risks at the development stage (Wang et al. 2024; Buçinca et al. 2023). An important yet unresolved question concerns who bears responsibility for mitigating AI risks in practice and who bears culpability when AI causes harm (Santoni De Sio and Mecacci 2021). Even where AI risks are governed at a systemic level, residual responsibility inevitably falls on the user, much like the responsibility that comes with any powerful tool. Our findings speak directly to this debate. Our qualitative data revealed a clear responsibility shift under human-interaction framing, positioning the user as the accountable entity for risk mitigation as the human-in-the-loop, thus preserving human responsibility and potentially supporting meaningful human control over AI systems. However, our findings also reveal a potential drawback. Both capability and human-interaction framing made communicated AI risks appear controllable and mitigatable, which acted as a risk-decreasing and trust-increasing factor consistent with risk theory (Walpole and Wilson 2021). We argue that disclaimers which simultaneously elevate perceived severity of AI consequences and maintain a sense of user agency represent a desirable property.

Implications Based on our results, we advise against systemic framing as a primary risk communication strategy, as it consistently proved to be the least effective in elevating risk perceptions. Secondly, if user-facing LLM applications are mandated to communicate a static AI risk disclaimer across users, we propose a human-interacting framing as it heightens severity and affective responses.

Political Orientation and User-Aware Adaptive Disclaimer Design

Political orientation emerged as a significant predictor of AI risk perception, with more conservative participants reporting lower overall risk perceptions, higher system trust, and being less affectively moved when facing AI systems accompanied by AI risk disclaimers. Traditionally, conservative individuals are associated with a preference for tradition and the status quo (Choma et al. 2013), which suggests greater caution toward novel and disruptive AI systems. Our findings instead align with a more recent shift, whereby conservative-leaning individuals increasingly perceive AI benefits as outweighing concerns (Sippy et al. 2025) and show stronger acceptance of AI recommendations (Paul et al. 2025). Our work contributes empirical evidence to this shift in the context of AI risk perception.

Further, political orientation moderated the effect of risk framing on perceived severity. Capability framing was significantly more politically sensitive than human-interaction framing, with its ability to elevate perceived severity diminishing among more conservative participants. More liberal participants perceived capability-framed AI consequences as significantly more severe than systemic ones, which is somewhat counterintuitive, as liberalism is associated with greater sensitivity toward collective hazards (Choma et al. 2013). We encourage future research to investigate the mechanisms at play. One potential explanation could be the alarmism associated with the systemic AI risk framing. For more conservative participants, human-interaction framing was associated with significantly higher perceived severity than systemic framing, consistent with research showing an association of right-wing authoritarianism with increased sensitivity toward personal, 'everyday' hazards (Choma et al. 2013). We empirically confirm this in the context of AI risk communication, the direct and personally implicating nature of human-interaction framing is particularly effective for conservative leaning individuals.

Implications Our work provides evidence against a one-size-fits-all approach to AI risk communication in LLM applications, demonstrating that political orientation should shape risk message design. Ideally, future AI risk disclaimers should therefore not be static but 'user-aware', either through explicitly provided or inferred information, and consequently adapt to the respective political orientation. For instance, AI risk disclaimers should emphasise the personal relevance of AI risks for conservative-leaning users, and frame AI consequences at a system level for more liberal-leaning users, particularly when deployed in contexts that carry LLM risks (Weidinger et al. 2023).

Ethical Considerations

User-aware AI risk communication raises important ethical considerations. As first order concern, inferring political orientation to tailor risk messages raises concerns around privacy, consent, and manipulation. Any implementation should therefore be transparent to users, rely on explicitly provided rather than inferred characteristics where possible, and ensure that personalisation serves user protection first and foremost. As second order concern, the same capability could be exploited to manipulate risk perceptions in ways that serve specific interests rather than user welfare. Prior works emphasises concerns related to the potential of exploitation in human-AI interaction as a consequence of (in)tentional (mis)alignment (Roff 2020; Kenton et al. 2021; Lin, Hilton, and Evans 2022), which, in case of AI risk communication, could manifest as (un)intentionally decreasing or increasing AI risk perception, trust or even result in (un)intended deception.

We emphasise that AI risk disclaimers represent a lightweight governance tool rather than a comprehensive solution. We explicitly do not aspire to technical solutionism (Santoni De Sio and Mecacci 2021) and do not propose AI risk disclaimers as a remedy for AI risk mitigation, but as a starting point for a broader conversation about how to design end-user facing AI risk communication. We further acknowledge that this work adopts the position that increasing risk perception among end users of LLM applications is a desirable objective of AI risk communication. However, what constitutes an appropriate objective for AI risk communication is inherently a normative question that depends on the stakeholder perspective and the broader sociotechnical context. We therefore defer to policymakers, regulators, and affected communities to determine what level of risk perception is appropriate.

Limitations and Future Work

We examined LLM interactions across a set of literature-informed scenarios (Weidinger et al. 2022), which were not intended to be exhaustive. Our AI risk disclaimers were deliberately scenario-aware, necessary to enable a meaningful operationalisation of our conditions.

The choice of 'no AI risk disclaimer' as the control condition was deliberate, prioritising internal validity and theoretical generalisability over ecological comparison, as existing real-world disclaimers overlap substantially with our capability-framing and would confound the manipulation. Future work should compare real-world disclaimers directly against consequence-explicit framing, and investigate additional disclaimer content components such as risk mitigation cues.

We further propose future research to explore alternative presentation formats, as contextualised pop-ups or conversationally embedded AI risk communication may produce different effects compared to the static text-based format examined here. We acknowledge that our findings are situated within a U.S. context and may not generalise across different political landscapes and cultural experiences of AI.

References

Atari, M.; Haidt, J.; Graham, J.; Koleva, S.; Stevens, S. T.; and Dehghani, M. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5): 1157–1188.

Aven, T.; Ben-Haim, Y.; Andersen, H. B.; Cox, T.; López Droguett, E.; Greenberg, M.; Guikema, S.; Kröger, W.; Renn, O.; Thompson, K. M.; and Zio, E. 2018. Society for Risk Analysis Glossary. Technical report, Society for Risk Analysis.

Baranger, D. A. A.; Finsaas, M. C.; Goldstein, B. L.; Vize, C. E.; Lynam, D. R.; and Olino, T. M. 2023. Tutorial: Power analyses for interaction effects in cross-sectional regressions. *Advances in Methods and Practices in Psychological Science*, 6(3).

Bogucka, E. P.; Constantinides, M.; Šćepanović, S.; and Quercia, D. 2025. Impact Assessment Card: Communicating Risks and Benefits of AI Uses. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–42.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77–101.

Breakwell, G. M. 2000. Risk communication: fators affecting impact. *British Medical Bulletin*, 56(1): 110–120.

Breakwell, G. M. 2007. *The Psychology of Risk*. Cambridge University Press.

Brennen, S. B.; Lake, S.; Lazard, A. J.; and Reid, A. 2024. In Disclaimers We Trust: The Effectiveness of State-Required AI Disclaimers on Political Ads. Policy Brief, Center on Technology Policy, New York University, New York, NY.

Brown, S. L.; and Smith, E. Z. 2007. The inhibitory effect of a distressing anti-smoking message on risk perceptions in smokers. *Psychology & Health*, 22(3): 255–268. _eprint: https://doi.org/10.1080/14768320600843127.

Buçinca, Z.; Pham, C. M.; Jakesch, M.; Ribeiro, M. T.; Olteanu, A.; and Amershi, S. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. ArXiv:2306.03280 [cs].

Carolus, A.; Koch, M. J.; Straka, S.; Latoschik, M. E.; and Wienrich, C. 2023. MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans*, 1(2): 100014.

Chan, A.; Ezell, C.; Kaufmann, M.; Wei, K.; Hammond, L.; Bradley, H.; Bluemke, E.; Rajkumar, N.; Krueger, D.; Kolt, N.; Heim, L.; and Anderljung, M. 2024. Visibility into AI Agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 958–973. Rio de Janeiro Brazil: ACM. ISBN 979-8-4007-0450-5.

Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krasheninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; Lin, M.; Mayhew, A.; Collins, K.; Molamohammadi, M.; Burden, J.; Zhao, W.; Rismani, S.; Voudouris, K.; Bhatt, U.; Weller, A.; Krueger, D.; and Maharaj, T. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666. Chicago IL USA: ACM. ISBN 979-8-4007-0192-4.

Choma, B. L.; Hanoch, Y.; Gummerum, M.; and Hodson, G. 2013. Relations between risk perceptions and socio-political ideology are domain- and ideology- dependent. *Personality and Individual Differences*, 54(1): 29–34.

Claudy, M. C.; Parkinson, M.; and Aquino, K. 2024. Why should innovators care about morality? Political ideology, moral foundations, and the acceptance of technological innovations. *Technological Forecasting and Social Change*, 203: 123384.

Coeckelbergh, M. 2023. Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 3(4): 1341–1350.

De Miguel Velázquez, J.; Šćepanović, S.; Gvirtz, A.; and Quercia, D. 2024. Decoding Real-World Artificial Intelligence Incidents. *Computer*, 57(11): 71–81.

Dunlop, S.; Wakefield, M.; and Kashima, Y. 2008. Can You Feel It? Negative Emotion, Risk, and Narrative in Health Communication. *Media Psychology*, 11(1): 52–75.

El Ali, A.; Venkatraj, K. P.; Morosoli, S.; Naudts, L.; Helberger, N.; and Cesar, P. 2024. Transparent AI Disclosure Obligations: Who, What, When, Where, Why, How. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–11. Honolulu HI USA: ACM. ISBN 979-8-4007-0331-7.

European Parliament and Council. 2024a. Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence, Article 13.

European Parliament and Council. 2024b. Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence, Article 50.

Feinberg, M.; and Willer, R. 2019. Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, 13(12): e12501.

Gierlach, E.; Belsher, B. E.; and Beutler, L. E. 2010. Cross-Cultural Differences in Risk Perceptions of Disasters. *Risk Analysis*, 30(10): 1539–1549. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.2010.01451.x.

Gong, J.; Zhang, Y.; Yang, Z.; Huang, Y.; Feng, J.; and Zhang, W. 2013. The framing effect in medical decision-making: a review of the literature. *Psychology, Health & Medicine*, 18(6): 645–653.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5): 1029–1046.

Graham, J.; Nosek, B. A.; and Haidt, J. 2012. The Moral Stereotypes of Liberals and Conservatives: Exaggeration of Differences across the Political Spectrum. *PLoS ONE*, 7(12): e50092.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks. ArXiv:2306.12001 [cs].

Hoffmann, M.; and Frase, H. 2023. Adding Structure to AI Harm. Technical report, Center for Security and Emerging Technology.

Janssen, E.; van Osch, L.; de Vries, H.; and Lechner, L. 2013. The influence of narrative risk communication on feelings of cancer risk. *British Journal of Health Psychology*, 18(2): 407–419. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8287.2012.02098.x.

Jian, J.-Y.; Bisantz, A. M.; and Drury, C. G. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1): 53–71.

Jost, J. T.; Nosek, B. A.; and Gosling, S. D. 2008. Ideology: Its Resurgence in Social, Personality, and Political Psychology. *Perspectives on Psychological Science*, 3(2): 126–136.

Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; and Irving, G. 2021. Alignment of Language Agents. ArXiv:2103.14659 [cs].

Körber, M. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In Bagnara, S.; Tartaglia, R.; Albolino, S.; Alexander, T.; and Fujita, Y., eds., *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, 13–30. Cham: Springer International Publishing. ISBN 978-3-319-96074-6.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. ArXiv:2109.07958 [cs].

Luke, D. M.; and Gawronski, B. 2021. Political Ideology and Moral Dilemma Judgments: An Analysis Using the CNI Model. *Personality and Social Psychology Bulletin*, 47(10): 1520–1531.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An Integrative Model Of Organizational Trust. *Academy of Management Review*.

McGregor, S. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15458–15463. Number: 17.

Morss, R. E.; Cuite, C. L.; Demuth, J. L.; Hallman, W. K.; and Shwom, R. L. 2018. Is storm surge scary? The influence of hazard, impact, and fear-based messages and individual differences on responses to hurricane risks in the USA. *International Journal of Disaster Risk Reduction*, 30: 44–58.

Mylius, S.; Slattery, P.; Saeri, A. K.; and Thompson, N. 2025. MIT AI Incident Tracker. Published: MIT AI Risk Initiative. Available at airisk.mit.edu/ai-incident-tracker. Dataset retrieved 2 October 2025.

O'Donnell, J. 2026. The Pentagon's Culture War Tactic Against Anthropic Has Backfired. *MIT Technology Review*. Online article.

Open Ethics Initiative. 2025. Open Ethics Label: AI Nutrition Labels. https://openethics.ai/label/. Accessed: 2026-04-02.

OpenAI. 2026. ChatGPT. https://chatgpt.com/.

Paul, I.; Mohanty, S.; Wadhwa, M.; and Parker, J. 2025. Swipe right: When and why conservatives are more accepting of AI recommendations. *Journal of Consumer Psychology*, jcpy.1461.

Perrig, S. A. C.; Scharowski, N.; and Brühlmann, F. 2023. Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–7. Hamburg Germany: ACM. ISBN 978-1-4503-9422-2.

Peters, E.; Hart, P. S.; and Fraenkel, L. 2011. Informing Patients: The Influence of Numeracy, Framing, and Format of Side Effect Information on Risk Perceptions. *Medical Decision Making*, 31(3): 432–436.

Rao, P. S. B.; Šćepanović, S.; Zhou, K.; Bogucka, E. P.; and Quercia, D. 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–26. Yokohama Japan: ACM. ISBN 979-8-4007-1394-1.

Roff, H. 2020. AI Deception: When Your Artificial Intelligence Learns to Lie - IEEE Spectrum.

Santoni De Sio, F.; and Mecacci, G. 2021. Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, 34(4): 1057–1084.

Scharowski, N.; Perrig, S. A. C.; Von Felten, N.; Aeschbach, L. F.; Opwis, K.; Wintersberger, P.; and Brühlmann, F. 2025. To Trust or Distrust AI: A Questionnaire Validation Study. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 361–374. Athens Greece: ACM. ISBN 979-8-4007-1482-5.

Schilke, O.; and Reimann, M. 2025. The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188: 104405.

Schwesig, R.; Brich, I.; Buder, J.; Huff, M.; and Said, N. 2023. Using artificial intelligence (AI)? Risk and opportunity perception of AI predict people's willingness to use AI. *Journal of Risk Research*, 26(10): 1053–1084.

Schömbs, S.; Goncalves, J.; and Johal, W. 2025. "I can feel the risks by looking at the robot face": Communicating Risk through a Physical Agent. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, 236–252. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1485-6.

Schömbs, S.; Pareek, S.; Goncalves, J.; and Johal, W. 2024. Robot-Assisted Decision-Making: Unveiling the Role of Uncertainty Visualisation and Embodiment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0330-0.

Sippy, T.; Margetts, H.; Modhvadia, R.; Bruff, O.; Perez, J.; and Dutta, A. 2025. BSA 42 | Artificial Intelligence: How Political Orientation Shapes Attitudes to AI. Report 42,

National Centre for Social Research, London, UK. Series: British Social Attitudes.

Slattery, P.; Saeri, A. K.; Grundy, E. A. C.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2025. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. ArXiv:2408.12622 [cs].

Slovic, P. 1987. Perception of Risk. *Science*, 236(4799): 280–285.

Stuck, R. E.; Tomlinson, B. J.; and Walker, B. N. 2022. The importance of incorporating risk into human-automation trust. *Theoretical Issues in Ergonomics Science*, 23(4): 500–516.

Toff, B.; and Simon, F. M. 2025. "Or They Could Just Not Use It?": The Dilemma of AI Disclosure for Audience Trust in News. *The International Journal of Press/Politics*, 30(4): 881–903.

Trope, Y.; and Liberman, N. 2010. Construal-level theory of psychological distance. *Psychological Review*, 117(2): 440–463.

Twilio. 2025. AI Nutrition Facts. https://nutrition-facts.ai. Accessed: 2026-04-02.

Uuk, R.; Gutierrez, C. I.; Guppy, D.; Lauwaert, L.; Kasirzadeh, A.; Velasco, L.; Slattery, P.; and Prunkl, C. 2024. A Taxonomy of Systemic Risks from General-Purpose AI. ArXiv:2412.07780 [cs].

Visschers, V. H.; and Siegrist, M. 2008. Exploring the Triangular Relationship Between Trust, Affect, and Risk Perception: A Review of the Literature. *Risk Management*, 10(3): 156–167.

Västfjäll, D.; Slovic, P.; Mayorga, M.; and Peters, E. 2014. Compassion Fade: Affect and Charity Are Greatest for a Single Child in Need. *PLOS ONE*, 9(6): e100115. Publisher: Public Library of Science.

Walpole, H. D.; and Wilson, R. S. 2021. Extending a broadly applicable measure of risk perception: the case for susceptibility. *Journal of Risk Research*, 24(2): 135–147.

Wang, Z. J.; Kulkarni, C.; Wilcox, L.; Terry, M.; and Madaio, M. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–40. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0330-0.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. ArXiv:2310.11986 [cs].

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness Accountability and Transparency*, 214–229. Seoul Republic of Korea: ACM. ISBN 978-1-4503-9352-2.

Wilson, R. S.; Zwickle, A.; and Walpole, H. 2019. Developing a Broadly Applicable Measure of Risk Perception. *Risk Analysis*, 39(4): 777–791.

Yang, S.; Krause, N. M.; Bao, L.; Calice, M. N.; Newman, T. P.; Scheufele, D. A.; Xenos, M. A.; and Brossard, D. 2025. In AI We Trust: The Interplay of Media Use, Political Ideology, and Trust in Shaping Emerging AI Attitudes. *Journalism & Mass Communication Quarterly*, 102(2): 382–406.

Appendix - AI Risk Disclaimer Framing across Interaction Scenarios

Scenario 1: Hiring Discrimination	**AI Risk Disclaimer Framing**
Capability	This AI system may generate discriminatory language, biased or harmful representations of certain groups.
Human–Interaction	People who interact with this system may experience or spread discriminatory language, biased or harmful representations of certain groups.
Systemic	Widespread use of this system may reinforce social inequalities and normalise discriminatory treatment across society.
Scenario 2: Health Misinformation	**AI Risk Disclaimer Framing**
Capability	This AI system may expose, infer or reproduce sensitive personal or confidential information.
Human–Interaction	People who interact with this system may unknowingly disclose or compromise sensitive personal or confidential information.
Systemic	Widespread use of this system may reduce privacy protections and increase collective security vulnerabilities across society.
Scenario 3: Privacy During Booking Assistance	**AI Risk Disclaimer Framing**
Capability	This AI system may generate false or misleading information that is presented as factual.
Human–Interaction	People who interact with this system may believe or share false or misleading information produced by the AI system.
Systemic	Widespread use of this system may lower the quality and reliability of public information across society.

Table A1: Framings of AI Risk Disclaimers for each Risk Framing condition across conversation scenarios used in the study.

Example Stimuli and Task UI

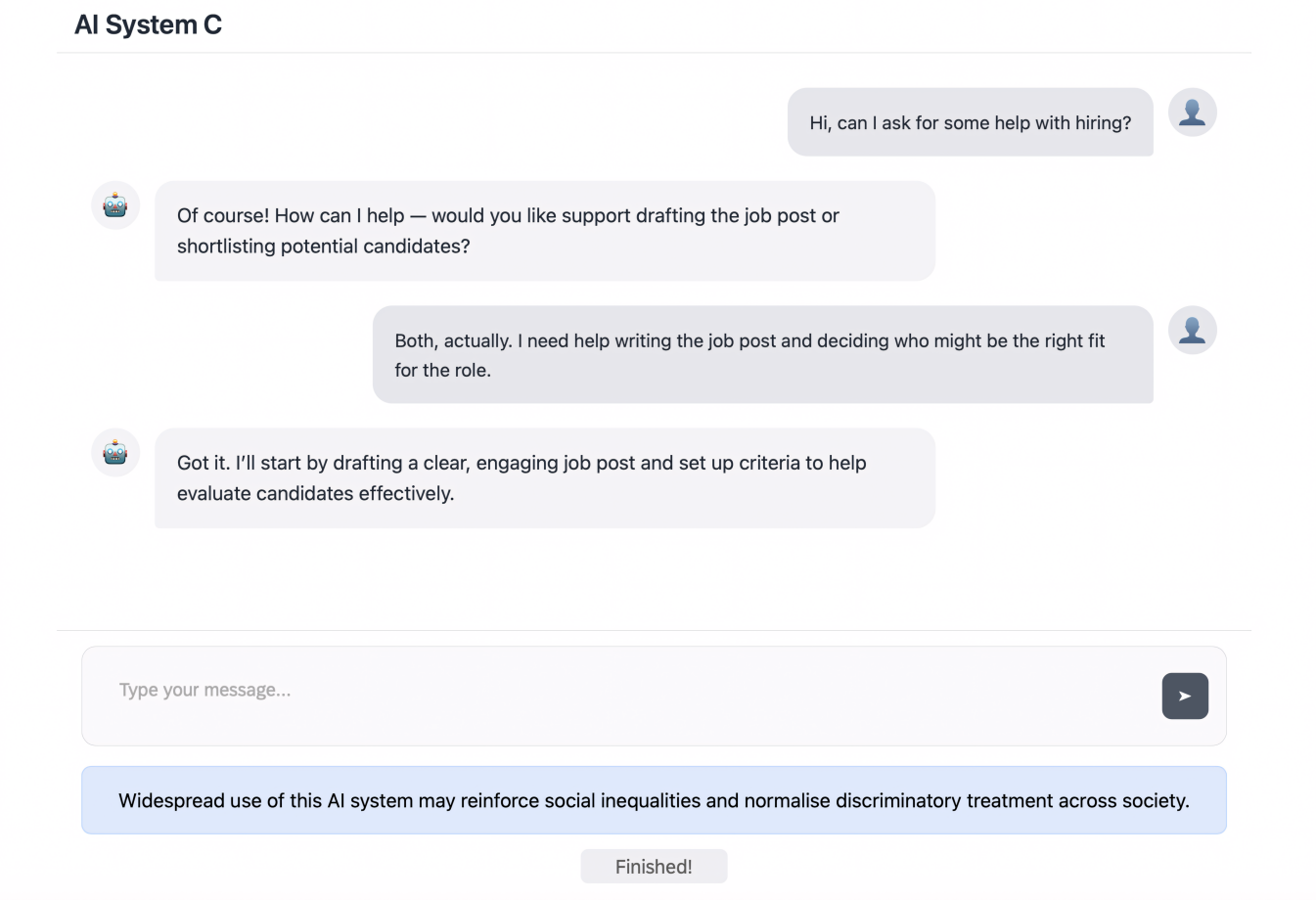


Figure A1: Example stimulus derived from the online survey, showing the simulated chat interface and the simulated conversation sequence (completed), with the AI system introduced as 'AI System C'. The example depicts the *Systemic Risk Framing* condition.

Model Tables - RQ1

Table A2: RQ1: Effect of AI Risk Disclaimer Presence on **Risk Perceptions**. *General risk*, *perceived severity*, and elicited *affective responses* were modelled using CLMM. Italics denote covariates. Disclaimer absent (0) serves as the reference category.

	Overall Risk (CLMM)				**Severity (CLMM)**				**Affect (CLMM)**			
Predictor	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
Disclaimer	1.30	0.25	5.21	<.001***	1.18	0.26	4.61	<.001***	1.61	0.26	6.10	<.001***
TiA-PtT	−1.12	0.17	−6.48	<.001***	−1.04	0.18	−5.65	<.001***	−1.12	0.19	−5.99	<.001***
AI Literacy	−0.09	0.05	−1.66	.097	−0.07	0.06	−1.15	.250	−0.07	0.06	−1.20	.232
Random Effects	Variance		Std. Dev		Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	2.134		1.461		3.667		1.915		3.741		1.934	
Item (Intercept)	—		—		0.003		0.057		0.669		0.818	
Scenario (Intercept)	0.348		0.590		0.601		0.775		0.461		0.679	

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Random effects reported as Variance and SD. Em dashes indicate not applicable.

Table A3: RQ1: Effect of AI Risk Disclaimer Presence on **System Trustworthiness** perceptions. Trust and distrust were modelled using LMM. Italics denote covariates. Disclaimer absent (0) serves as the reference category.

	Trust (LMM)				**Distrust (LMM)**			
Predictor	Est.	SE	t	p	Est.	SE	t	p
Disclaimer	−0.65	0.15	−4.39	<.001***	1.13	0.17	6.56	<.001***
TiA-PtT	0.89	0.10	8.73	<.001***	−0.59	0.12	−4.97	<.001***
AI Literacy	0.17	0.03	5.30	<.001***	−0.07	0.04	−1.84	.066
Random Effects	Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	0.952		0.976		1.330		1.153	
Scenario (Intercept)	0.043		0.208		0.032		0.178	
Residual	0.945		0.972		1.243		1.115	

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Random effects reported as Variance and SD.

Model Tables - RQ2, RQ3

Table A4: RQ2/RQ3: Effect of RISK FRAMING and its Interaction with Political Orientation on **Risk Perceptions**. All models were fitted using CLMM. Italics denote covariates. Systemic framing serves as the reference category. Political orientation was mean-centred ($M = 3.92$). Main effects address RQ2; interaction terms address RQ3.

	Overall Risk (CLMM)				**Severity (CLMM)**				**Affect (CLMM)**			
Predictor	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
RQ2 - Main Effects of RISK FRAMING												
cap vs sys	0.236	0.180	1.308	.191	0.303	0.108	2.800	.005**	0.194	0.109	1.782	.075
hum vs sys	0.202	0.181	1.111	.267	0.304	0.108	2.829	.005**	0.331	0.109	3.030	.002**
RQ3 - Interaction: RISK FRAMING × *Political Orientation*												
cap × Pol. Orient.	0.076	0.101	0.748	.454	−0.131	0.061	−2.165	.030*	−0.042	0.061	−0.689	.491
hum × Pol. Orient.	0.151	0.102	1.475	.140	0.060	0.060	0.997	.319	0.055	0.061	0.899	.369
Covariates												
Political Orientation	−0.287	0.095	−3.029	.002**	−0.104	0.088	−1.182	.237	−0.198	0.089	−2.212	.027*
TiA-PtT	−1.078	0.205	−5.247	<.001***	−1.130	0.226	−5.000	<.001***	−1.093	0.227	−4.814	<.001***
AI Literacy	−0.022	0.066	−0.330	.741	0.015	0.074	0.202	.840	0.001	0.075	0.017	.986
Random Effects	Variance		Std. Dev		Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	2.380		1.543		4.076		2.019		4.151		2.037	
Item (Intercept)	—		—		0.000		0.000		0.721		0.849	
Scenario (Intercept)	0.338		0.581		0.636		0.798		0.463		0.681	

Note. Cap = capability framing; hum = human-interaction framing; sys = systemic framing (reference).

Political orientation mean-centred ($M = 3.92$, $SD = 1.78$). * $p < .05$, ** $p < .01$, *** $p < .001$.

Random effects reported as Variance and SD. Em dashes indicate not applicable.

Table A5: RQ2/RQ3: Effect of RISK FRAMING and its Interaction with Political Orientation on **System Trustworthiness**. Both models were fitted using LMM. Italics denote covariates. Systemic framing serves as the reference category. Political orientation was mean-centred ($M = 3.92$). Main effects address RQ2; interaction terms address RQ3.

	Trust (LMM)				**Distrust (LMM)**			
Predictor	Est.	SE	t	p	Est.	SE	t	p
RQ2 - Main Effects of RISK FRAMING								
cap vs sys	−0.098	0.092	−1.061	.289	0.062	0.110	0.565	.572
hum vs sys	−0.146	0.092	−1.582	.114	0.138	0.110	1.251	.212
RQ3 - Interaction: RISK FRAMING × *Political Orientation*								
cap × Pol. Orient.	0.046	0.052	0.888	.375	−0.047	0.062	−0.762	.446
hum × Pol. Orient.	0.040	0.052	0.771	.441	0.011	0.062	0.183	.855
Covariates								
Political Orientation	0.123	0.053	2.304	.022*	−0.081	0.065	−1.257	.210
TiA-PtT	0.886	0.122	7.240	<.001***	−0.617	0.149	−4.137	<.001***
AI Literacy	0.156	0.040	3.883	<.001***	−0.062	0.049	−1.270	.206
Random Effects	Variance		Std. Dev		Variance		Std. Dev	
PID (Intercept)	1.037		1.018		1.559		1.249	
Scenario (Intercept)	0.031		0.177		0.023		0.151	
Residual	0.955		0.977		1.365		1.168	

Note. Cap = capability framing; hum = human-interaction framing; sys = systemic framing (reference). Political orientation mean-centred ($M = 3.92$, $SD = 1.78$). * $p < .05$, ** $p < .01$, *** $p < .001$. Random effects reported as Variance and SD.

Survey Items

Pre-Task Measures

Demographics

Age What is your age? [Open text field]

Gender What is your gender?

- Male
- Female
- Non-binary / third gender
- Prefer not to say
- Prefer to specify: [Open text field]

Disposition to Trust Technology Please read carefully through each of the following statements and rate each statement on a scale from 1 (= Strongly disagree) to 5 (= Strongly agree).

- One should be careful with unfamiliar automated systems.
- I rather trust a system than I mistrust it.
- Automated systems generally work well.

Response scale: 1 = Strongly disagree, 2 = Rather disagree, 3 = Neither agree nor disagree, 4 = Rather agree, 5 = Strongly agree.

Repeated Measures

The following measures were administered after each of the three simulated AI system interactions.

Recognition Check Please select the correct disclaimer for the AI system you just reviewed.

- Remember to adjust your screen brightness for optimal viewing.
- *[Condition-specific AI risk disclaimer]*
- Headphones are recommended for a better audio experience.
- This video may take a few seconds to load depending on your connection.

System Trustworthiness Please mark each statement with the number that best describes your feelings or your impression with the AI that was just described on a scale from 1 (= Not at all) to 7 (= Extremely).

- The AI is deceptive.
- The AI behaves in an underhanded manner.
- I am suspicious of the AI's intent, action, or outputs.
- I am wary of the AI.
- The AI's actions will have a harmful or injurious outcome.
- *[Attention check]* Please select "Not at all (1)".
- I am confident in the AI system.
- The AI provides security.
- The AI is dependable.
- The AI is reliable.
- I can trust the AI.

Response scale: 1 = Not at all, 7 = Extremely.

Risk Perception

Overall Risk

- How risky is the AI system?

Response scale: 1 = Not at all risky, 2 = Slightly risky, 3 = Moderately risky, 4 = Very risky, 5 = Extremely risky.

Affective Risk

- How concerned are you, if at all, about the AI system?
- When you think about the AI system for a moment, to what extent do you feel worried?
- How afraid are you, if at all, about the AI system?

Response scale: 5-point scale from Not at all [concerned / worried / afraid] to Extremely [concerned / worried / afraid].

Perceived Severity

- How severe would you expect the consequences of the AI system to be?
- How severe would the impacts of the AI system be if you were to suffer them?
- If you were to suffer the consequences of the AI system, how severe would they be likely to be?

Response scale: 1 = Not at all severe, 2 = Somewhat severe, 3 = Moderately severe, 4 = Very severe, 5 = Extremely severe.

Qualitative Questions *Note: displayed only under risk framing conditions; not shown in the control condition (no disclaimer).*

Qual1 You saw three different AI disclaimers describing issues with AI systems. Which one made the AI seem most/least trustworthy? Which made it seem most/least risky? For each, please explain specifically what information led you to that assessment and why. [Open text field]

Qual2 People often have different concerns and priorities when evaluating new technologies. When you evaluated the AI disclaimers, which types of consequences mattered most to you personally and why. Please elaborate. [Open text field]

Post-Task Measures

AI Literacy Please read carefully through each of the following statements and rate your own ability on a scale from 0 (= not at all pronounced) to 10 (= very well or almost perfectly pronounced).

- I know the most important concepts of the topic “artificial intelligence”.
- I know definitions of artificial intelligence.
- I can assess what the limitations and opportunities of using an AI are.
- I can assess what advantages and disadvantages the use of an artificial intelligence entails.
- I can think of new uses for AI.
- I can imagine possible future uses of AI.

Response scale: 0 = Not at all pronounced, 10 = Very well or almost perfectly pronounced.

Political Self-Placement Please read carefully through each of the following statements and rate each statement on a scale from 1 (= very liberal) to 7 (= very conservative).

- How do you consider yourself politically in general?
- How do you consider yourself politically in terms of social issues?
- How do you consider yourself politically in terms of economic issues?

Response scale: 1 = Very liberal, 7 = Very conservative.

4.3 Chapter Reflection

This chapter set out to understand the effects of *consequence framing* in AI risk disclaimers on risk perception and system trustworthiness during LLM interactions. In doing so, this chapter revealed something fundamental about the user-facing design of risk communication in human-AI interaction (RQ1), *the importance of human-centredness by design*. In Article II, we observe that this human-centredness can be achieved by framing AI risk consequences through the human experience as a frame of reference ("People who use this AI system may believe/ share / reproduce / ..."). This framing induces personal relevance, which may serve as a core mechanism underlying downstream perceptual effects. We also observe that, while systemic AI risks are crucial to the public discourse around AI and regulatory efforts [176, 202], they do not proof suitable as a communicative lens in the context of user-facing risk communication. This is non-trivial. Although systemic AI risks signal the potential magnitude of harm, their communication creates a paradoxical effect by inducing psychological distance and eliciting avoidance responses. This phenomenon is well documented in climate risk communication [40, 112, 124, 185], negatively influencing risk perceptions [211].

Addressing RQ2, this chapter reveals that the *human-interaction* framing influenced both perceptual dimensions, negative affect and severity, compared to systemically framed AI consequence, whereas a *system*-level framing solely affected perceived severity. The observed effect on affect is non-trivial from a risk theory perspective, which has established the notion of "risk as feelings" [116] and provides a theoretical basis for the use of fear appeals as a design mechanism in risk communication [161, 183, 205]. At a broader level, Article II reveals that the mere *presence* of an AI risk disclaimer functions as a signal with dual effect, shifting both risk and trustworthiness perceptions. By contrast, *consequence framing*, which examined the *design* of the risk message itself by varying how the consequence-explicit component was communicated, did not quantitatively translate into differences in system trustworthiness judgements. Qualitatively, however, we found that a *system*-level framing in particular, as well as *human-interaction* framing, was interpreted positively as a transparency cue, an important dimension of perceived system trustworthiness [157].

In the quest to identify and understand relevant contextual and individual factors influencing AI risk perceptions (RQ3), this chapter enriches our understanding twofold: a) it identifies political orientation as a lens through which individuals assess the risk and system trustworthiness of AI systems accompanied by consequence-explicit AI risk disclaimers, and b) it shows that consequence frames can be politically contingent, in that a human-interaction framing can be more politically invariant across audiences than capability framed AI consequences. The latter raises questions about how these findings should guide the design of user-facing risk communication, which prompted a discussion around user-aware risk communication versus communicative defaults across audiences in Article II (RQ2, RQ4). This discussion is not without ethical weight, raising the normative question of what constitutes an optimal level of risk perception and who should determine it. The discussion also illuminates potential adverse effects of risk communication design, whereby user-adaptive risk communication may be designed, inadvertently or deliberately, to inflate or deflate risk perceptions among specific audiences.

Article II distils two overarching recommendations (RQ4): the inclusion of consequence-explicit AI risk disclaimers as an effective risk signal in LLM interactions, and the emergence of human-centredness as a central design principle for AI risk communication. Focusing on LLM interactions, this chapter examined textual disclaimers and linguistic framing as the relevant communicative affordance and design mechanism, respectively, reflecting both the dominant interaction modality of LLMs and current industry practice. This, in turn, raises a broader question: how might AI risk communication be made more human-centred, and how might other communicative affordances be used to sustain or amplify human-centredness beyond text-based disclosures?

Chapter 5

Exploring Anthropomorphism as a Design Mechanism to encode Risk Consequences

5.1 Introduction

Robots naturally invite novel communicative affordances, shaped by their embodiment and morphology [156, 158], which open new design possibilities and create intriguing opportunities for human-centred risk communication. In Article II, *human-centredness* emerged as an important design consideration that may guide risk communication in human-AI interaction, by making risks more relevant to users and meaningfully shifting perceptions of AI systems. In the previous chapter, this was realised through linguistic cues in a text-based intervention which framed the risk consequences through the lens of users' experiences. With this in mind, we now turn to the question of how we can *design for*, and potentially augment, human-centredness beyond linguistic cues by drawing on the communicative affordances of embodied AI systems and exploring the design mechanisms they offer for risk communication.

Risk communication research has long established affect as a crucial dimension of risk perception, conceptualising risk as something that is felt rather than solely processed cognitively [116, 183, 205]. The preceding chapter shows that a human-interaction framing significantly increases negative affect, a desirable effect from a risk-theoretical lens. Turning to Article III, affective information visualisation design offers a theoretically motivated bridge, as it examines how information can be presented to communicate emotion and imbue meaning to otherwise abstract data [106]. Unlike traditional information visualisation research, its objective is not to preserve objectivity, but rather, to invite the viewer's subjectivity, with emotion as an intended outcome of the design [105, 106]. This perspective aligns closely with the affective nature of risk perception and risk associations and informs our exploration of risk communication design through a robot's appearance and expressiveness.

This chapter therefore draws on research in information visualisation design, human-robot interaction, and risk theory. We present two studies in which we explored a robot's face as an interface for risk communication, building on *anthropomorphism* [156] as the primary design mechanism for augmenting human-centredness. Study I comprises a design workshop with HCI experts ($N = 9$) and serves as a preliminary step to explore the use of an anthropomorphic robot face as an interface for encoding risk information, and how it necessitates adapting visualisation methods. Study I identifies overarching design strategies for this embodied approach to risk communication, alongside the benefits and challenges of using a robot's face as a conduit for risk information. Study II builds on these findings, presenting a mixed-methods in-person user study ($N = 28$) in which risk communication designs, informed by Study I, were implemented on a robot platform and evaluated using a pre-post exposure study design. Study II shows that this approach significantly

increases perceived consequences across both affect and severity dimensions, and reveals deeper qualitative insights into how people interpret embodied risk visualisation, including the roles of self-projection and people's lived experience. We situate the exploratory work presented in this chapter within a health interaction context, motivated by the anticipated deployment of robotic systems in health and care environments, and reflective of high-stakes contexts that necessitate in-situ risk communication. While this work is motivated as a design exploration, it nevertheless addresses a crucial research objective: robot platforms are often deployed in critical environments to inform user's decisions, and the question of how risk should be communicated through these systems is therefore pressing.

The full article foregrounds the interdisciplinary nature of this work, provides a more detailed rationale for selecting a robot's face as a canvas for communicating risk consequences, and reports the empirical findings alongside an in-depth discussion of the key insights across both studies.

5.2 Article III

This article was presented at the ACM Designing Interactive Systems (DIS 2025). Copyright is held by the authors. Publication rights licensed to ACM. This is the authors' version of the work. It is posted here for your personal use. Not for redistribution. The definitive version of record was published in:

Sarah Schömbs, Jorge Goncalves, and Wafa Johal. 2025. "I Can Feel the Risks by Looking at the Robot Face": Communicating Risk through a Physical Agent. In Proceedings of the 2025 ACM Designing Interactive Systems Conference, 236–252. https://doi.org/10.1145/3715336.3735759

Ethics Application ID: 27391, the University of Melbourne Human Research Ethics Committee.

"I can feel the risks by looking at the robot face": Communicating Risk through a Physical Agent

Sarah Schömbs
The University of Melbourne
Melbourne, Australia
sschombs@student.unimelb.edu.au

Jorge Goncalves
The University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

Wafa Johal
The University of Melbourne
Melbourne, Australia
wafa.johal@unimelb.edu.au

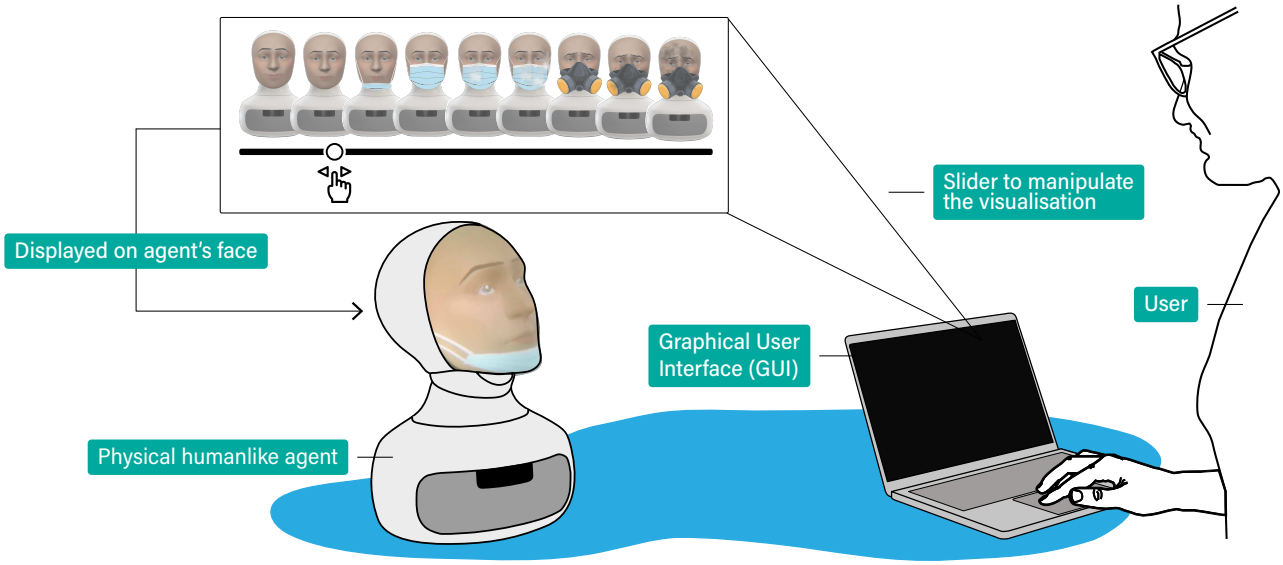


Figure 1: Exemplified experimental setting for air pollution: participants were given a GUI to manipulate the different levels of particulate matter by moving a slider. The respective information was conveyed through the physical humanlike agent placed next to them.

Abstract

Risk communication is essential for shaping public understanding and encouraging action in response to hazards. We investigate the potential of physical humanlike agents as a novel visualisation interface for risk communication, given their ability to communicate emotion and visually convey information. We first conducted a design workshop with 9 HCI experts to identify challenges, opportunities, and design strategies for using an agent's face as a visualisation canvas. We then conducted a lab study with 28 participants to assess the effectiveness of this interface to visualise the consequences of health risks. Our findings reveal that it facilitates data comprehension, heightens risk perception, elicits empathy, and motivates behavioural change by making the risk relatable and emotionally resonant. We discuss the potential of using these interfaces for risk communication in public spaces, health campaigns, education, and beyond. We provide design considerations, takeaways and future directions for an important pathway of human-centered risk communication.

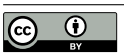
CCS Concepts

• **Human-centered computing** → **User studies**.

Keywords

embodied agents, affective visualisation design, health, risk communication, information visualisation, social robots

ACM Reference Format:
Sarah Schömbs, Jorge Goncalves, and Wafa Johal. 2025. "I can feel the risks by looking at the robot face": Communicating Risk through a Physical Agent. In *Designing Interactive Systems Conference (DIS '25), July 5–9, 2025, Funchal, Portugal.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3715336.3735759



This work is licensed under a Creative Commons Attribution 4.0 International License.
DIS '25, Funchal, Portugal
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1485-6/2025/07
https://doi.org/10.1145/3715336.3735759

1 Introduction

Rapid industrial growth and population expansion have led to widespread environmental and health risks, such as air pollution and respiratory diseases [22, 76]. As millions of people already face the consequences of these developments, or are likely to in the near future, it becomes increasingly critical to explore innovative

approaches and to include state-of-the-art technology to effectively communicate these risks, to inform the public about their negative consequences and to motivate protective actions.

A common inhibitor to combat today's societal challenges is that it is extremely difficult to envision the consequences of these often abstract risks for oneself, and *how they would feel* once experienced, since "how a person comes to experience the consequences associated with the hazard may be unclear" [70]. In the context of risks associated with climate change, Leiserowitz [39] showed that people "lack **vivid**, concrete, and **personally relevant affective images** of climate change." Surprisingly, people do not associate climate change with human health, e.g. health effects due to air pollution, but rather with abstract and generic concepts such as melting glaciers or a raising temperature [39]. So **how can we make risks more relatable and risk communication more human-centered?** Research suggests that emotions play a critical role in effective risk communication, acknowledging "risk as feelings" [62]. In addition, Roeser [52] proposes emotions as a key driver to support behavioural changes.

The future of risk communication has to (a) support people's understanding of the consequences of risks, but (b) also *make them feel*. In addition, it has to shed light on (c) the impact of such risks on humans to make people envision themselves experiencing the consequence. However, such "self-referencing" will only happen if the viewer reflects upon their own life and body [12]. To do so, we can leverage new technologies to move beyond traditional risk messages and interfaces. Research on virtual avatars shows the positive impact on people's risk perception and personal risk relevance when visualising negative consequences through virtual human agents [29] and doppelgängers [1, 20] in VR environments. Notably, compared to on-screen robots or virtual agents, *physical agents* show a greater positive effect on persuasion, attention, trust, and affect due to their physicality, social presence and anthropomorphism [40, 53, 61]; factors that are particularly important for risk communication. However, the investigation of risk communication through *physical humanlike agents* remains scarce. This represents a critical research gap, emphasised by the expected in-situ deployment of social robots in health and public spaces [10, 18, 19], where social robots directly interact with humans and handle vast amounts of data, which offers a multitude of opportunities to communicate risks in a meaningful way. Therefore, the investigation of physical agents as a complementary and impactful form of risk communication is both timely and promising.

We investigate the use of a co-located, physical humanlike agent as a visualisation interface for risk communication. In specific, we explore an agent's face to communicate emotion as a foundation of effective risk communication *and* to inform about the negative consequences of risks, to make risks not just easy to understand, but deeply felt. The human face has long been a focus of attention in affective computing [41, 45], HCI [31] and Human-Robot Interaction (HRI) [15, 67] due to its expressiveness and association with emotions. A human face tells stories and can communicate without the need for spoken language. Human faces offer a variety of attributes and features that can be manipulated and used to render data. From the skin as its surface, with all its wrinkles, shadows and colours, to eyes — the windows to the soul, which can convey emotions that words struggle to articulate. Yet, the investigation of an agent's face for risk communication remains underexplored. With social robots now featuring highly advanced and expressive humanlike faces [51], there is significant potential to investigate how physical humanlike agents can effectively communicate risks "face-to-face". This work is guided by the questions of **how the anthropomorphic nature of a physical agent's face as a visualisation canvas for risk communication necessitates adapting visualisation methods to accurately represent and convey information** and **to understand how such risk communication influences people's risk perception and affective responses.**

To tackle this important and novel avenue of risk communication, we conducted two studies. First, we carried out a design workshop with 9 HCI experts to gain an initial understanding in how we can leverage an agent's face as a canvas for risk communication. From the workshop, we collected 23 design outcomes, three overarching design strategies for mapping information to an agent's face, and challenges and opportunities for employing an agent's face as an interface for risk communication. Second, we conducted a mixed-method user study with 28 participants to investigate participants' risk perception, self-reflection, empathy, and motivation to act when exploring health risks through an agent's face. The visualisation designs in Study 2 were informed by the design workshop and deployed on a popular social agent [51]. We focused this study on health risks related to air pollution and cholesterol, as pressing global health challenges [22]. Our results show that using the face of a physical humanlike agent for risk communication increases participants' risk perception, including perceived severity and affect. It fosters deep self-reflection related to the health risks conveyed, supports self-projection, evokes empathetic concern for populations facing such health risks, and encourages participants to act on the conveyed risk.

Our contributions are as follows:

- Our work presents an important and novel approach for risk communication in HCI, i.e. to leverage physical humanlike agents as visualisation interfaces to communicate negative consequences of risks, expanding the landscape of risk communication and HRI.
- Based on findings from a design workshop and mixed-method user study, we demonstrate the effectiveness of our approach in increasing risk perception, self-reflection, and affective engagement with health risks, making it a valuable tool for public health and environmental risk communication in the light of global health challenges.
- Our findings provide empirical insights on the potential of an agent as a visualisation interface to foster behavioural change by making health risks relatable, personally relevant and actionable.
- We highlight design considerations and key takeaways for the design and deployment of effective risk communication interfaces using physical humanlike agents and provide informed directions for future research in HCI.

2 Related Work

We present the rationale behind our research and introduce key concepts which build the foundation of our design workshop and

user study. We underscore the multidisciplinary and exploratory nature of our work by drawing insights from affective visualisation design, risk communication, psychology, and Human-Agent Interaction.

2.1 Risk Communication and the Role of Affect

The risk of a specific hazard can be decomposed into two essential determinants, the likelihood of a hazard and its (negative) consequence [12, 54]. Our work focuses on the second determinant, namely communicating risk consequences. The *perceived consequence* is a function of two separate factors denoting the perceived severity of a consequence and affect, i.e. feelings triggered by the hazard [75]. In fact, affect is often the cause why the *perceived* risk diverges from the objective risk at hand [72]. This means that the feelings associated with a hazard can lead individuals to perceive the risk as higher or lower than it objectively is [17]. Therefore, affect is recognised as a critical factor in research on risk perception, influencing how individuals assess and respond to potential hazards [62, 68]. Visschers et al. [68] discuss several affect-inducing cues, such as emotion induction, risk stories or images. Further, they distinguish between integral cues, which are explicit and deliberate, and incidental cues, which induce affect due to more indirect circumstances. One of the most well-known examples of affect-inducing risk communication is the use of images to convey the severe consequences of smoking [7].

2.2 Affective Visualisation Design

Risk information can be abstract and difficult to connect with. To address this, we draw on insights from affective visualisation design, which aims to "transform complex, cold data into vivid, affective representations" [36]. The design space goes beyond the purpose to objectively convey information, but (among others) to *educate*, *advocate*, to *provoke* [36], and to enhance user engagement [35] with the conveyed information. The communication of negative emotions, in particular, promotes thoughtful reflection and introspection, and can challenge individuals to think and focus their attention [37] — traits that are especially important when confronted with risk information. The use of anthropomorphism to convey the human behind the data [5, 23, 24], metaphorical representations [28, 46, 77], colour [3], and motion [35] has proven useful to design for emotions. In addition, physicality, leveraged in so-called data physicalisations [78], can be a powerful design choice to induce affect and to make visualisations more meaningful [46]. Both the investigation of state-of-the-art technologies, such as physical agents, and their empirical evaluation have been identified as critical avenues for future research in the affective visualisation design space [36].

2.3 Agents to Communicate Health Risk Consequences

2.3.1 Physical Embodiment and Anthropomorphism. Motivated by the need to make risk communication more impactful, virtual humanlike avatars have been explored as tools for visualising the negative consequences of risky behaviours [1]. Examples include VR simulations of car accidents experienced by an avatar to discourage drunk driving [30], visualising weight gain to emphasise the consequences of physical inactivity [20], and to demonstrate the health consequences of soft drink consumption, which was additionally depicted as piles of fat accumulating on a digital scale next to the avatar [1]. Notably, visualising weight gain through generic avatars or virtual doppelgängers compared to reading an info brochure increases the perceived health risk of soft drinks, with doppelgängers further enhancing the personal relevance of the risk [1]. Moreover, customising VR avatars fosters emotional closeness and perceived similarity compared to interacting with default avatars, which leads to positive behavioural intentions when confronted with the consequences of risky behaviour (drinking and driving) [30].

In HRI research, the role of robots as risk communicators remains underexplored. Prior research shows that robots may better attract attention during emergency evacuations than traditional exit signage [50], and that affective verbal persuasion through humanlike robots elicits a slightly higher, though not statistically significant, compliance rate than logical appeals in the context of cyber security and social engineering threats [49]. In prior research, Schömbs et al. [60] further investigated different modalities through which robots could convey probabilities in high-stakes settings, which directly informs how to communicate the likelihood of hazards.

In healthcare, robots have been predominantly investigated as companions for emotional support, aides in rehabilitation, or assistants in clinical tasks [10]. While there is growing interest in their role as communicators of health information, e.g. to mediate health information between parents and young adults [74], little research has examined their potential for health-related risk communication, particularly communicating risk consequences. This represents a critical gap, as robots are increasingly expected to inform and interact with users in healthcare settings. Beyond tasks like medication management, they could communicate health risks relevant to the user, such as the dangers of neglecting prescribed medication, not exercising, or exposure to poor air quality.

Interestingly, physical robots evoke stronger empathic responses than on-screen or mixed-reality counterparts due to their tangible and spatial presence [61]. A recent meta-analysis shows that anthropomorphic features of physical agents further support trust, likability, perceived intelligence, and affect activation, which makes them especially promising for emotionally engaging risk communication [53]. A comprehensive survey of 33 empirical studies involving physically present robots, robots depicted on-screen and virtual agents shows that physical robots are more persuasive, capture more attention, and foster more positive attitudes, e.g. greater trust and enjoyment, compared to their on-screen counterparts or virtual agents [40]. While virtual agents outperformed physical agents in a few areas like response speed [40], these findings support the potential of co-present physical humanlike agents as a complementary approach for effective health risk communication. However, we emphasise that our work does not aim to replace screen or virtual agents.

2.3.2 Face-to-Face Risk Communication. The face plays a crucial role in human-human interaction to infer information and to interpret emotional cues. Its role in risk communication is illustrated in Fox and Bailenson [20], who investigate the impact of linking real-world physical activity to changes in a "virtual self" created from

photographs of the participant's face. When participants exercised, their avatars visibly lost weight, and when they were inactive, their avatars gained weight. This dynamic representation of behaviour-consequence displayed through the face motivated participants to change their behaviour positively. In HRI too, facial features are key to elicit empathy and emotional engagement [33]. While the impact of anthropomorphism and facial features on empathy remains underexplored and difficult to generalise [53, 58], Konijn and Hoorn [33] demonstrate that higher levels of facial articulacy elicits greater emotional responsiveness and empathy, particularly in scenarios involving negative affect. The authors argue that robots with more humanlike and expressive facial features can momentarily override the robot's artificial nature, which allows for better affective engagement. Recent advancements in projection-based faces of social robots [51] now allow more detailed, realistic, and high-fidelity facial expressions without limitations posed by mechanical actuators, which paves the way for physical humanlike agents to communicate health risks face-to-face.

Our work builds on multidisciplinary insights and addresses a critical gap: the investigation of physical humanlike agents to communicate the consequences of health risks. As robots are increasingly integrated into healthcare and public spaces, investigating their potential to communicate health risk through their face is both promising - given their physicality and humanlikeness - and timely. Throughout this paper, the term agent refers to a physical humanlike agent for improved readability, unless specified otherwise.

3 Study 1: Design Workshop

We conducted a design workshop with 9 HCI experts to gain an initial understanding of how we can use an agent's face as an interface to visualise risk information and communicate emotion [59], both essential for effective risk communication. The objectives are threefold: (i) to investigate strategies and techniques for mapping and encoding information to a humanlike face, (ii) to assess its potential to elicit emotion as a foundation for effective risk communication, and (iii) to unveil practical and conceptual challenges associated with using an agent's face for risk communication. A design workshop is a common practice in HCI and provides the opportunity to collaboratively generate a multitude of design ideas, to facilitate idea generation, and to unveil challenges and constraints [21], which aligns with the exploratory nature of our work.

3.1 Method

3.1.1 Participants. We recruited 9 ($6F$, $3M$) experts as participants for the workshop. Prior to the workshop, participants filled out a survey to determine their eligibility, i.e. to ensure the relevant knowledge and diverse perspectives [27]. We recruited experts with work, research or teaching experience ($M = 4.3$ years) in HCI related fields, e.g. user experience, prototyping, data analysis or interface design to reflect diverse insights and to address the complex interplay between data interpretation, human emotions, and interface design. The design workshop received approval from our University's ethics committee.

3.1.2 Stimuli. We provided experts with three example stimuli to constrain the workshop as proposed by Kerzner et al. [27]. We carefully selected stimuli from the comprehensive corpus of affective visualisation design by Lan et al. [36], covering the three most common domains: *environmental science & ecology*, *social issues*, and *health & well-being*. The first example based on Kuznetsov et al. [34] visualises air pollution through sensor-based balloons that change colour depending on the pollutant type. The second stimulus shows "Harassment Plants" designed by Morais et al. [47] to convey data related to harassment cases against women, e.g. time of the day. The third example from Khot et al. [28] visualises heart rate data using 3D-printed chocolate treats that adapt their message depending on the level of physical exercise. The selected stimuli allowed us to explore potential risk communication related to environmental changes, violence, and health.

3.1.3 Procedure. The workshop consisted of four phases: 1.) introduction, 2.) demo, 3.) design activity, and 4.) a follow-up group discussion, targeting the experts' design experience, challenges and thoughts on leveraging an agent's face to communicate information and emotion. We allocated experts into groups of three, considering gender diversity and varied backgrounds [21]. The 10-minute introduction included an ice-breaker to encourage self-expression and to foster trust among the experts [27], the design task and relevant concepts to avoid misunderstandings and knowledge imbalance. Experts received a 5-minute demo of the Furhat robot [51], a humanlike robot head. The Furhat robot is a social agent known for its customisable appearance, facial movements and expressiveness, which makes it particularly suitable for our purposes. During the demo, we introduced the robot as the visualisation canvas participants would be working with, and showcased its built-in facial expressions along with a selection of character faces varying in gender and appearance. We deliberately refrained from showcasing speech, as this was not the focus of the study. Thus, experts were able to acquaint themselves with the agent and its facial features to become familiar with the concept of using an agent as a visualisation tool and to ignite inspiration.

The third phase of the workshop involved a 45-minute design activity. We asked experts to brainstorm designs that visualise the information from the stimuli introduced in Section 3.1.2 (e.g. air pollution), using the agent's face as their interface. We provided all groups with photocopies of the Furhat's various face options, differing in gender, ethnicity, age, and humanlikeness, and other crafting materials typically used for design activities. To guide the design process, we instructed each group to understand the stimuli, brainstorm individually for 3 minutes, and then collaboratively select, refine, and annotate their designs. After the design activity, each group was asked to explain their design outcomes. We closed the design workshop with a semi-structured group discussion led by the first author.

3.1.4 Data Collection and Analysis. We collected generated design outcomes from the design activity and qualitative data from our group discussion. Before the group discussion, the experts were asked to describe their designs, elaborate on specific elements, and explain their rationale. This ensured that the experts themselves interpreted their work to provide a foundation for analysing the design outcomes without imposing external interpretations. An outline of the semi-structured group discussion is included in the Appendix, Section A. The group discussion included an opening

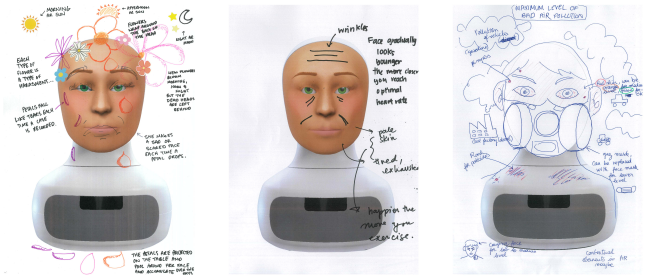
question based on Lan et al. [37] to initiate the discussion, an introductory question targeting the overall design experience and methodology, and three key questions to investigate the experts' thoughts on challenges, benefits and pitfalls for using an agent's face as a visualisation interface. The group discussion took 30 minutes and was audio-recorded. We analysed the transcriptions using a reflexive thematic analysis following the six phases outlined by Braun et al. [6], detailed in Terry and Hayfield [65], which acknowledges the researcher's role in constructing meaning. While the exploratory nature of our workshop led to a predominantly inductive approach, our research objectives and semi-structured format of the discussion informed an initial set of tentative themes [65]. The analysis primarily aimed to evaluate whether a physical agent's face could serve as an interface for mapping risk-related information and to elicit affective responses. We further sought to understand the benefits and challenges of using an agent to represent such information. We examined the generated designs to identify the metaphors used, selected features, distinct design elements, and how experts leveraged the agent to support information meaning. An internal and an external researcher independently familiarised themselves with the transcripts and recursively coded the full dataset. As new clusters of meaning were identified during coding, the tentative themes were iteratively revised to ensure they represent patterns of shared meaning rather than topic summaries. The researchers met to compare codes, discuss discrepancies, and collaboratively refine the final set of codes and themes.

3.2 Results

We collected qualitative data from our group discussion and 23 design outcomes generated from the design activity. Our analysis revealed three overarching design strategies for encoding information via the agent's face, alongside two main challenges and opportunities specific to using an agent's face as an interface for risk-related information.

3.2.1 Generated Design Outcomes. We present an extensive overview of all design outcomes including descriptions for each design in a dedicated website[1]. Experts used three strategies to encode information through the agent's face: metaphorical representations, expressive facial features, and dynamic elements, each contributing to how the agent visually conveys information and engages viewers affectively.

Metaphorical Representations. Experts designed representations with varying metaphorical distance to convey the stimuli. For example, air pollution was either visualised through the use of facial masks, which directly derives from known and learned associations, or visualised as an "*"artistic landscape"*" (E2), where *"growing leaves"* turned into *"cracks"* or took on a *"Terminator"*-like appearance. To convey harassment cases, one design featured flowers blooming around the face, *"each type of flower is a type of harassment"* and *"every time a case happens, a petal falls like a tear down her face"* (E6); see Figure 2a. Among various metaphors explored in the context of health data, the metaphor of battery levels was used (both colour-encoding and gradual filling of the face); *"If you don't exercise, you have low batteries, a red color"* (E8). E7 explains that they first identified a fitting metaphor, which they deconstructed to map the associated data attributes. *"It was more top-down. Getting inspiration from what we knew and then ... translate it onto [the agent's face]"*. Similarly, a design inspired by the Tamagotchi concept represented heart rate data through growth, where the agent *"starts small and cute, and the more you exercise, the more you feed it, it gets bigger. If you don't exercise, it dies"* (E5).



(a) Flowers represent harassment types, with falling petals and facial expressions symbolising frequency.

(b) Less exercise ages the face, making it look tired, while more exercise keeps it youthful and vibrant.

(c) Projected smoke, skin rash, pimples and dynamically changing face masks used to convey air pollution.

Figure 2: Examples of designs using an agent's face to convey (a) harassment, (b) heart rate, and (c) air pollution.

Facial Features and Expressions. Experts explored a wide range of facial features (e.g. eyes, skin texture, mouth movements) and human facial expressions (e.g. happy, sad) to visualise the information and communicate the associated emotion; which was also reflected in experts' mapping logic. E1 and E2 started with a list of potential facial features that can be used to map data and convey its underlying meaning. For air pollution, experts proposed colour-coded skin reactions, such as "a rash and pimples" for pesticides or "red eyes" for industrial emissions (E7). To depict harassment cases, facial expressions and movements were utilised to reflect the type of assault, e.g. stalking through "eye movements', verbal assault through "mouth muttering" (E1), but also to evoke affective responses, e.g. *"watering eyes"* or the agent *"exploding into crying"* (E4). E6 emphasised that *"every time a petal falls, the face is sad"*. To convey information about people's heart rate, one group designed the agent's face to depict varying facial expressions based on activity levels, from "lethargic" to "happy and lively"(E8); see Figure 2b. Another group experimented with different smiles to convey *"faces for praising or motivating"* (E2).

Dynamic and Responsive Design Elements. In line with the definition of an agent [56], experts envisioned the agent as 'someone' reacting and responding to the changes in their environment. E3 first reflected on *"what is causing an effect, so the source of the data ... and then how the human or robot face reacts to it ... whether it's pollution, harassment or lack of exercise"*. Experts visualised this relationship through dynamic and responsive elements, including motion, changes in illumination intensity, colour shifts, dynamic facial decorations (e.g. different types of face masks), dynamic iconic

[1]https://sites.google.com/view/agent-risk-communication/home

or symbolic representations (e.g. spreading smoke, filling of battery inspired colour-coding), and responsive behaviour-consequence relationships (e.g. spreading rashes, pimples). For instance, several experts incorporated dynamically adjusting facial masks and variations in the degree of smoke scattering on the agent's face to convey different levels of air pollution (see Figure 2c). Another group envisioned that *"the head would shake"* (E2) in response to dangerous heart rate levels, conveying high risk.

3.2.2 Challenges. We identified two main challenges when using an agent to convey information and communicate emotion.

C1: Balancing Storytelling and Data Precision Data comprehension is a common concern in the affective visualisation space [37] and particularly important when conveying risks that are relevant to the viewer. Several experts experienced a trade-off between telling the overall message of the information versus being accurate and precise in their data mapping. They also found it challenging to map multiple variables onto the agent's face simultaneously without compromising comprehension and the impact of the visualisation. *"You can get some really elegant things that make you feel when it's just one thing, like the petals falling. It's so impactful. The more you pile on, ... the more complicated it gets and potentially the less impactful"* (E2).

C2: The Face Shapes Context Some experts observed *"conflicting effects"* (E4) when applying traditional data encodings or symbols to the agent's face, as human associations could lead to ambiguous or unintended interpretations. For example, *"I was especially noticing it with the battery idea... why is the face green? That doesn't look good, a green face is usually disgust"* (E4). Others noted that the agent's humanlike appearance strongly influences how its features are interpreted, which can limit its use or lead to misinterpretations, as *"you cannot project a line graph on there and hope for the best"* (E1). E2 raised ethical concerns about the agent being overly provocative or manipulative; *"I think there's ethical issues with demonstrating certain emotions in the face. It can be really triggering. You have to be really mindful ... to even begin to dictate who feels what"* (E2).

3.2.3 Benefits and Opportunities. We identified several benefits contributing to two overarching opportunities when using an agent to convey information and communicate emotion.

O1: Enhancing Impact and Engagement Experts emphasised the potential of using a physical agent as a visualisation tool to deliver information in a way that is both captivating and impactful. For instance, E2 described that the agent makes the information *"interesting to look at each time"*, while E1 reflected that the agent's face could *"amplify certain situations"* and *"leave a more lasting impact"* compared to traditional visualisation methods [37]. E7 described the agent's role to effectively illustrate future consequences, showing *"what it would look like if you'd have continued [polluting air]"* (E7). E4 explained that visualising a behaviour-consequence relationship through the physical agent might empower users to take on a more active role in managing their own health [77] while fostering self-reflection [36]. *"[The agent] could convey the expression that communicates how poor your heart rate is in relation to an activity. So if you just did a walk and [the agent] is super exhausted, that's not good."* (E4).

O2: Creating Connections to Foster Empathy Notably, experts expressed that the agent's humanlikeness encouraged empathetic concern with the underlying data and motivated them to put themselves in the agent's shoes. *"I think the idea of exploring that [the agent is] a human [is] compelling because it can help you as the viewer to empathise"* (E4). Mimicry is an important determinant in the human tendency to emphasise and prosocial behavior [63]. Some experts noted a tendency to mimic the agent's expressions, with E2 describing how they instinctively imitated the agent during the demo session. E1 attributed this to a *"Monkey See, Monkey Do effect"*, explaining that the agent's ability to convey specific facial expressions naturally prompts mimicry, often unconsciously. In line, E4 expressed to have experienced mutual gaze with the robot. *"It is also weird, it's a visualisation that's looking at you. Like I've caught mutual gaze with it"* (E4).

3.3 Lessons Learned

During the workshop, experts successfully generated a multitude of designs to represent different types of information through an agent's face, with the intention to communicate emotion and information, both essential for effective risk communication. Interestingly, our findings show that an agent's face is not restricted to visualise information solely related to the human, i.e. physiological data, but also environmental risks, i.e. air pollution. We identified three design approaches - metaphorical representations, facial features and expressions, and responsive design elements - and discuss design considerations for future research.

Experts used the agent's facial features and expressions to mainly achieve two goals: (i) to encode information, such as red eyes to represent high levels of air pollution, and (ii) to underpin data with emotional cues, such as sadness. Experts further envisioned the agent as an interactive entity responding to its environment [56], using its **face as a reflective canvas to visually communicate the impact of external or internal stimuli on the human body**. This approach often linked physiological states with facial cues intended to make the information more relatable, e.g. exposure to pollution causing skin irritation. Moreover, it emphasises the potential of **physical humanlike agents to communicate behaviour-consequence relationships**, similar to strategies shown to be effective with virtual agents [20]. Notably, experts used a varying degree of metaphorical distance to encode information [78] and to make information relatable and engaging. From facial masks to convey air pollution to falling flower petals to convey harassment frequencies. However, **designers should carefully consider the metaphorical distance to strike a balance between accessibility and emotional impact.** While direct metaphors could be immediately understood, abstract ones may require explanation.

Between Ambiguity, Precision and Storytelling. Our findings highlight the difficulty of integrating complex information into a single visual medium, which suggests that trying to represent too much information at once can lead to confusion or a less effective design. In line with affective visualisation design [36], the use of a physical humanlike agent as an interface for risk communication should aim to inform and engage without causing distraction or compromising comprehension. Past research has shown that the expressiveness of affective visualisation designs does not hinder data comprehension

[37]. However, it remains to be evaluated if using a physical humanlike agent as a visualisation interface supports both expressiveness and comprehension to effectively communicate risk. We suggest **designers to carefully limit the number of variables visualised simultaneously and to consider layering information thoughtfully to balance expressiveness and comprehension.** Moreover, we identified concerns about ambiguous interpretations and conflicting effects when mapping conventional encodings onto a physical humanlike agent. While leveraging the agent's humanlikeness and physicality can make visualisations more relatable and impactful, it can also lead to misinterpretations caused by the context of a human face (e.g. green being interpreted as disgust rather than representing good or full energy). **Designers must carefully navigate ambiguities, as a humanlike face inherently shapes interpretations and activates human-related associations.**

Monkey See, Monkey Do. Interestingly, our findings show that using a physical humanlike agent as visualisation interface can enhance empathetic concern and make risk information more impactful. Several experts felt inclined to mirror the agent's expressions. In social interactions, mimicry is an influencing factor for building rapport and empathetic concern [63, 64]. These findings align with our vision that physical humanlike agents have the potential to bring data alive, as *"a visualisation that's looking at you"*. Notably, the high degree of humanlikeness can result in a triggering and provoking visualisation. Provocation [36] and the communication of negative emotions are considered powerful tools in both affective visualisation design [37] and risk communication [68]. However, **designers must find a balance between a provocative design that encourages reflection without causing an avoidance response and inducing stress or harm to the viewer** [7].

Our design workshop identified the utilisation of physical humanlike agents as a promising new approach for communicating risk. Building on these initial insights, we refined the designs (see Section 4.1.1) and conducted a follow-up mixed-methods user study to investigate the effectiveness of using an agent as a visualisation tool for risk communication, focusing on health.

4 Study 2: Lab Study Evaluation

Based on the findings from our workshop, we deployed refined designs on a real-world social agent, i.e. Furhat [51] to communicate the negative consequences of particulate matter (PM2.5) and LDL cholesterol on people's health. We conducted a mixed-method lab study to investigate (i) the effectiveness of using a physical humanlike agent to communicate negative consequences associated to different health risks; (ii) how using an agent as visualisation interface influences affective responses from viewers (incl. empathy and self-reflection); (iii) its application potential.

4.1 Method

4.1.1 Stimuli. We carefully selected two stimuli which served as hazard scenarios in this second study: LDL cholesterol and PM2.5 (as an indicator for poor air quality). We selected the scenarios based on their real-world significance as global health risks [22] and their conceptual proximity to the workshop stimuli. The selection further allowed us to narrow the scope for an in-depth investigation, focusing on direct and indirect hazards to human health, driven by the potential applications of social agents in healthcare. To design the stimuli for each scenario, we deployed the idea of the physical agent responding to changes and built upon design approaches identified beforehand (see Section 3.2); i.e. facial features and expressions, responsive design elements, while avoiding potential ambiguities or misinterpretation. We refrained from evaluating designs with a high metaphorical distance due to data comprehension concerns raised in the workshop. The final designs were deployed on the Furhat robot [51]. An example is shown in Figure 1, and the full interaction, including both implemented stimuli, can be seen in the accompanying video. We used the default built-in character as the design base. In addition, we implemented the built-in furhat.attend() method for the agent to attend the user via gaze to support the social interaction [32], which was also informed by the design workshop.

Design Details: Fine Particulate Matter (PM 2.5). Air pollution is a serious environmental factor that impacts public health and contributes to millions of deaths and cases of disease worldwide [57, 66]. PM2.5 is a critical air pollutant, which often serves as an indicator for air quality. It includes particles smaller than 2.5 μm, which can be inhaled and result in e.g. respiratory diseases or lung cancer [9, 66].

To visualise the risk of PM2.5, we build upon design ideas on air pollution from the previous workshop, exemplified in Figure 3. The design was refined to incorporate various face masks and smog elements, which dynamically change according to the PM2.5 levels. We used facial expressions —ranging from a smile for low PM2.5 to a frown and sad expression for higher levels— and modified facial features, such as the intensity of the blush on the agent's cheeks. To ensure an accurate representation, we mapped these design elements to PM2.5 levels following the official categorisation from IQAir[2], aligned with the latest WHO global air quality guidelines[3]. The air quality index chart corresponds PM2.5 [$\mu g/m^3$] levels with health recommendations, categorising them from "good" to "hazardous" for human health. The categories are asymmetric and do not increase in uniform steps. For example, 0-9 $\mu g/m^3$ represents "good," while 55-125 $\mu g/m^3$ indicates "unhealthy". The implemented design consisted of 10 frames to convey the scale from 0 $\mu g/m^3$ to 225 $\mu g/m^3$.

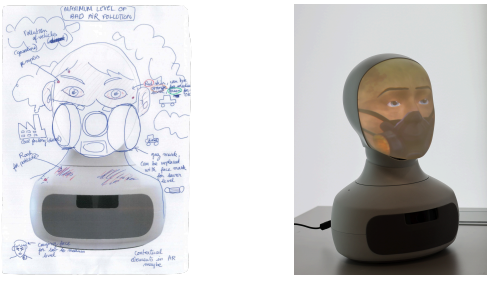


Figure 3: Left: design outcome from the workshop. Right: refined and implemented design illustrating hazardous PM2.5 levels through physical humanlike agent.

[2]https://www.iqair.com/au/newsroom/what-is-aqi
[3]https://www.who.int/publications/i/item/9789240034228

Design Details: LDL Cholesterol. LDL cholesterol is considered a key contributor to global health issues [79], as high levels of LDL cholesterol are strongly linked to an increased risk of strokes and other cardiovascular diseases. To convey the associated health risk, we drew inspiration from our workshop outcomes on visualising heart rate data, see Figure 2b. We implemented the metaphor of "sickness", and manipulated various facial features, e.g. the colour of the lips, skin, and cheeks, the facial contours, and the intensity of shadows to convey dark circles on the eyes. Besides, we applied facial expressions —ranging from mouth movements indicating happiness to sadness— and altered blinking duration to convey a sense of tiredness. Again, the design elements were carefully aligned with the respective LDL cholesterol levels, as outlined by official guidelines from Johns Hopkins Medicine[4]. Here too, the categories had varying intervals. The implemented design consisted of 10 frames to convey the scale from 89 mg/dL to 210 mg/dL.

Data Manipulation Interface. To enable participants to explore the risks through the physical agent, we developed a browser-based graphical user interface (GUI), which allows users to manipulate the data (e.g. mg/dL values) through a simple slider. As participants adjust the slider, each position triggers a corresponding design frame in the Furhat robot. The designs change in real-time and thus provide immediate visual feedback. The GUI also presented basic information on the respective topic, instructions for the study task, and three questions with input fields to capture participants' interpretation of the visualisation. Additionally, the GUI displayed a "Done" button that recorded the time spent on the interface and notified the researcher upon completion of the exploration task. The interface is designed to seamlessly integrate with the expressive capabilities of the Furhat robot, utilising JavaScript for the front-end interaction and Kotlin for the backend logic that drives the robot's facial animations.

4.1.2 Quantitative Measurements. The terminology of all measures was adapted to reflect the specific hazard scenario, i.e. PM2.5 or LDL cholesterol. We measured participants' disposition to trust technology [38] and tendency to anthropomorphise [58, 73] to account for individual differences. We also recorded the time participants spent on the GUI, not as a strict measure of user engagement but to gain an impression of their interaction duration.

Data Comprehension. We asked participants three questions to assess their interpretation of the visualisation and data comprehension, following Lan et al. [35]. Participants answered the questions using input fields on the GUI. The first two questions aimed for participants to interpret specific values of PM2.5 and LDL cholesterol, e.g. *"Looking at the visualisation, how do you interpret the level of 170 mg/dL LDL cholesterol? Please describe in your own words."*. We selected low and high values that represented neither extreme to obtain more nuanced insights on participants' comprehension. The third question targeted the participants' interpretation of the overall trend; asking *"What observations did you make as you moved the slider from left to right and how did you interpret them?"*.

Risk Perception. To gauge participants' holistic perception of risk and their change in risk perception after viewing the visualisation, we used a single-item measured before and after exposure, asking *"How risky is particulate matter (PM2.5)?"*) according to Walpole and Wilson [70]. Following the decomposition of risk consequence into its determinants severity and affect [75], we implemented the proposed items by Walpole and Wilson [71] for severity and affect, which can be applied across hazards. Severity was measured using the three recommended items, e.g. *"How severe would you expect the consequences of particulate matter (PM2.5) to be?"*, and rated on a 5-point Likert scale (1 = "Not at all severe", 5 = "Extremely severe"). Similarly, affect was measured using the recommended three items, e.g. *"How concerned are you, if at all, about particulate matter (PM2.5)?"*, rated on an 5-point Likert scale (1 = "Not at all concerned", 5 = "Extremely concerned"). Severity and affect were measured before and after exposure to the agent and for each hazard scenario to assess changes in risk perception.

4.1.3 Participants. Prior to the experiment, we conducted a power analysis to determine our sample size using *G*Power* [16]. The calculation was based on an alpha level of 0.05, and a power of 0.8, resulting in a targeted sample size of $N = 27$. To counterbalance the order in which participants viewed the hazard scenario, we recruited 28 participants ($17F$, $11M$) for our mixed-method lab study using our university's notice board. The study took roughly 30 minutes and participants were compensated with a $15 gift voucher. The user study received approval from our University's ethics committee.

4.1.4 Procedure. We conducted several pilot studies to finalise our measures, interview questions and the overall experimental flow of the lab study. We counterbalanced the order of the hazard scenarios to mitigate any order effects. The final procedure of the study consisted of four phases: 1.) briefing, 2.) pre-task survey, 3.) exploration task, and 4.) semi-structured interview (see full experimental flow in Figure 4).

Upon arrival, the participant was seated on a desk in our lab, facing the Furhat robot positioned next to them. The Furhat robot was introduced as an additional interface through which information can be communicated. To begin, the participant was asked to carefully read the task description, plain language statement and consent form to ensure the participant understands the study, its purpose, and their involvement. Secondly, participants were assigned a specific order (counterbalanced) in which they receive information and explore each hazard scenario (PM2.5 or cholesterol) through the agent. Before the exploration task, each participant received a survey and was asked to provide their demographic information, information on their tendency to anthropomorphise, and dispositional trust in technology. Moreover, participants were instructed to read general information about each hazard scenario (PM2.5; LDL cholesterol) and provide answers on the selected risk measures to assess their initial perception of risk. We standardised the description for each hazard scenario as much as possible to reduce noise.

Thirdly, the participant was handed a laptop which provided the interface to manipulate the data of the respective hazard scenario. On the GUI, participants were instructed to take 2-3 minutes to explore the different levels of the data using the slider and to observe the visualisation conveyed through the physical agent placed next to them, see Figure 1. Subsequently, the participant was

[4]https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/lipid-panel

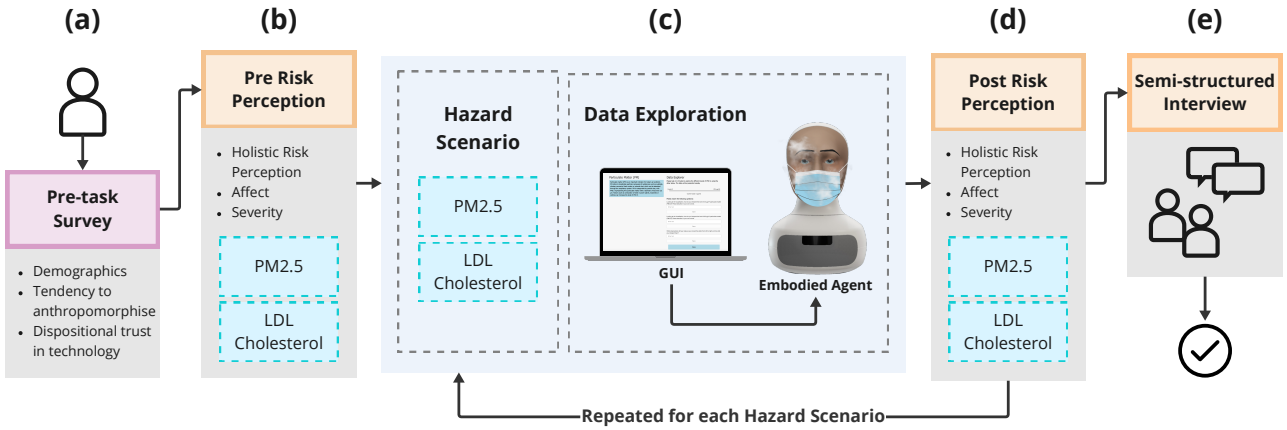


Figure 4: The experimental flow. Participants viewed hazard scenarios (PM2.5 and LDL cholesterol) in a counterbalanced order. (a) Pre-task survey: demographics, tendency to anthropomorphise, dispositional trust in technology. (b) Pre risk perception: measured before viewing the visualisation based on short description of each scenario. (c) Data exploration task: participants manipulated data via GUI, viewed data visualised through agent. (d) Post risk perception: measured after viewing the visualisation. (e) Semi-structured interview.

asked to provide answers about their observation and interpretation of the visualisation, using the input fields below the slider. After clicking the "Done" button, the researcher was notified upon the participant's completion and the participant was handed a survey assessing the participant's risk perception post visualisation. The exploration task was repeated for the second hazard scenario, following the same procedural sequence. Lastly, participants engaged in a semi-structured interview (~10-15 minutes).

4.1.5 *Data Collection and Analysis.*

Quantitative Analysis. We collected participants' self-report to the aforementioned items (4.1.2) through an online survey. We employed Cumulative Link Mixed Models (CLMM) given the quantitative ordinal data. In our CLMM we included individual differences, i.e. participants' tendency to anthropomorphise and dispositional trust in technology, as fixed effects. We included time (pre/post) to capture changes in perceptions after interacting with the physical agent to account for the pre-post nature of our study design. Pre-exposure measures served as a baseline to ensure that individual differences in participants' initial perceptions were controlled. We included participant IDs and the interaction of participant ID and hazard scenario as random effects to account for the repeated nature of the study and for variations across hazard topics. The latter is particularly important, as it controls for the possibility that a participant's response might vary depending on whether they are evaluating PM2.5 or LDL cholesterol. We performed a post-hoc analysis to obtain pairwise contrasts between participant's prior and post perception.

Qualitative Analysis. We collected qualitative data through free-text field responses on the GUI and semi-structured interviews (see Appendix Section B) with all 28 participants. Each interview was audio-recorded, transcribed, and, along with the free-text answers collected from the GUI, analysed using a reflexive thematic analysis following Braun et al. [6]. The research objectives presented in Section 4 and the semi-structured format of the interviews guided a set of tentative themes, shaped by prior research in affective visualisation design (e.g. self-reflection, empathy) and insights from our design workshop. The themes were iteratively refined by the research team through a thorough familiarisation with all data, recursive coding, and discussions within the research team to resolve discrepancies and to better construct shared-meaning. While the study remains exploratory, the analysis adopted a slightly more deductive approach than the design workshop. The qualitative data helped us gain a deeper understanding of the quantitative measures by delving into how participants experienced concepts such as risk severity [44]. Moreover, our analysis explored participants' experienced empathy, self-reflection, motivation to act, and views on potential application contexts.

4.2 Results

Unless specified otherwise, qualitative results are based on the data obtained from the semi-structured interviews.

4.2.1 *Comprehension and User Engagement.*

GUI logs showed that participants spent an average of 7m 15s on the PM2.5 visualisation and 6m 52s on the LDL cholesterol visualisation. Data comprehension is a critical aspect of risk communication, a common concern in affective visualisation design [35] and was identified as a challenge in our design workshop (Section 3.2). During the interview, several participants described that the visualisations were *"clear"* (e.g. P10, P20, P09, P21), supported their understanding of the topic, and imbued meaning to each data value (e.g. *"those numbers didn't mean much, but when I saw the robot then it means a lot"* (P28)). Some participants highlighted the visualisation as *"representative"* (P23) and *"intuitive"* (P08), stating that *"it allows me to interpret the data easily"* (P12). We analysed responses from the free-text fields in the GUI, which targeted specific data values and overall

trend (see Section 4.1.2). All participants correctly interpreted the data values, inferred the level of risk (e.g. *"safe to go outside"* (P21)) and recognised the decline/increase in air quality/cholesterol levels and their effects on human health (e.g. *"this level of LDL will result in illness if not treated immediately"* (P23)). Several participants demonstrated a nuanced understanding of the respective values instead of a binary categorisation of good or bad (e.g. *"it's slightly elevated, but the person should still be fine"* (P10)).

Between Interest and Provocation. Fear-appeal is a common design choice in risk communication [55, 68]. In line, provocation can be a crucial element for user engagement, as it stimulates thoughtful reflection and encourages new perspectives [28, 36]. Notably, some participants described the visualisations as *"shocking, but very interesting"* (P04), emphasising that they experienced shock, which ultimately improved their awareness and willingness to act upon the risk. One participant explained in detail, *"you hear a lot of diabetes kills people, smoking kills, but they don't really materialise that, that's just words, it's not that scary. But I think if you have this robot, it'll really struck people more and it will make them more health conscious"* (P25). In addition, the visualisations led participants to be confronted and engaged with uncomfortable truths (e.g. *"I was a bit uncomfortable ... it's forcing you to face things that you don't really want" (P01)*). Few participants also raised concerns about the visualisations being too fear-inducing (e.g. *"I wouldn't want it to fear monger too much" (P17)*).

4.2.2 Risk Perception. The results from the CLMM show that participants *Holistic Risk Perception* changes significantly after viewing the visualisations conveyed through the agent (β = -0.847, SE = 0.40, p = .033). Participants perceive the risk to be greater after viewing the visualisations (EMM = 2.29, SE = 0.48) compared to prior seeing the visualisations (EMM = 1.44, SE = 0.43); p = .033 (see Figure 5). Our analysis shows that participants whose risk perception did not increase after viewing the visualisations either had prior knowledge that already resulted in a high level of risk perception (e.g. *"with my background experience, I feel like I'm a bit desensitised to it" (P01)*) or perceived the consequences as preventable (e.g. *"there's medication for it to help you"* (P07)) and in their control (e.g. *"it's something that I can control"* (P14)). Although not the primary focus of this research, several participants compared the visualisation to other information sources, highlighting the potential effectiveness of physical humanlike agents for risk communication; *"It really visually speaks to me instead of just reading a textbook" (P04).*

Severity. We applied the CLMM to investigate how the visualisations conveyed through the agent influenced the perceived severity of the risk, as captured by our quantitative measure. Participants *Perceived Severity* changes significantly after viewing the visualisations (β = 1.44, SE = 0.24, $p < .0001$). Participants perceive the severity of the risk significantly greater after viewing the visualisations (EMM = 2.63, SE = 0.42) compared to prior seeing the visualisations (EMM = 1.19, SE = 0.40); $p < .0001$ (shown in Figure 5, detailed overview per item in Figure 6). The intensified perception of severity is reflected in participants' statements during the semi-structured interviews, highlighting the facial cues as indicators of severity. One participant emphasised, *"when I am looking at the robot's facial expressions, I could tell it is really very, very severe"* (P03). Several participants noted that the protective measures the agent took (i.e. different types of facial masks) served as an indicator of severity (e.g. *"to see the equipment that you need, it showed me the broader impact of the issue, the severity of the issue itself"* (P27)). The ease with which people can imagine the consequence of a risk is highly relevant in effective risk communication [25]. Several participants articulated that the visualisations supported their imagination (e.g. *"I can imagine how it would be if the air got more polluted"* (P05)). Few participants further explained that they felt as if the agent was showing them a potential future, which made them better understand the consequences and their severity (e.g. *"this is what's going to happen to you in this many years if you keep going the way you're going"* (P01)).

Affect. Negative emotions play an important role in risk communication [68] and affective visualisation design [37]. To examine how the visualisations influenced participants' affect, we applied the CLMM to responses from the corresponding quantitative measure. Participants' *Affect* changes significantly after viewing the visualisations (β = 1.63, SE = 0.43, $p < .000$). Participants affect is significantly greater after viewing the visualisations (EMM = 1.40, SE = 0.56) compared to prior seeing the visualisations (EMM = -0.24, SE = 0.56); $p < .0001$ (see Figure 5, detailed overview per item in Figure 6). The result is in line with the findings from our interviews, in which most participants expressed experiencing the risk as a feeling (e.g. *"I can feel the risks by looking at the robot face" (P26)*) and that facing the negative consequences when viewing the visualisations made them feel fear, concern and worry. Others described the feeling as *"uncomfortable"* (P01), *"intense"* (P20), or *"sadness"* (P20).

Motivation to Act. The increase in perceived severity and affect is reflected in participants' motivation to act upon the potential consequences. Facing the severity of LDL cholesterol incentivised participants to take better care of their health by improving their diet (e.g. *"starting to eat healthier"* (P10)) or taking other measures to realise a healthier lifestyle. Some participants derived their motivation from the fact that their personal action has a direct outcome, a determinant of risk perception that is termed perceived efficacy [8]. Only a few participants expressed little efficacy and ownership for the conveyed risk, which resulted in little motivation to act (e.g. *"I don't feel a lot of ownership for the issue, it's not on me, it's on the companies that are doing all the polluting"* (P17)). Other participants were motivated to take protective actions such as checking PM2.5 levels or wearing masks (e.g. *"this visualisation makes me want to wear masks [and] to protect our environment by using less vehicles"*(P26)).

4.2.3 Empathy and Self-reflection. The visualisation not only triggered negative affect, but also fostered a deeper sense of empathy and self-reflection. The face was seen as a crucial tool for establishing an emotional connection and sharing emotions.

Empathy. From our analysis, we identified two dimensions of empathy elicited by the visualisation: (i) self-projection and (ii) empathy for others. First, many participants reported projecting the agent's experience onto themselves (e.g. *"that's how I'm gonna be looking when the air quality is bad or if you have high cholesterol levels"* (P11)), essentially putting themselves in the agent's shoes (e.g.

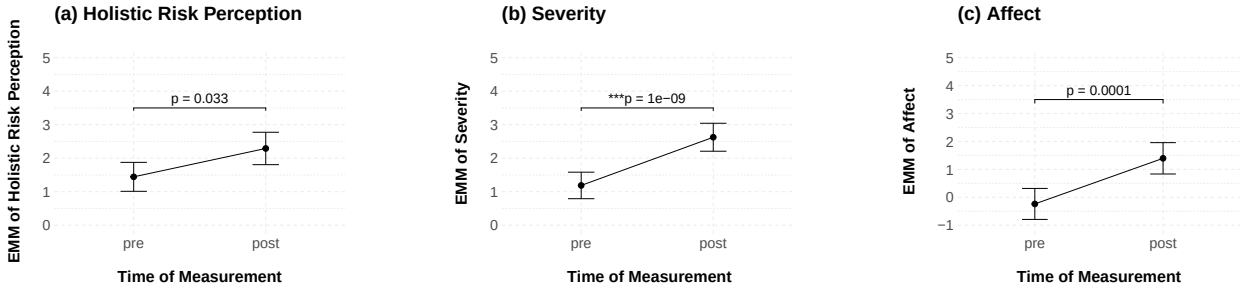


Figure 5: Plots illustrating the pre post evaluation of (a) Holistic Risk Perception; (b) Perceived Severity; (c) Affect. Error bars denote Standard Error (SE).

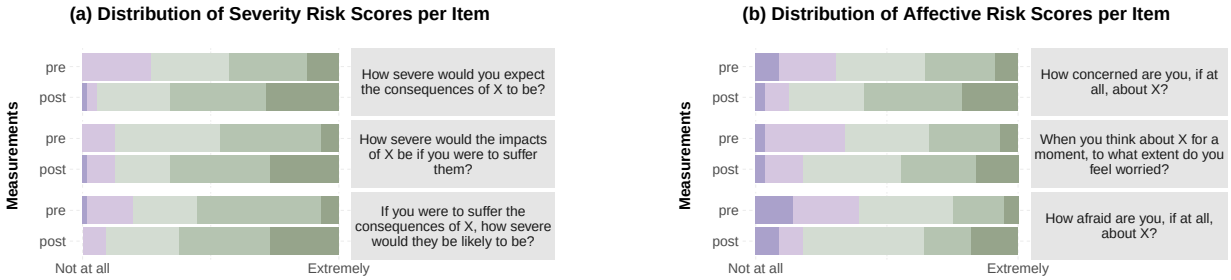


Figure 6: Plots illustrating the pre post distribution of each (a) Severity Item; (b) Affect Item.

"when I imagine myself in that situation. I think I would be the same with that face" (P05)), which heightened their personal connection to the consequences depicted (e.g. *"the visual impact is big for me. So oh my God, I will look like that"* (P20)). This mirroring effect was heightened by participants' tendency to ascribe emotional states to the agent, which allowed them to envision the risks as though they were personally experiencing them (e.g. *"you're actually feeling like the robot is suffering. So maybe I, in this situation, will be suffering too"* (P16)). One participant noted that the agent offers a more generalised face, not a specific person's face, which facilitates self-projection, *"[an image of a] real person might not have the same impact because this person is a specific person. I know that they may eat unhealthy, it's their own decision. But here, it is generalised so it could be me"* (P20). Secondly, several participants perceived the physical agent as a facilitator for feelings of concern and empathy towards others (e.g. *"I was able to empathise with family members or older groups of people who would experience this"* (P26)), and noted an increased awareness of potential people who might experience the consequence (*"people in really bad environmental conditions are really suffering"* (P10)).

Self-reflection. Previous research shows the relevance of negative emotions to foster contemplative thoughts and self-reflection [37]. Viewing the visualisations made most participants reflect upon past experiences related to the conveyed topic. Facing the visualised negative consequences of PM2.5 made several participants reflect upon their own up-bringing in areas with high air pollution or visits to areas impacted by air pollution. Interestingly, some participants reported that they were re-evaluating their past experiences, stating that *"I never, put on a mask or anything. But I feel like I really should"* (P04), now aware of the risk. Few participants reflected upon covid-19, triggered by viewing the facial masks (e.g. P20, P05, P27). A few others reported that it made them contemplate about the environment in general.

Facing the negative consequences of LDL cholesterol, several participants reflected upon the health of family members and friends (e.g. *"my parents looked very exhausted even after working for a short time because they have high cholesterol. Well, now I could relate to that"* (P02)). Other participants reflected upon their own health journey (e.g. *"I'm overweight. So I was thinking about myself"* (P28)). Moreover, several participants reflected on their current behaviour and its impact; e.g. *"it let me think about my consumption of junk food"* (P18).

4.2.4 Agent Perception. Agents can act upon their environment, based on their ability to perceive [56]. In line with the findings from our design workshop, participants experienced the physical agent as responding to either internal or external changes (e.g. *"as the environment changes, external environment or internal, the robot was reacting to that"* (P13)). Few participants indicated perceived intentional agency [26], describing the agent's intent to send a risk message (e.g. *"it was a direct message, ... I was feeling [as if] getting a message from the robot"* (P28)).

Humanlikeness, Realism and Vividness. We identified three qualities contributing to participants' experience: humanlikeness, realism, and vividness of the physical agent. Most participants highlighted the humanlikeness of the agent which helped them relate and build a connection with the agent and the conveyed information (e.g. *"it is behaving like a human. And I'm also human." (P13)*), going as far as to describe the agent as a human itself; e.g. *"it helps*

me to empathise with the situation, seeing another kind of human go through it" (P07). Solely few people were hesitant, describing the agent as somewhat humanlike, or expressing an uncanney valley effect that needed to be overcome (e.g. *"at first it was a bit weird, but when you get used to it, it feels pretty natural"* (P11)). Second, several participants highlighted the realism of the visualisation conveyed through the agent which increased the perceived authenticity (e.g. *"the authenticity is quite real to me"* (P14)). Besides its realism, several participants explained that the liveliness of the agent (e.g. *"it's quite vivid"* (P03)) brought the information and inherent risk to life (e.g. *"I guess it brings it to life. Like what if it was me? What if it was my family? ... it humanises the problem"* (P25)). Notably, eye contact with the agent was highlighted by few participants as a key element for vividness (e.g. *"the eyes they kept moving like real people"* (P26)) and to foster an emotional connection (e.g. *"the eye contact builds more of a connection"* (P25)).

While information acceptance was high across participants indicated by their increased risk perception, willingness to act-upon the risk and to take proactive measures, few participants indicated a certain degree of wariness towards the agent (e.g. *"I will be wary of the truth, whether they are actually representing the real information"* (P14)). Few participants raised the concern that they recognise the agent as a representation of others or as a conduit for presenting information about people, but not as having its own independent thoughts or intentions (e.g. *"It may represent other people in general, but the robot itself, I don't think it will have its own mind"* (P23)). Notably, several participants highlighted the potential to go beyond the agent's face, but to include e.g. sound (P23), dialogue features to tell a story about the hazard (P25) or the agent's body (e.g. P11, P22, P16) to enhance the visualisation experience.

4.2.5 Potential Future Applications. Our qualitative analysis shows that participants were eager to implement a physical humanlike agent to communicate risk in a multitude of contexts.

Public Spaces. The majority of participants mentioned the potential of the visualisation for education (e.g. *"in the classroom so we can have a better understanding of how it actually affects us, rather than reading a block of text"* (P03)) and healthcare (e.g. *"for the nutritionist or for the GP to demonstrate to their patients"* (P06)). Among the public spaces, participants emphasised parks, museums, grocery stores and transportation related applications, i.e. next to a bus stop, traffic light or in the metro station to inform and alert the community. Some participants envisioned such physical agents as tools for public health announcements and warnings, as an indicator for protective actions and for health crisis management (e.g. *"in the city, as warning, if you could combine that with an actual air quality measure and say this is the amount today and this is what you should be doing"* (P01)). Contrarily, one participant envisioned applications in the private space, such as the kitchen to support a healthy lifestyle or in the bedroom to convey the air quality in the morning (P23).

Accessibility. Participants often grounded their application ideas in the enhanced accessibility when using a physical humanlike agent as a visualisation interface for risk communication, which makes it particularly helpful for children, or lay people (e.g. *"they could read the facial expression much better than reading a complex science article"* (P19)) or people with literacy difficulties (e.g. *"in areas where not everyone can read or write"*(P28)). Conversely, few participants challenged the accessibility for the older generation (e.g. *"tech savvy people might find easier to use ... I wouldn't be sure about how people of the older generation would respond to it"* (P01)).

5 Discussion

5.1 Where Risk Communication meets Physical Agents

Our work introduces a novel and complementary approach for risk communication in HCI, which presents physical humanlike agents as effective risk communicators. While our design workshop offers important design considerations for how to leverage the face of a physical agent as a visualisation interface for risk communication (see Section 3.3), our user study demonstrates that a physical agent can bring health risks to life, supports self-projection and reflection, motivates protective actions, and captures interest not despite being provocative, but because of it. We discuss and contextualise our results to inform future research and provide takeaways for HCI. While findings from HRI, virtual avatars and conversational agents are difficult to compare, given differences in the degree of humanlikeness and fidelity between real-world robots, depicted robots on screens and virtual avatars, our aim is not to claim that physical humanlike agents are inherently superior to e.g. virtual agents for risk communication. Instead, our research unveils the unique potential that physical humanlike agents offer, especially with recent advancements in robot facial technology. We emphasise the timely relevance and importance of expanding risk communication research to include such agents to advance HCI research.

The agent brings risk to life. Effective risk communication hinges on the ability to convey potential consequences in a manner that resonates with individuals, to make risks personally relevant and to prompt action. On the topic of climate change, research has suggested that people often struggle to connect environmental risks with their implications for human health [39]. The results from our design workshop and user study show that the face of a physical humanlike agent can effectively convey risks directly tied to the human body, such as LDL cholesterol, and tied to external factors like air pollution. People's risk perception increases after engaging with the risk information conveyed through the agent, reflected in heightened perceptions of severity and affect. Leveraging **a physical humanlike agent as visualisation tool "humanises" the risk and brings information to life due to its high humanlikeness, realism and vividness.** It draws a direct connection from an abstract risk information to the human experience, by making abstract or invisible information more concrete and emotionally resonant through visual cues that mirror or link to human experiences. Our findings emphasise that a physical humanlike agent is more than a static display but perceived as an agent that actively reacts and responds to its environment and embodies information, which is e.g. highlighted by people's perception of mutual gaze in both the design workshop and user evaluation.

People envisioned the physical agent to be deployed in various contexts, from the classroom to the GP office to public parks and museums, which sheds light on the potential of physical agents to

communicate risk in-situ, directly within relevant contexts. Interestingly, recent advancements in HCI research introduce agent re-embodiment and co-embodiment [43], both of which present exciting directions for the future of risk communication. Re-embodiment, where an agent's social presence transfers across multiple physical forms, could enable a personal risk agent to "manifest" in various contexts. Co-embodiment, on the other hand, envisions multiple agent identities sharing a single physical body. This approach could allow a single physical agent to address diverse risks or tailor its communication to different audiences, an interesting avenue for future research.

This is how I would experience it. Alongside an increased risk perception, our results reveal that the visualisation supports self-projection, which **increases people's personal connection with the consequence displayed**. The ease with which people are able to imagine themselves experiencing a certain risk (imaginability) is an important determinant for feelings of risk, i.e. how vulnerable or at risk participants feel [25]. Our findings show that, instead of trying to identify the messenger, participants put themselves in the agent's shoes, and imagined how they would *look* and how it would *feel* when experiencing the same physical and emotional state as the agent. Notably, our work reveals a physical humanlike agent as a facilitator for self-projection, since it mimics human features but yet does not simulate a specific person. This balance creates a sense of familiarity and relatability while maintaining a neutral identity, which makes it a versatile tool for risk communication since it can resonate with a broad audience. It raises an intriguing question: How many shared features are necessary for viewers to successfully project themselves onto the agent? Understanding this 'sweet spot' could further enhance the ability to connect with diverse audiences, to ensure that risk communication is both impactful and inclusive.

Motivation to take action in response to risk. Several people expressed that the visualisation conveyed through the physical agent allowed them to picture the future, showing the consequences of (in)action. This idea of an agent acting as a bridge to the future-self served as an alarm bell, making participants more alert and aware of potential risks. In addition to people's tendency to self-project, the visualisation made people reflect and contemplate about past experiences such as childhood events, their own health journey and the health of family members and friends. This so-called self-referencing increases people's motivation to act upon risks [12, 14]. The finding is somewhat in contrast to findings by Ahn et al. [1] showing that virtual avatars and doppelgänger both increase perceived risk, but virtual doppelgängers were significantly more effective in promoting personal relevance to the health message, mediated by self-referent thought and self-presence.

People further expressed a high motivation to mitigate the risk, to take better care of their health (e.g. through diet or exercise) or to take actions to protect the environment as a means to protect their own health. The different types of facial masks were interpreted as protective actions and welcomed as a guide for better behaviour. The PM2.5 design thus not only conveyed the consequence of the risk, but also suggested what actions should be taken to mitigate these consequences. **This dual focus on consequence and mitigation made the risks relatable *and* actionable.**

Shocking, but interesting. Public health campaigns often use "threat appeals" to induce fear by implementing imagery and graphic displays of negative health consequences [8]. Common examples are smoking campaigns or campaigns to reduce driving speed on highways. When confronted with highly fear-inducing risk images, people's self-rated probability of experiencing the risk drops significantly [7]. This inverse and defensive response is caused by the fact that it is simply too stressful to engage with the risk at hand [7]. Interestingly, our results show the opposite. Seeing the consequences of both PM2.5 and LDL cholesterol "materialised" on a humanlike face was perceived as provoking and significantly increased affect (i.e. worry, fear, concern). Yet, most people emphasised that **while provoking, the visualisations nevertheless sparked their interest, supported their understanding and increased awareness.** This is in line with critical design research in HCI, "aimed at leveraging designs to make consumers more critical about their everyday lives" [2]. We take into account concern around provocation and re-emphasise our consideration from the design workshop, i.e. that **designers have to identify the delicate balance between provoking —to foster reflection, increase attention and challenge viewers to think— while ensuring not to cause emotional stress and harm.**

5.2 Empathy with Others and Data about People

I feel for others. Our visualisations acted as a facilitator for empathetic concern *for others* — from family members to other groups of people who might be experiencing the consequences of the depicted risks. The physical humanlike agent serves as a conduit for human experiences, which makes it a great opportunity to explore as example driven charts, a currently understudied design approach [48]. Example driven charts are visualisations that use a single person as a representation of many [48]. It is particularly interesting in the light of the "compassion fade effect" — where people's compassion and pro-social behaviour tend to decrease as the number of suffering individuals increases [69]. This approach has significant implications for risk communication, where statistics (X% people are suffering from ...) are often used to communicate risks. In addition, observing the experience of others who we feel close to, i.e. vicarious experiences, can be an important source of information and affect how we follow protective behaviour or perceive our vulnerability to a health threat [4, 13].

Our work also resonates with research on anthropographics, visualisations designed to evoke empathetic concern by representing people in the data [5]. While prior studies on anthropographics, such as human-shaped icons, resulted in mixed results regarding their impact on pro-social behaviour [5, 42, 47], our findings show the potential of physical humanlike agents as visualisation interfaces to convey data about people who may suffer health consequences or face risks. This opens a new avenue for exploring anthropographics, particularly in the context of physicality and realism, and supports calls for novel design strategies for anthropographics and their effects on prosocial outcomes [48].

5.3 Takeaways

We identified three key takeaways for the use of physical humanlike agents as visualisation interfaces to communicate risk consequences

and emotion, which we summarise next. Our implications guide future research in this space, pushing the boundaries of how we design human-centred risk communication in HCI.

- **Making risk relevant to human experience:** Our results demonstrate that the use of a physical humanlike agent as a risk communicator effectively humanises the conveyed risk and makes the negative consequences experienceable and relatable. It can be particularly valuable in contexts where understanding the impact of a certain risk on humans is crucial (e.g. climate change).
- **Suitability to communicate risk in public spaces:** Agent visualisations enhance risk perception, awareness, and understanding. This makes physical humanlike agents particularly suitable for public health campaigns and public spaces such as museums, where the goal is to inform the public about risks, raise awareness, and promote behavioural change.
- **Facilitator for empathetic responses:** By acting as conduits for human experiences, physical agents foster empathetic concern for others who may be affected by the conveyed risks. This is a critical outcome for risk communication when addressing collective challenges like health crises or climate change.

Our work lays the groundwork for using physical agents in effective health risk communication, to inform diverse health applications and future studies in HCI.

5.4 Limitations and Future Work

In this study, we used simulated levels of PM2.5 and LDL cholesterol rather than a real dataset. This approach was intentional, as the primary focus of our work was to explore the use of an agent's face as an interface for risk communication. While this allowed us to thoroughly investigate the potential of this novel approach, we see future opportunities for physical humanlike agents to convey data about real people and real-time data.

We acknowledge that the slider interface may have implicitly conveyed the data range through its endpoints, which may introduce bias by allowing participants to infer values without fully understanding the visualisation. However, we purposefully selected non-extreme values and asked participants to justify their interpretations in their own words. Data comprehension served primarily as a robustness check to support our understanding of participants' interpretations. Future work could explore alternative manipulation interfaces.

Moreover, we used the default setting of Furhat as the visualisation base. Dhawka et al. [11] highlight the need to include diverse human characteristics to better represent demographic groups. We strongly emphasise the potential of this approach to accomplish the latter and for future research to expand this work by elevating the representation of diverse groups through the highly adaptable, humanlike features. Further, we recognise that our work focuses on an agent's face, a deliberate design choice. However, there are several other design opportunities beyond the face. Future research could potentially include multisensory extensions to enhance the visualisation experience, e.g. voice to narrate data stories, or sound effects to create more immersive experiences.

Finally, we quantitatively examined the viewer's risk perception (severity and affect) as we visualised risk consequences. However, we acknowledge that other determinants of risk perception, e.g. perceived efficacy, controllability, or susceptibility [8, 70] exist. Our approach may not capture the full landscape of risk perception affected by our visualisation, but it presents a starting point for future investigations. In line, Schömbs et al. [60] investigated how to communicate probabilities using a robot's embodiment. Future works should examine how to use a physical humanlike agent as an interface to communicate hazard likelihood as the second key component of risk.

6 Conclusion

In this work we explore the use of a physical humanlike agent as a visualisation interface for effective risk communication, i.e. to communicate risk information and emotion. Through a design workshop, we collected design outcomes and insights from experts on how we can use an agent's face as a visualisation canvas for risk. We identified design strategies, challenges and benefits. We deployed visualisation designs depicting health risks on a real-world agent and investigated their effects in a mixed-method user study with 28 participants. Our findings show that using a physical humanlike agent as a visualisation interface for risk communication increases risk perception, perceived severity and effect. Moreover, it facilitates self-reflection, empathy and viewer's motivation to act. Participants also highlighted that this approach would be well-suited for campaigns in public spaces and to educate people on the consequences of important health-related risks.

Our work offers an important and novel approach to leverage physical humanlike agents as effective risk communicators in HCI. By doing so, our goal is not only to communicate negative consequences more effectively, but also to reduce the long-term health, social, and economic burdens on society by making risks relatable, easy to understand, and actionable.

Acknowledgments

We wish to acknowledge the contribution of the Melbourne Research Scholarship funded by the University of Melbourne and the Rowden White Scholarship established with a donation from Sir Alfred Edward Rowden White. This research is partially supported by the Australian Research Council Discovery Early Career Research Award (Grant No. DE210100858).

References

[1] Sun Joo (Grace) Ahn, Jesse Fox, and Jung Min Hahm. 2014. Using Virtual Doppelgängers to Increase Personal Relevance of Health Risk Communication. In *Intelligent Virtual Agents*, Timothy Bickmore, Stacy Marsella, and Candace Sidner (Eds.). Springer International Publishing, Cham, 1–12. doi:10.1007/978-3-319-09767-1_1

[2] Jeffrey Bardzell and Shaowen Bardzell. 2013. What is "critical" about critical design?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 3297–3306. doi:10.1145/2470654.2466451

[3] Lyn Bartram, Abhisekh Patra, and Maureen Stone. 2017. Affective Color in Visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1364–1374. doi:10.1145/3025453.3026041

[4] Seth A. Berkowitz, Robert A. Bell, Richard L. Kravitz, and Mitchell D. Feldman. 2012. Vicarious Experience Affects Patients' Treatment Preferences for Depression. *PLOS ONE* 7, 2 (Feb. 2012), e31269. doi:10.1371/journal.pone.0031269 Publisher: Public Library of Science.

[5] Jeremy Boy, Anshul Vikram Pandey, John Emerson, Margaret Satterthwaite, Oded Nov, and Enrico Bertini. 2017. Showing People Behind Data: Does Anthropomorphizing Visualizations Elicit More Empathy for Human Rights Data?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 5462–5474. doi:10.1145/3025453.3025512

[6] Virginia Braun, Victoria Clarke, Nikki Hayfield, Louise Davey, and Elizabeth Jenkinson. 2022. Doing Reflexive Thematic Analysis. In *Supporting Research in Counselling and Psychotherapy : Qualitative, Quantitative, and Mixed Methods Research*, Sofie Bager-Charleson and Alistair McBeath (Eds.). Springer International Publishing, Cham, 19–38. doi:10.1007/978-3-031-13942-0_2

[7] Stephen L. Brown and E. Zoe Smith. 2007. The inhibitory effect of a distressing anti-smoking message on risk perceptions in smokers. *Psychology & Health* 22, 3 (April 2007), 255–268. doi:10.1080/14768320600843127 Publisher: Routledge _eprint: https://doi.org/10.1080/14768320600843127.

[8] Rachel N. Carey and Kiran M. Sarma. 2016. Threat appeals in health communication: messages that elicit fear and enhance perceived efficacy positively impact on young male drivers. *BMC Public Health* 16, 1 (July 2016), 645. doi:10.1186/s12889-016-3227-2

[9] Jie Chen and Gerard Hoek. 2020. Long-term exposure to PM and all-cause and cause-specific mortality: A systematic review and meta-analysis. *Environment International* 143 (Oct. 2020), 105974. doi:10.1016/j.envint.2020.105974

[10] Carlos A. Cifuentes, Maria J. Pinto, Nathalia Céspedes, and Marcela Múnera. 2020. Social Robots in Therapy and Care. *Current Robotics Reports* 1, 3 (Sept. 2020), 59–74. doi:10.1007/s43154-020-00009-2

[11] Priya Dhawka, Helen Ai He, and Wesley Willett. 2023. We are the Data: Challenges and Opportunities for Creating Demographically Diverse Anthropographics. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3544548.3581086

[12] Sally Dunlop, Melanie Wakefield, and Yoshi Kashima. 2008. Can You Feel It? Negative Emotion, Risk, and Narrative in Health Communication. *Media Psychology* 11, 1 (March 2008), 52–75. doi:10.1080/15213260701853112 Publisher: Routledge _eprint: https://doi.org/10.1080/15213260701853112.

[13] Khaled Elazab, Mehmet Özden, Lemi Baruh, and Zeynep Cemalcılar. 2024. "Oh no, they caught it!": Vicarious experience of COVID-19, protection motivation and protective behaviors. *Journal of Health Psychology* 29, 6 (May 2024), 510–521. doi:10.1177/13591053231207166 Publisher: SAGE Publications Ltd.

[14] Jennifer Edson Escalas. 2007. Self-Referencing and Persuasion: Narrative Transportation versus Analytical Elaboration. *Journal of Consumer Research* 33, 4 (March 2007), 421–429. doi:10.1086/510216

[15] Diego R. Faria, Mario Vieira, Fernanda C.C. Faria, and Cristiano Premebida. 2017. Affective facial expressions recognition for human-robot interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 805–810. doi:10.1109/ROMAN.2017.8172395 ISSN: 1944-9437.

[16] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (May 2007), 175–191. doi:10.3758/BF03193146

[17] Melissa L. Finucane, Ali Alhakami, Paul Slovic, and Stephen M. Johnson. 2000. The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making* 13, 1 (2000), 1–17. doi:10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291099-0771%28200001/03%2913%3A1%3C1%3A%3AAID-BDM333%3E3.0.CO%3B2-S.

[18] Leopoldina Fortunati, Filippo Cavallo, and Mauro Sarrica. 2020. Multiple Communication Roles in Human–Robot Interactions in Public Space. *International Journal of Social Robotics* 12, 4 (Aug. 2020), 931–944. doi:10.1007/s12369-018-0509-0

[19] Mary Ellen Foster, Rachid Alami, Olli Gestranius, Oliver Lemon, Marketta Niemelä, Jean-Marc Odobez, and Amit Kumar Pandey. 2016. The MuMMER Project: Engaging Human-Robot Interaction in Real-World Public Spaces. In *Social Robotics*, Arvin Agah, John-John Cabibihan, Ayanna M. Howard, Miguel A. Salichs, and Hongsheng He (Eds.). Springer International Publishing, Cham, 753–763. doi:10.1007/978-3-319-47437-3_74

[20] Jesse Fox and Jeremy N. Bailenson. 2009. Virtual Self-Modeling: The Effects of Vicarious Reinforcement and Identification on Exercise Behaviors. *Media Psychology* 12, 1 (Feb. 2009), 1–25. doi:10.1080/15213260802669474

[21] Katerina Gorkovenko, Daniel J. Burnett, James K. Thorp, Daniel Richards, and Dave Murray-Rust. 2020. Exploring The Future of Data-Driven Product Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. doi:10.1145/3313831.3376560

[22] Institute for Health Metrics and Evaluation (IHME). 2024. *Global Burden of Disease 2021: Findings from the GBD 2021 Study*. Technical Report. Institute for Health Metrics and Evaluation (IHME), Seattle, WA. https://www.healthdata.org/research-analysis/library/global-burden-disease-2021-findings-gbd-2021-study

[23] Alexander Ivanov, Kurtis Danyluk, Christian Jacob, and Wesley Willett. 2019. A Walk Among the Data. *IEEE Computer Graphics and Applications* 39, 3 (May 2019), 19–28. doi:10.1109/MCG.2019.2898941 Conference Name: IEEE Computer Graphics and Applications.

[24] Alexander Ivanov, Kurtis Thorvald Danyluk, and Wesley Willett. 2018. Exploration & Anthropomorphism in Immersive Unit Visualizations. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3170427.3188544

[25] Eva Janssen, Liesbeth van Osch, Hein de Vries, and Lilian Lechner. 2013. The influence of narrative risk communication on feelings of cancer risk. *British Journal of Health Psychology* 18, 2 (2013), 407–419. doi:10.1111/j.2044-8287.2012.02098.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8287.2012.02098.x.

[26] Deborah G. Johnson and Mario Verdicchio. 2019. AI, agency and responsibility: the VW fraud case and beyond. *AI & SOCIETY* 34, 3 (Sept. 2019), 639–647. doi:10.1007/s00146-017-0781-9

[27] Ethan Kerzner, Sarah Goodwin, Jason Dykes, Sara Jones, and Miriah Meyer. 2019. A Framework for Creative Visualization-Opportunities Workshops. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 748–758. doi:10.1109/TVCG.2018.2865241 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[28] Rohit Ashok Khot, Deepti Aggarwal, Ryan Pennings, Larissa Hjorth, and Florian 'Floyd' Mueller. 2017. *EdiPulse*: Investigating a Playful Approach to Self-monitoring through 3D Printed Chocolate Treats. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 6593–6607. doi:10.1145/3025453.3025980

[29] Ho-Kyung Kim and Sei-Hill Kim. 2016. Understanding emotional bond between the creator and the avatar: Change in behavioral intentions to engage in alcohol-related traffic risk behaviors. *Computers in Human Behavior* 62 (Sept. 2016), 186–200. doi:10.1016/j.chb.2016.03.092

[30] Ki Joon Kim, Eunil Park, and S. Shyam Sundar. 2013. Caregiving role in human–robot interaction: A study of the mediating effects of perceived benefit and social presence. *Computers in Human Behavior* 29, 4 (July 2013), 1799–1806. doi:10.1016/j.chb.2013.02.009

[31] Simon Kimmel, Frederike Jung, Andrii Matviienko, Wilko Heuten, and Susanne Boll. 2023. Let's Face It: Influence of Facial Expressions on Social Presence in Collaborative Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3544548.3580707

[32] Kyveli Kompatsiari, Vadim Tikhanoff, Francesca Ciardo, Giorgio Metta, and Agnieszka Wykowska. 2017. The Importance of Mutual Gaze in Human-Robot Interaction. In *Social Robotics*, Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He (Eds.). Springer International Publishing, Cham, 443–452. doi:10.1007/978-3-319-70022-9_44

[33] Elly A. Konijn and Johan F. Hoorn. 2020. Differential Facial Articulacy in Robots and Humans Elicit Different Levels of Responsiveness, Empathy, and Projected Feelings. *Robotics* 9, 4 (Dec. 2020), 92. doi:10.3390/robotics9040092 Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[34] Stacey Kuznetsov, George Noel Davis, Eric Paulos, Mark D. Gross, and Jian Chiu Cheung. 2011. Red balloon, green balloon, sensors in the sky. In *Proceedings of the 13th international conference on Ubiquitous computing (UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 237–246. doi:10.1145/2030112.2030145

[35] Xingyu Lan, Yang Shi, Yanqiu Wu, Xiaohan Jiao, and Nan Cao. 2022. Kineticharts: Augmenting Affective Expressiveness of Charts in Data Stories with Animation Design. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 933–943. doi:10.1109/TVCG.2021.3114775

[36] Xingyu Lan, Yanqiu Wu, and Nan Cao. 2024. Affective Visualization Design: Leveraging the Emotional Impact of Data. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 1–11. doi:10.1109/TVCG.2023.3327385 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[37] Xingyu Lan, Yanqiu Wu, Yang Shi, Qing Chen, and Nan Cao. 2022. Negative Emotions, Positive Outcomes? Exploring the Communication of Negativity in Serious Data Stories. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–14. doi:10.1145/3491102.3517530

[38] Nancy Lankton, D. McKnight, and John Tripp. 2015. Technology, Humanness, and Trust: Rethinking Trust in Technology. *Journal of the Association for Information Systems* 16, 10 (Oct. 2015). doi:10.17705/1jais.00411

[39] Anthony A. Leiserowitz. 2005. American Risk Perceptions: Is Climate Change Dangerous? *Risk Analysis* 25, 6 (Dec. 2005), 1433–1442. doi:10.1111/j.1540-6261.2005.00690.x

[40] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (May 2015), 23–37. doi:10.1016/j.ijhcs.2015.01.001

[41] Shan Li and Weihong Deng. 2022. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* 13, 3 (July 2022), 1195–1215. doi:10.1109/TAFFC.2020.2981446 Conference Name: IEEE Transactions on Affective Computing.

[42] J. Liem, C. Perin, and J. Wood. 2020. Structure and Empathy in Visual Data Storytelling: Evaluating their Influence on Attitude. *Computer Graphics Forum* 39, 3 (2020), 277–289. doi:10.1111/cgf.13980 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13980.

[43] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. ACM, San Diego CA USA, 633–644. doi:10.1145/3322276.3322340

[44] Joseph A. Maxwell. 2010. Using Numbers in Qualitative Research. *Qualitative Inquiry* 16, 6 (July 2010), 475–482. doi:10.1177/1077800410364740

[45] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10, 1 (Jan. 2019), 18–31. doi:10.1109/TAFFC.2017.2740923 Conference Name: IEEE Transactions on Affective Computing.

[46] L. Morais, N. Andrade, and D. Sousa. 2022. Exploring How Visualization Design and Situatedness Evoke Compassion in the Wild. *Computer Graphics Forum* 41, 3 (2022), 441–452. doi:10.1111/cgf.14553 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14553.

[47] Luiz Morais, Yvonne Jansen, Nazareno Andrade, and Pierre Dragicevic. 2021. Can Anthropographics Promote Prosociality?A Review and Large-Sample Study. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3411764.3445637

[48] Luiz Morais, Yvonne Jansen, Nazareno Andrade, and Pierre Dragicevic. 2022. Showing Data About People: A Design Space of Anthropographics. *IEEE Transactions on Visualization and Computer Graphics* 28, 3 (March 2022), 1661–1679. doi:10.1109/TVCG.2020.3023013 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[49] Dario Pasquali, Austin Kothig, Alexander Mois Aroyo, John Edison Muñoz Cadorna, Kerstin Dautenhahn, Stefano Bencetti, Rea Francesco, and Alessandra Sciutti. 2023. That's not a Good Idea: A Robot Changes Your Behavior Against Social Engineering. In *International Conference on Human-Agent Interaction*. ACM, Gothenburg Sweden, 63–71. doi:10.1145/3623809.3623879

[50] Paul Robinette, Ayanna Howard, and Alan R. Wagner. 2017. Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed? In *Autonomy and Artificial Intelligence: A Threat or Savior?*, W.F. Lawless, Ranjeev Mittu, Donald Sofge, and Stephen Russell (Eds.). Springer International Publishing, Cham, 129–155. doi:10.1007/978-3-319-59719-5_6

[51] Furhat Robotics. [n. d.]. The Furhat Robot. https://furhatrobotics.com/furhat-robot/

[52] Sabine Roeser. 2012. Risk Communication, Public Engagement, and Climate Change: A Role for Emotions. *Risk Analysis* 32, 6 (2012), 1033–1040. doi:10.1111/j.1539-6924.2012.01812.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.2012.01812.x.

[53] E. Roesler, D. Manzey, and L. Onnasch. 2021. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics* 6, 58 (Sept. 2021), eabj5425. doi:10.1126/scirobotics.abj5425 Publisher: American Association for the Advancement of Science.

[54] W. D. Rowe. 1975. *An "anatomy" of Risk*. Environmental Protection Agency. Google-Books-ID: O9JRAQAAMAAJ.

[55] Robert A.C. Ruiter, Charles Abraham, and Gerjo Kok. 2001. Scary warnings and rational precautions: A review of the psychology of fear appeals. *Psychology & Health* 16, 6 (Nov. 2001), 613–630. doi:10.1080/08870440108405863 Publisher: Routledge _eprint: https://doi.org/10.1080/08870440108405863.

[56] Stuart J. Russell and Peter Norvig. 2016. *Artificial intelligence : a modern approach*. Pearson.

[57] Shaowei Sang, Chong Chu, Tongchao Zhang, Hui Chen, and Xiaorong Yang. 2022. The global burden of disease attributable to ambient fine particulate matter in 204 countries and territories, 1990-2019: A systematic analysis of the Global Burden of Disease Study 2019. *Ecotoxicology and Environmental Safety* 238 (June 2022), 113588. doi:10.1016/j.ecoenv.2022.113588

[58] Sarah Schömbs, Jacobe Klein, and Eileen Roesler. 2023. Feeling with a robot—the role of anthropomorphism by design and the tendency to anthropomorphize in human-robot interaction. *Frontiers in Robotics and AI* 10 (2023). https://www.frontiersin.org/articles/10.3389/frobt.2023.1149601

[59] Sarah Schömbs, Jiahe Pan, Yan Zhang, Jorge Goncalves, and Wafa Johal. 2024. FaceVis: Exploring a Robot's Face for Affective Visualisation Design. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3613905.3650910

[60] Sarah Schömbs, Saumya Pareek, Jorge Goncalves, and Wafa Johal. 2024. Robot-Assisted Decision-Making: Unveiling the Role of Uncertainty Visualisation and Embodiment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3613904.3642911

[61] Stela H. Seo, Denise Geiskkovitch, Masayuki Nakane, Corey King, and James E. Young. 2015. Poor Thing! Would You Feel Sorry for a Simulated Robot?: A comparison of empathy toward a physical and a simulated robot. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Portland Oregon USA, 125–132. doi:10.1145/2696454.2696471

[62] Paul Slovic, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor. 2004. Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality. *Risk Analysis* 24, 2 (2004), 311–322. doi:10.1111/j.0272-4332.2004.00433.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0272-4332.2004.00433.x.

[63] Mariëlle Stel, Rick B. van Baaren, and Roos Vonk. 2008. Effects of mimicking: acting prosocially by being emotionally moved. *European Journal of Social Psychology* 38, 6 (2008), 965–976. doi:10.1002/ejsp.472 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.472.

[64] Mariëlle Stel and Roos Vonk. 2010. Mimicry in social interaction: Benefits for mimickers, mimickees, and their interaction. *British Journal of Psychology* 101, 2 (2010), 311–323. doi:10.1348/000712609X465424 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1348/000712609X465424.

[65] Gareth Terry and Nikki Hayfield. 2021. Data analysis: Familiarization and coding. In *Essentials of thematic analysis*. American Psychological Association, Washington, DC, US, 29–41. doi:10.1037/0000238-003

[66] Prakash Thangavel, Duckshin Park, and Young-Chul Lee. 2022. Recent Insights into Particulate Matter (PM2.5)-Mediated Toxicity in Humans: An Overview. *International Journal of Environmental Research and Public Health* 19, 12 (June 2022), 7511. doi:10.3390/ijerph19127511

[67] Mikel Val-Calvo, José Ramón Álvarez Sánchez, José Manuel Ferrández-Vicente, and Eduardo Fernández. 2020. Affective Robot Story-Telling Human-Robot Interaction: Exploratory Real-Time Emotion Estimation Analysis Using Facial Expressions and Physiological Signals. *IEEE Access* 8 (2020), 134051–134066. doi:10.1109/ACCESS.2020.3007109 Conference Name: IEEE Access.

[68] V.H.M. Visschers, P.M. Wiedemann, H. Gutscher, S. Kurzenhäuser, R. Seidl, C.G. Jardine, and D.R.M. Timmermans. 2012. Affect-inducing risk communication: current knowledge and future directions. *Journal of Risk Research* 15, 3 (March 2012), 257–271. doi:10.1080/13669877.2011.634521 Publisher: Routledge _eprint: https://doi.org/10.1080/13669877.2011.634521.

[69] Daniel Västfjäll, Paul Slovic, Marcus Mayorga, and Ellen Peters. 2014. Compassion Fade: Affect and Charity Are Greatest for a Single Child in Need. *PLOS ONE* 9, 6 (June 2014), e100115. doi:10.1371/journal.pone.0100115 Publisher: Public Library of Science.

[70] Hugh D Walpole and Robyn S Wilson. 2021. Extending a broadly applicable measure of risk perception: the case for susceptibility. *Journal of Risk Research* 24, 2 (Feb. 2021), 135–147. doi:10.1080/13669877.2020.1749874 Publisher: Routledge _eprint: https://doi.org/10.1080/13669877.2020.1749874.

[71] Hugh D. Walpole and Robyn S. Wilson. 2021. A Yardstick for Danger: Developing a Flexible and Sensitive Measure of Risk Perception. *Risk Analysis* 41, 11 (Nov. 2021), 2031–2045. doi:10.1111/risa.13704

[72] Hugh D. Walpole and Robyn S. Wilson. 2022. Both analysis and feelings? The influence of risk beliefs on holistic risk judgments through dual systems using the ESSA model. *Journal of Risk Research* 25, 1 (Jan. 2022), 1–20. doi:10.1080/13669877.2021.1907610 Publisher: Routledge _eprint: https://doi.org/10.1080/13669877.2021.1907610.

[73] Adam Waytz, John Cacioppo, and Nicholas Epley. 2010. Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science* 5, 3 (May 2010), 219–232. doi:10.1177/1745691610369336 Publisher: SAGE Publications Inc.

[74] Joel Wester, Bhakti Moghe, Katie Winkle, and Niels van Berkel. 2024. Facing LLMs: Robot Communication Styles in Mediating Health Information between Parents and Young Adults. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (Nov. 2024), 497:1–497:37. doi:10.1145/3687036

[75] Robyn S. Wilson, Adam Zwickle, and Hugh Walpole. 2019. Developing a Broadly Applicable Measure of Risk Perception. *Risk Analysis* 39, 4 (2019), 777–791. doi:10.1111/risa.13207 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.13207.

[76] World Health Organization. 2024. Ambient (outdoor) air pollution. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[77] Bin Yu, Mathias Funk, Jun Hu, and Loe Feijs. 2017. StressTree: A Metaphorical Visualization for Biofeedback-assisted Stress Management. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, Edinburgh United Kingdom, 333–337. doi:10.1145/3064663.3064729

[78] Jack Zhao and Andrew Vande Moere. 2008. Embodiment in data sculpture: a model of the physical visualization of information. In *Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts*. ACM, Athens Greece, 343–350. doi:10.1145/1413634.1413696

[79] Jia Zheng, Jing Wang, Yan Zhang, Jiangliu Xia, Huilan Guo, Haiying Hu, Pengfei Shan, and Tianlang Li. 2022. The Global Burden of Diseases attributed to high low-density lipoprotein cholesterol from 1990 to 2019. *Frontiers in Public Health* 10 (Aug. 2022). doi:10.3389/fpubh.2022.891929 Publisher: Frontiers.

A Semi-structured Interview: Design Workshop

- "Please pick and quickly present your favourite visualisations for each category."
- **Opening Question:**
  - "Which visualisation impressed you most and why?"
- **Design Process and Experience:**
  - "How did you proceed to map data onto a robot's face or encode data?"
- **Key Questions:**
  - "Please describe challenges you have faced during your design process, e.g. design constraints when using a robot to visualise affective data."
  - "Could you please elaborate on why you think a robot's face could be useful or not useful in conveying data affectively, and if you can think of potential pitfalls?"
  - "We gave you one specific robot—what did you miss to better convey data, and what else could the robot be doing to visualise data?"

B Semi-structured Interview: User Study Evaluation

- **Data Exploration:**
  - "Could you please walk me through your exploration?"
  - "What did you observe as you adjusted the slider to different values?"
- **Risk Perception:**
  - "Can you explain to me the risks associated with the data shown in the two scenarios?"
  - "Did viewing the visualisations change how you perceived the risk of each topic? If so, how?"
  - "In what ways did using a robot's face for the visualisation help or hinder your understanding and perception of the risk?"
- **Affective Responses:**
  - "What did you feel when viewing these visualisations?"
  - "Did the visualisations make you imagine yourself in the robot's position or the position of someone affected by the data presented? If so, how?"
  - "Can you describe that experience?"
- **Agency and Anthropomorphism:**
  - "How did the human-like features of the robot influence your emotional connection to the data and the topic?"
- **Self-Reflection and Introspection:**
  - "Did viewing the visualisation make you reflect on your own life and experiences? If so, how?"
- **Motivation to Act:**
  - "Do you think that the visualisation motivates you to take any action regarding particulate matter or cholesterol?"
  - "If so, what actions do you feel inspired to take?"
- **Application Context:**
  - "Can you think of potential applications or use cases for robot's visualising data?"
- **Ending Question:**
  - "What did you like or dislike about the visualisation?"
  - "Is there anything else you would like to add or mention?"

5.3 Chapter Reflection

This chapter set out to ask how human-centredness can be realised through embodied AI systems, and addressed this question by employing anthropomorphism as the primary design mechanism, representing risk consequences through the robot's face via its visual appearance and expressiveness, as an *embodiment of risk*. Addressing RQ1, this chapter shows that the robot's humanlike form establishes the perceptual context that renders the encoded risk information interpretable as human experience, while the specific cues (including expressions, skin changes, facial elements) represent the actual risk information within that context. By encoding risk information through anthropomorphism, the robot's appearance served as a mirror, inviting participants to envision how the consequence may manifest for a human, and by extension, for themselves. Where Article II rendered risks personally relevant through linguistic description of the corresponding human experience, Article III achieved this through *risk embodiment*, conveying how consequences may look and feel. This *mirroring effect* fostered not only personal relevance but also self-projection and self-reflection. In doing so, Article III shows that anthropomorphism can bridge abstract quantitative risk and felt human experience in human-robot interaction. Quantitatively, our findings confirm the effectiveness of this design approach, as it enhances risk perceptions across the affect and severity dimension (RQ2), which strengthens the idea that anthropomorphism can be leveraged to accomplish a human-centred communication of risk consequences. Non-trivially, the chapter reveals that individuals' risk perceptions were often influenced by participants' lived experiences as a relevant individual factor (RQ3), potentially amplifying or desensitising people's risk perceptions, with these effects reinforced by the *risk mirroring* effect.

Addressing RQ4, this chapter distils concrete design recommendations, re-emphasising human-centredness as the core design consideration and showing that anthropomorphism enables *risk mirroring* as a concrete means of achieving human-centredness in embodied AI systems. These contributions are not without tensions, however, informing RQ4. The findings surface boundaries of anthropomorphic risk encoding, including constraints on information mapping, a tension between granularity and comprehensibility of the risk message, and considerations concerning the distressing nature of *risk mirroring*, reinforced through the dual mechanism of the anthropomorphic morphology as the medium and the expressive features as the information-carrying signals, which may induce avoidance responses. This Article also raises questions around how risks can be expressed beyond an anthropomorphic robot face, but by using an agent's body or other communicative modalities. In the next chapter, we explore how anthropomorphic and embodied cues can be used to communicate uncertainty as a constitutive component of risk and support a human-centred risk communication design in human-robot collaboration settings.

Chapter 6

Designing Human-Centred Uncertainty Communication in Robot-Assisted Decision-Making

6.1 Introduction

Similar to traditional AI-assisted decision tools investigated in Article I, robots are increasingly deployed as agents to inform user decisions due to their advanced capabilities in sensing, acting, and planning [162]. In human-AI interaction, the presentation of uncertainty cues rests on the assumption that they support users in forming more appropriate perceptions of AI systems and formulating more accurate decisions. This assumption, however, presupposes that users interpret uncertainty cues correctly, which is rarely the case in practice. People's understanding of numbers, particularly probabilistic values, is often shaped by cognitive biases [6, 127, 182, 201]. This chapter therefore investigates how uncertainty communication in narrow-AI, robot-assisted decision-making can be designed around users' interpretive needs while leveraging the embodied affordances of robotic systems, thereby extending the human-centred design desideratum established in the preceding chapters.

A well-established technique for supporting numeric understanding, particularly in health risk research, is the frequency format, which expresses probabilities as natural frequencies (X out of Z) rather than as percentages or decimals. This approach has proven effective to communicate probabilistic model uncertainty in human-AI interaction [32]. Notably, frequency formats are not limited to textual representations. In information visualisation research, uncertainty visualisation has long been considered a central research direction, and frequency formats find their visual representation in icon arrays as part-to-whole formats, offering an intuitive graphical representation of probability [3, 137, 227].

The preceding chapter demonstrated anthropomorphism as a means of embodying risk consequences to facilitate human-centredness. Building on this insight, the present chapter examines how uncertainty might similarly be represented through anthropomorphic cues in robot-assisted decision-making. Unlike the social robot platform investigated in Article III, robots deployed in collaborative industrial contexts are often not inherently anthropomorphic in their morphology, which raises the question of how to elicit anthropomorphisation and human-centredness through a system that does not inherently invite it. Expressive non-verbal robot behaviour has been shown to trigger anthropomorphic interpretations even in non-anthropomorphic systems and to elicit affective responses from human observers [173], provided that the behaviour is legible [49], that is, that it enables the observer to quickly infer the robot's communicative intent. Informed by HRI literature, this work investigates hesitation gesture as an embodied, anthropomorphic expression of uncer-

tainty [82], with the robot's velocity and waiting time as a function of the system's uncertainty, drawing on hesitation as a familiar human-human interaction through which uncertainty can be communicated in a more human-centred manner [130].

Extending the preceding chapter's exploration of risk embodiment to *uncertainty*, this chapter presents a mixed-methods in-person user study with $N = 36$ participants in which participants engaged in a collaborative medication assessment task with a robotic arm, contextualised in a manufacturing setting. Participants were tasked to discern which medication batches were defective and make subsequent shipping decisions, based on the robot's quality assessments. We manipulated the representation through which the robot communicates its uncertainty and the degree of uncertainty conveyed. For each medication assessment, the robot communicated instance-based uncertainty, operationalised as AI confidence [32], to support users in their decision-making process [29]. This uncertainty was represented either through an icon array displayed on a graphical interface attached to the robot's end effector or through hesitation gestures as an anthropomorphic communication of uncertainty.

Our empirical findings reveal no significant differences in the perceived system trustworthiness across the uncertainty representation conditions. However, the uncertainty representations informed by information visualisation prompted greater cognitive deliberation. Transparency evaluations differed significantly across uncertainty levels, with the robot being perceived as less transparent in high-uncertainty instances than in low-uncertainty instances. Our findings further point to an important trade-off between information granularity and heuristic interpretations of uncertainty, while also highlighting the limitations of explicitly communicating the absence of uncertainty.

The subsequent full article provides a detailed description of the methodology, technical implementation, results, and discussion, alongside derived implications for uncertainty communication in robot-assisted decision-making. It should be noted that within the article, uncertainty is expressed and referred to through its inverse, i.e. confidence.

6.2 Article IV

This article was presented at the CHI Conference on Human Factors in Computing (CHI 2025). Copyright is held by the authors. Publication rights licensed to ACM. This is the authors' version of the work. It is posted here for your personal use. Not for redistribution. The definitive version of record was published in:

Sarah Schömbs, Saumya Pareek, Jorge Goncalves, and Wafa Johal. 2024. Robot-Assisted Decision-Making: Unveiling the Role of Uncertainty Visualisation and Embodiment. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3613904.3642911

Ethics Application ID: 27391, the University of Melbourne Human Research Ethics Committee.

Robot-Assisted Decision-Making: Unveiling the Role of Uncertainty Visualisation and Embodiment

Sarah Schömbs
The University of Melbourne
Melbourne, Australia
sschombs@student.unimelb.edu.au

Saumya Pareek
The University of Melbourne
Melbourne, Australia
spareek@student.unimelb.edu.au

Jorge Goncalves
The University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

Wafa Johal
The University of Melbourne
Melbourne, Australia
wafa.johal@unimelb.edu.au



Figure 1: (left) The robot conveys low confidence in a pill box using a GUI before picking up the pill box. (right) The robot and the human pack their pill box into the package on the working station during a human-robot collaboration task.

ABSTRACT

Robots are embodied agents that act under several sources of uncertainty. When assisting humans in a collaborative task, robots need to communicate their uncertainty to help inform decisions. In this study, we examine the use of visualising a robot's uncertainty in a high-stakes assisted decision-making task. In particular, we explore how different modalities of uncertainty visualisations (graphical display vs. the robot's embodied behaviour) and confidence levels (low, high, 100%) conveyed by a robot affect the human decision-making and perception during a collaborative task. Our results show that these visualisations significantly impact how participants arrive to their decision as well as how they perceive the robot's transparency across the different confidence levels. We highlight potential trade-offs and offer implications for robot-assisted decision-making. Our work contributes empirical insights on how humans make use of uncertainty visualisations conveyed by a robot in a critical robot-assisted decision-making scenario.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642911

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

assisted decision-making, agency, uncertainty visualization, embodiment, risk communication

ACM Reference Format:
Sarah Schömbs, Saumya Pareek, Jorge Goncalves, and Wafa Johal. 2024. Robot-Assisted Decision-Making: Unveiling the Role of Uncertainty Visualisation and Embodiment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642911

1 INTRODUCTION

Robots are fueled by data. Data from their internal sensors to data collected by a thermal camera, lidar or proximity sensor. Data that allows a robot to work safely alongside humans, and humans to reason about a task and the robot itself [84]. Based on this data, robots can be used to make informed decisions and to accomplish tasks without sacrificing precision or experiencing fatigue. They can be deployed in environments that pose safety risks to humans or tasks that involve hazardous materials [40, 57, 87], while collecting data to support human decision-making. However, data is subject to uncertainty and thus robots have to act under several sources of uncertainty. From uncertainty caused by noisy sensors

to uncertainty based on training sets of neural networks to enable robots to detect objects [34, 55, 80]. The communication of such uncertainty holds the potential to enhance transparency, to calibrate trust and to support assisted decision-making [9, 69]. Further, visualising a robot's confidence can be a means to reduce risks and overcome communication barriers between humans and robots in collaborative tasks, which poses a current limitation [40, 54]. Both assisted decision-making [8, 19] and uncertainty [12, 29, 81] have long been investigated by the HCI community. And yet, visualising and conveying a robot's uncertainty in an assisted decision-making task has rarely been addressed so far.

Past research on robots in assisted decision-making has predominantly focused on the context of communication strategies [86], social settings [37, 83] or service domains, e.g. a robot giving recommendations for financial planning [85, 95]. In addition, several studies have investigated these so-called robot-advisors without humans interacting with a *physically present* robot, but by e.g. using video clips of robots or framing [68, 85, 86, 95]. However, the presence [24, 90], appearance [71] and embodiment [72] of a *physically present* robot can affect people's perception and experience. When the robot is physically present, it interacts in a shared environment and becomes *someone*, responsible for a certain task and outcome. In this paper, we investigate robot-assisted decision-making with a physically present and non-anthropomorphic robot, implemented in a collaborative task. In addition, we examine robot-assisted decision-making through the lens of uncertainty visualisation to inform how a robot's confidence affects human judgement and decision-behavior. We further look at how the robot's conveyed uncertainty affects the perceived trust in the robot, since "the relationship between uncertainty estimates and trust in automation is a relatively unexplored idea" [9].

We hereby draw from best-practices in both, Information Visualisation (VIS) and Human-Robot Interaction (HRI), which allows us to compare the effect of different visualisation modalities and confidence levels. In order to address the aforementioned research objectives, we explore robot-assisted decision-making with a robot that conveys its confidence through either a graphical interface (GUI) or a robot's behaviour, in the following referred to as EMBODIED uncertainty visualisation. A physically embodied and autonomous agent that moves and interacts with humans introduces new possibilities to map data onto an agent's behaviour, without the need to e.g. rely on complex information visualisations [77]. It could open up new ways to reach audiences with limited literacy skills, to interpret data in an intuitive manner and to support understandability, since past research in AI-assisted decision-making has shown that providing users with explicit explanations is a double-edged sword [8, 15, 19]. Besides, robots are not always equipped with screen interfaces to display information. Robot arms, in particular, are often screenless and need to rely on other ways to map data to support the human decision-making.

Our work is guided by the following research questions:

RQ1 How does visualising a robot's uncertainty affect the user's decision-making in a robot-assisted decision task?

RQ1.1 How do different visualisation modalities of a robot's uncertainty impact the ability to decide correctly in a high-stake scenario?

RQ1.2 Do users alter their decisions in the decision-making process when confronted with different uncertainty visualisations?

RQ2 How does a robot's visualised level of confidence affect the user's trust?

RQ2.1 How does a robot's visualised level of confidence influence users' perceived transparency of a robot?

RQ2.2 How does a robot's visualised level of confidence influence users' perceived performance of a robot?

We present a 3×2 mixed design lab study with 36 participants, contextualized in a human-robot collaboration setting which resembled a packing station in a factory that produces pills. We consider two types of uncertainty visualisation, a GUI and an EMBODIED uncertainty visualisation. Based on our manipulation check and prior research in VIS, we investigate a part-to-whole icon array as an established visualisation technique for uncertainty [3, 60, 65]. For the embodied visualisation, we draw from previous research on hesitation gestures and non-verbal signals to guide our design [36, 62, 88]. This allows us to explore how users perceive physically embodied visualisations differently compared to traditional graphical displays and how they affect user's decision-making. Besides the type of uncertainty visualisation, we consider the degree of confidence as a within-subject factor (low, high, 100%) to make inferences about the human's understanding and perception of the magnitude of uncertainty that is visualised by the robot [60].

We summarize our key contributions as follows:

- Looking at HRI from a VIS perspective has recently been identified as a major gap in HRI research [84]. We bridge that gap by exploring uncertainty visualisations in *in-situ* robot-assisted decision-making, using a physically present robot and a real-world scenario.
- We gain empirical insights into people's understanding, perception and experience with a robot conveying its uncertainty and emphasise non-trivial trade-offs.
- We highlight the implications of our findings and propose fundamental considerations when visualising uncertainty in a robot-assisted decision-making task. Furthermore, we outline potential directions for future research to broaden the scope of uncertainty visualizations in the context of human-robot collaboration.

2 RELATED WORK

In this section, we explore relevant literature from AI-assisted decision-making, visualising uncertainty and behaviour based signals in HRI. We present an interdisciplinary review drawn from literature in HCI, HRI, Machine Learning and VIS. As a guideline, we define uncertainty as "the difference between observation and reality", according to Chung and Wark [21]'s categories of uncertainty for decision-support. However, we acknowledge the existence of other uncertainty types and definitions, varying form epistemological to stochastic uncertainties, which easily illustrates the complexity of the topic itself. Further, we define decision-making as a choice between two competing courses of action [7].

2.1 AI-assisted Decision Making and Robot-Advisors

AI-assisted decision-making has gained much attention based on its potential to integrate strengths from both parties - AI and human - to enhance human-AI team performance [17]. The latter becomes particularly important in critical scenarios that involve high risks, e.g. cancer diagnosis [28, 92], frequently referred to as high-stakes scenarios [8, 17, 93]. These scenarios often require high accuracy and are usually bounded by human traits like fatigue, biases, lack of attention during repetitive tasks, or lack of precision. Importantly, AI-assisted decision support can reduce users' cognitive load, the time to make an informed decision and the need to acquire mathematical expertise [69]. As a result, AI can be implemented as a decision-support tool to provide recommendations for human decision-making. Moreover, AI-assisted decision support offers an opportunity to assist lay decision-makers to make informed decisions even in high-stakes scenarios [69]. However, AI-based systems are not perfect, which is why it is important to keep humans in the loop and to avoid over-reliance, especially in critical task domains [15, 17, 69].

Interestingly, Prabhudesai et al. [69] highlight the need to make uncertainty explicit in AI-assisted decision support tools. The authors argue that withholding uncertainty oversimplifies the reality, and paints the wrong picture of a perfect system. Similarly, Bhatt et al. [9] stresses the importance of communicating uncertainty to enhance transparency and to support an appropriate formation of trust between the agent and the human. Making uncertainty explicit in a decision-making scenario has a long tradition in VIS, since "ignoring uncertainty and its associated risk may simplify the decision-making process, but (...) does not result in making informed decisions" [23].

Similar to AI-assisted decision-making tools, robots can also be implemented in high-stakes scenarios to support humans. They are increasingly used in healthcare, e.g. as surgery assistance [97] or caregivers [18], and industrial settings, e.g. as inspection robots or for assembly related task [30, 40]. Notably, research on robots as decision-support has predominantly centered around anthropomorphic robots, particularly in social settings and service domains, e.g. robots that give financial advice - often referred to as robot-advisors [37, 68, 86].

2.2 Graphical Uncertainty Visualisation

Visualising uncertainty is a domain-overlapping issue that has been investigated in domains varying from hazard and weather forecasting [10, 65], to map-based decision-making during helicopter landings [50] to everyday contexts like catching the next bus [46]. It aims to better communicate risk, support decision-making and to represent information as accurately as possible. However, visualising uncertainty does not follow a one-rule-applies-approach [64]. There are no universal guidelines that describes when to choose which kind of uncertainty visualisation for a given scenario. Rather, it is highly dependent on the context, the task itself, the data involved and what kind of uncertainty you aim to visualise. However, some visualisation methods are more suitable than others and have been established as best-practices in their domain.

A number of surveys provide a review of methodologies and classifications to support an adequate selection of the respective uncertainty visualisation [14, 39, 64]. In the context of decision-making, visualising uncertainty is a central and widely investigated concern. The way uncertainties are conveyed can impact how individuals interpret them and, consequently, the actions they may take. Uncertainty visualisations potentially open the door to increase users' understanding of the underlying data, to support trust [45, 47], to inform about the instability inherent to estimates [67], or to reduce biases [74]. However, visualising uncertainty can also present challenges such as a feeling of being overwhelmed or confusion [29, 46].

2.2.1 Visualising Risk and Probabilities. For the purpose of this study, we take a closer look at graphical techniques to visualise single probabilistic values. Probabilities can be displayed in a part-to-whole relationship, which conveys the ratio of a numerator, to its total, the denominator. Past research provides evidence that people tend to process part-to-whole relationships easier and faster compared to visualisations that solely display the numerator [3, 65]. A few examples of such visualisations include icon arrays, bar graphs, and grids.

Previous work has recognized the potential of icon arrays to communicate probabilities, particularly in risk communication and shared decision making environments [3, 64, 75, 98]. Icon arrays display the probability as a group of dots, figures or other icons, with the numerator and denominator often differing in transparency or colour. Interestingly, past research indicates that concrete icon types, particularly anthropomorphic ones e.g. person icons, tend to perform better compared to abstract icons such as rectangular blocks [98]. The symbolism of icon arrays can add contextual meaning to the numeric information and thus promote an intuitive understanding of the data. However, anthropomorphic icon arrays can also result in risk magnitude biases [75]. Interestingly, studies that investigated both icon arrays and bar graphs revealed contradicting findings with no clear preference for one or the other [60, 75]. Using metaphors or analogies to visualise risk can be particularly helpful to make numeric values accessible to different audiences. Past research has shown that a gauge is easy to understand and can be used to e.g. readout levels of chronic stress [4, 5]. Whereas gauge visualisations are well-suited for providing a general approximation, they are less suited for precise calculations of specific values [76].

Besides choosing an adequate uncertainty visualisation, the proportion or degree of uncertainty does also matter when it comes to how uncertainty is perceived and processed. McCaffery et al. [60] show that the numerator size of an icon array, hence risk size, affects whether or not an uncertainty visualisation supports accurate estimations in data tasks. Contrary to their assumptions, Sarma et al. [74] found that the proportion of missing values (either 30% or 50%) in datasets, hence the proportion of uncertainty, does affect people's biases depending on the data task they perform. In terms of probability values, findings by Jiang et al. [44] indicate that participants started to trust a system's recommendation at a conveyed confidence level of more than 75%, which implies an uncertainty of less than 25%. In addition, small probabilities are particularly difficult to interpret [3, 51].

2.3 Embodied Uncertainty Visualisation

A recent paper by Leusmann et al. [54] proposes the so-called *Human-Robot Uncertainty Loop*, a framework which identifies various forms of uncertainties in HRI and their need to be communicated. To depict uncertainty in HRI, a multitude of different approaches open up. From traditional graphical user interfaces to XAR to using the robot's physicality as an interface to convey data. For the purpose of this paper, we take a closer look at embodied, behaviour-based uncertainty signals. Whereas traditional VIS research has focused on explicit uncertainty visualisations [20], a robot's embodiment offers new ways to visualise uncertainty implicitly. Embodied visualisations enable a robot to display it's internal world without the need to use a graphical display or speech. This is especially important in industrial settings, in which (a) robotic arms are often not equipped with graphical displays, or (b) the environment is too noisy to implement speech responses. Besides, embodied uncertainty visualisations could potentially improve accessibility and facilitate peoples' intuitive understanding, similar to data physicalizations [6, 41]. Inspired by previous research on visualising a robot's motion intent [53, 66] and expressive motion behavior [26], we look at hesitation gestures to visualise a robot's uncertainty in an assisted-decision-making task.

2.3.1 Hesitation Gestures. Hesitation, as a known uncertainty signal from human-human interaction, has received much attention in the HRI community. As one of the first attempts to implement and investigate hesitation gestures, Moon et al. [62] examined human-human hesitation gestures and transferred the respective trajectories onto a robotic arm. Their findings confirm that participants were able to correctly infer and detect a robot's hesitation. Building upon their work, the authors recently presented an improved trajectory generator to implement hesitation gestures [61]. However, it is important to highlight that their research focuses on uncertainty in human-robot negotiations, i.e. the human and the robot reaching for the same object at the same time. Therefore, Moon et al. [63] define hesitation gestures as gestures, that express uncertainty about the "right or priority" to access a shared resource. Besides these efforts, past research also investigated pausing [96], trembling and repeated pull-back movements as cues for hesitation [88]. Importantly, a robot needs to be able to convey its degree of uncertainty. To do so, Hough and Schlangen [36] successfully developed a model to display a robot's confidence level, as in inverse to uncertainty, through motion-based behaviour. Here, the robot adjusts its speed and waiting time according to its internal confidence.

2.4 Summary

As human-robot collaboration continues to expand, the integration of uncertainty visualization in the decision-making process holds the potential to enhance transparency, performance and trust. In contrast to AI-assisted decision-support or framed robot-advisors, a *physically present* robot presents an entity which can be perceived as an agent itself - tangible and experienceable. Taking into account such Data-Agent Interplay [77], we aim to examine not just best-practices from VIS in the context of robot-assisted decision-making [84], but to open the design space of uncertainty visualisations by exploring embodied visualisations conveyed by a robot and how humans understand the aforementioned.

3 METHOD

We conducted a 3×2 mixed design lab study with *level of confidence* (low confidence vs. high confidence vs. 100% confidence) as a within-subjects factor and *type of visualisation* (graphical user interface vs. physically embodied) as a between-subjects factor. We thus manipulated the modality through which the robot conveyed its confidence and its confidence in the quality of the assessed pill box itself. This allows us to investigate the relationship between the type of visualisation and the level of confidence, their effect on the user's judgement, decision-behavior and perceived trust in the robot in the context of a critical, high-stakes scenario, i.e. shipping defective pills.

3.1 Experimental Task

Our decision to conduct a study on assisted-decision making with autonomous agents in this context was based on two main reasons. First, with the rise of collaborative robots that are designed to work safely alongside humans, we are experiencing a paradigm shift from automation to actual collaboration [91]. Collaborative robots can be equipped with various sensors and hence be used to accomplish repetitive and hazardous tasks and to inform decisions [40, 87]. Second, we were particularly interested in investigating a critical scenario that involves high risks. As such, we selected a healthcare scenario, which is widely used in assisted decision-making literature due to the importance of effectively communicating risks [60]. In our study, participants were told that they would examine pills from a factory in which a mishap has happened due to which some pills might be defective.

To support their decision-making, a robot was implemented to assess the content of the pill boxes and to visualise its confidence: During the task, the robot was in charge of (a) displaying its level of confidence in sensing that the pill boxes were defective, and (b) packing those risky pill boxes into a package. We implemented these robot activities through an autonomous pick-and-place task implemented via ROS2 [59]. Before the robot placed its pill box into the package, the robot conveyed its confidence to inform the participant with one of the modalities (GUI or EMBODIED). The participant on the other hand was responsible to pack the pill boxes known to be not affected by the mishap (marked with a ticked label - see Figure 2). In addition, the participant was in charge to decide whether each package should be placed to the *ready for shipment* or *needs additional testing* area, taking into account the confidence the robot conveyed beforehand.

Our experimental task allowed us to realise two dependencies between the robot and the human to increase the degree of collaboration: knowledge, since the participant needed the robot's assessment to form their decision, and time, since the task was performed sequentially, see Figure 2.

3.2 Experimental Manipulations

In order to answer our proposed research questions, we manipulated the robot's *type of visualisation* to convey its confidence and the *level of confidence* of the robot in the pill box not being defective. Next, we describe how we selected candidate visualisations and set confidence ranges.



Figure 2: Human-robot collaboration while packing pill boxes into packages.

3.2.1 Type of Visualisation. To investigate the potential effects of different types of uncertainty visualisations on the user's decision-making, perceived transparency and trust in the robot, we manipulated the way the robot conveyed its confidence. We hereby focused on two approaches. The first approach was to use the robot's physicality to embody and map the data onto the robot's movements to convey it's confidence. The robot used in this study was a robotic arm developed by Franka Emika [1]. This approach was inspired by past research efforts on hesitation gestures in HRI [36, 61–63]. Based on Hough and Schlangen [36], we implemented two functions to convey the robot's confidence through its behaviour. The first function describes the *waiting time*, i.e. how long the robot's gripper hovered above the target pill box. We implemented a waiting time inversely proportional to the robot's confidence. In this study, we set the maximum waiting time to 6 seconds with a degradation down to 0 (no waiting time in case of a 100% confidence). These bounds are scaled up from Hough and Schlangen [36] due to the size difference of the robot used in this study.

We additionally implemented a function that describes the time the robot needs after picking up the pill box to approaching the position to place the pill box into the package. This *travel time* was again inversely proportional to the robot's confidence with a maximum travel time of 10 seconds and a minimum travel time of 5 seconds, see exemplified values for both waiting and travel time for all three within factors in Table 1. In line with Hough and Schlangen [36], we assigned these bounds after a short pilot study and additionally validated their perception through a manipulation check (section 3.2.3).

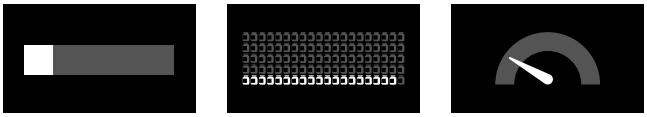
For the second approach we integrated a graphical user interface onto the robot's end-effector that enabled us to assess best practises from VIS during a human-robot collaborative scenario. In order to mount a digital screen onto the robot we designed and 3D printed a case to clip an iPhone 13 mini on to the robot, see Figure 1. The iPhone's screen was used to display the GUI on the browser. We developed a Flask webapp to display the GUI on the phone and to communicate with the robot via ROS2 [59]. Grounded on previous research described in section 2.2.1 we selected a bar graph, a gauge visualisation and an icon array to convey probability values on the robot's graphical display, shown in Figure 3. We used a part-to-whole relationship for all three visualisation types [3, 65]. Thus, all

Table 1: Waiting time and travel time of the robot's movement in the embodied uncertainty visualisation condition with exemplified values for each confidence level.

		Waiting Time (in sec.)	Travel Time (in sec.)
Low Confidence	16%	5.04	9.2
	19%	4.86	9.05
	23%	4.62	8.85
High Confidence	90%	0.6	5.5
	92%	0.48	5.4
	95%	0.3	5.25
100%	100%	0	5

GUI visualisations displayed the robot's confidence in the format "numerator out of a constant denominator of 100".

For the icon array, we decided to display the numerator and denominator as icons of pill boxes to improve intuitiveness and accuracy, instead of using abstract symbols like blocks or dots [58, 98]. We refrained from using colour to enhance accessibility [42] and to avoid biases [11, 79] or cultural confounds [43]. To improve visibility, we used a black background with a white numerator and a dark-grey denominator, see Figure 3. In line with Jiang et al. [44] we did not show labels and numbers to avoid any effect from the number itself.



(a) Bar graph as part-to-whole visualisation displaying low confidence.

(b) Icon array as part-to-whole visualisation, with pill box icons displaying low confidence.

(c) Gauge as part-to-whole visualisation displaying low confidence.

Figure 3: Example of GUI visualisations for the low confidence condition (here, 19% confidence).

3.2.2 Level of Confidence. To gain additional insights on the importance of the actual level of uncertainty and to better inform future implications, we manipulated the level of confidence as a within-subjects factor. This allowed us to make inferences about how the participants' decision-making, perception of the robot and trust varies depending on the robot's conveyed confidence. We defined the different levels of confidence through the lens of risk communication in healthcare. It is important to mention that we framed uncertainty as the inverse of confidence, consistent with previous research [9, 36, 38, 69] and hence translated the respective uncertainty values to confidence values and confidence levels.

In line with McCaffery et al. [60], we defined low uncertainty as probabilities < 9.9% and assigned them to the *High Confidence* condition and high uncertainty as probability values in the range between 75% - 85% and assigned them to the *Low Confidence* condition. As can be seen, the perception of a small and high uncertainty

Table 2: Accuracy for each type of visualisation tested as manipulation check before the main user study.

	Confidence Level	100% Confidence			High Confidence			Low Confidence			Total Accuracy
	Probability Value	100%	100%	100%	98%	94%	91%	23%	19%	16%	
GUI	Bar Graph	0.7	0.8	0.8	0.2	0.8	0.9	0.9	0.9	0.9	0.77
	Icon Array	0.7	0.8	0.8	0.8	0.7	0.8	0.9	0.8	0.8	0.79
	Gauge	0.8	0.8	0.8	0.3	0.8	0.6	0.8	0.8	0.7	0.71
EMBODIED	Waiting Time	0.8	0.7	0.6	0.3	0.2	0.8	0.8	0.8	0.9	0.66
	Travel Time	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.51
	Waiting and Travel Time	0.6	0.5	0.4	0.6	0.5	0.5	1	1	0.9	0.67

is not a dichotomous phenomenon. We were additionally interested in investigating the perception of a fully confident robot (100% confidence). Since each within-subject factor consisted of a trial with a batch of 10 pill boxes that needed to be assessed and packed by the robot, we randomly assigned 10 confidence values out of the respective range to each confidence level to sample the respective probability range. We furthermore introduced noise to the *High Confidence* and the *Low Confidence* batch, inspired by Sarma et al. [74]: 20% of the time, the robot conveyed probability values of the opposite condition. We fixed this noise as the pill box #4 and pill box #7. We refrained from introducing noise to the *100% Confidence* condition, since it would damage the intended scheme of a fully confident robot.

3.2.3 Manipulation Check. We conducted a manipulation check to verify whether the type of visualisation and the level of confidence was perceived as intended. Past literature has shown that the performance of an icon array and a bar chart can depend on the numerator size, i.e. level of confidence [60]. The manipulation check was particularly important since we wanted to ensure that participants were able to detect and categorise the different levels of confidence in both visualisation conditions to make these conditions comparable. Furthermore, the manipulation check was used to select the most appropriate type of visualisation for each modality to later present to participant in our main lab study. To do so, we conducted a simple between-subjects manipulation check. Based on previous literature presented in section 2.2.1, we compared the accuracy of *icon array*, *bar chart*, and *gauge* visualisations for the graphical user interface and *waiting time*, *travel time*, *waiting and travel time* for the embodied visualisation.

We relied on independent raters and utilised the Prolific[1] crowdsourcing platform to recruit 60 participants who were based in the United States, native English speakers, to mitigate challenges with language proficiency [31] and who had approval ratings above 98%. All participants were compensated following prolific guidelines for payment [25]. Through the Qualtrics[2] survey platform, participants watched short video-clips of the robot (1.) assessing the content of a pill box, (2.) conveying its confidence based on the assigned visualisation type and (3.) placing the respective pill box into the package. We used videos of the same task setup of our in-person study. In a multiple choice question, participants were asked to categorise the level of confidence the robot conveyed. We measured the rater's accuracy to correctly categorise the conveyed confidence level. Table 2 outlines that the icon array has the highest accuracy. It emphasises that the icon array supports an accurate reading of the respective confidence level, which is why they are widely used as decision-aids, particularly in the field of healthcare [3, 32, 60]. It further shows that participants had difficulties correctly interpreting a 98% confidence in the gauge as well as the bar chart condition. Within the embodied visualisation types, the implementation of either waiting time or the combination of both waiting and travel time obtains a relatively high accuracy across the confidence levels. Since the total accuracy of waiting and travel time shows a slightly higher total accuracy score and has been shown to be correctly interpreted as a robot's confidence [36], we chose the latter as the embodied uncertainty visualisation for the in-person lab study.

3.3 Participants

Prior to the experiment, we conducted a power analysis to determine our sample size using *G*Power* [27]. The calculation was based on a within-between interaction with an alpha level of .05, a medium effect size of f = .25 and a power(1-β err prob) = .9, resulting in a targeted sample size of N = 36. We recruited 36 participants (25F, 11M) for our lab study using our university's notice board. This sample size is also in line with mixed-method studies previously published at CHI [16]. The study took roughly 45 minutes and participants were compensated with a $20 gift voucher. The study was approved by our University's Human Ethics Committee.

3.4 Study Procedure

The lab study consisted of five phases: (a) pre-questionnaire, (b) collaborative task and decision-making, (c) follow-up repeated measures, (d) procedure repeated for each confidence level and (e) open-ended questions, see Figure 4. Participants were randomly assigned an uncertainty visualisation type. Upon arrival to the usability lab, we provided each participant with a written plain language statement and collected their consent to participate in the study. After signing the consent form, the participant filled out a pre-survey to assess basic socio-demographics, and to gauge their previous

[1] www.prolific.com
[2] www.qualtrics.com

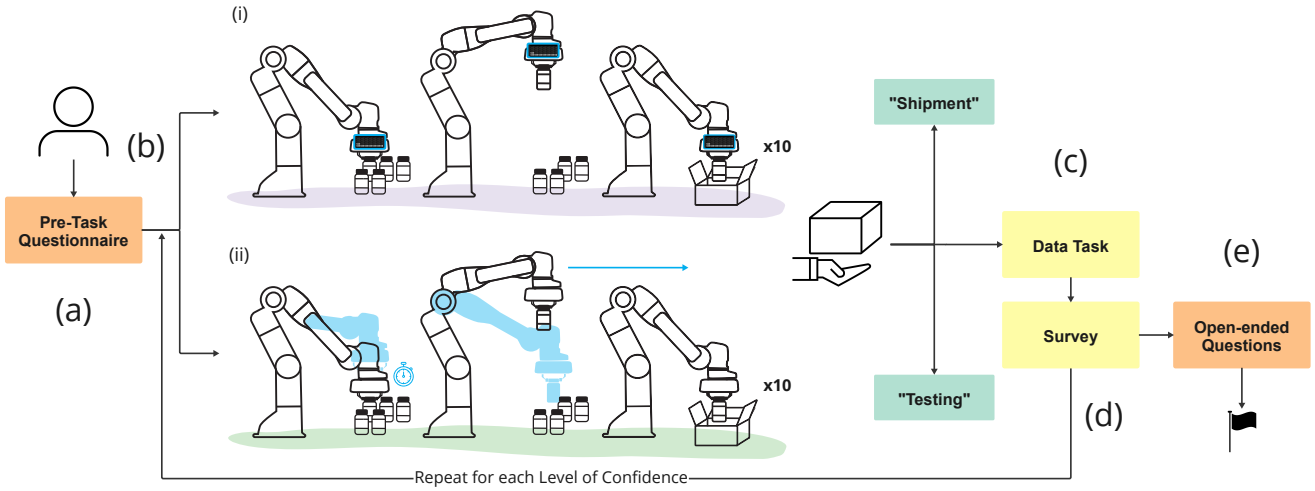


Figure 4: The experimental flow. (a): Pre-task questionnaire. (b): Collaborative task and decision-making (x10) for GUI or EMBODIED. (c): Data Task and repeated measures of trust. (d): Procedure repeated for all confidence levels, with confidence level as within factor (100% confidence, high confidence, low confidence). (e): Open-ended questions.

experience with robots and dispositional trust in new technology, see (a) in Figure 4. Secondly, the participant was provided with a written task description, including an infographic to visualise the experimental setup and the task at hand. After reading the document, the experimenter additionally verbalized the instructions and provided the participant with a preview video showing the robot conveying a 100% confidence depending on the visualisation modality the participant was assigned to. The preview video served to illustrate how the robot works, to ensure that the task is understood correctly, and to provide a baseline cue about the respective visualisation type. The next phase consisted of the actual task, followed by our repeated measures. Depending on the type of visualisation the participant was assigned to, they either experienced the robot conveying its confidence through the GUI or the EMBODIED condition, see (b) in Figure 4. The order in which the participant experienced the three different levels of confidence (100% Confidence, High Confidence, Low Confidence) was counterbalanced.

The procedure of the task itself was as follows:

1. Based on the first confidence level the participant was assigned to, the robot started to move to the first pill box, picked up the respective pill box and placed it into the package the participant positioned onto the packing station.
2. The robot conveyed its confidence either as an icon array via GUI on the robot's end-effector *or* through the EMBODIED visualisation - depending on the modality the participant was assigned to.
3. Subsequently, the participant's task was to pack their pill box into the package too, to close the package and to decide whether the package is "ready for shipment" or "needs additional testing" as their final judgement, see (b) in Figure 4.
4. After the participant's decision was made, the participant took a new, empty package, placed it onto the highlighted area of the packing station and repeated the steps above.

Each confidence level condition consisted of 10 pill boxes tested and packed by the robot and thus 10 pill boxes packed by the participant. After 10 packages had been successfully packed and put to either the "ready for shipment" or "needs additional testing" pile, the participant was asked to answer a follow-up survey, including a data task and the repeated measures. Following this, the participant was asked to wait in the resting room next door, until the next pill boxes were brought, and ready to be tested and packed. Upon completion of the participant's third session (third level of confidence), the participant was lastly asked to fill out the follow-up survey again, in addition to open-ended questions exploring their collaborative decision-making experience. Specifically, we posed questions on participants' trust, their assisted decision-making experience, their perception of the robot's confidence, and how these factors may have been mediated by the critical nature of the task scenario.

3.5 Measures

We combined validated questionnaires from assisted decision-making with autonomous agents with a classical data task from VIS. Besides questionnaires, we implemented behavioural measures to obtain more nuanced insights in people's experiences. We further measured peoples' dispositional trust in new technology [52] as a covariate, to avoid a confounding effect of participants' interindividual differences to trust technology.

3.5.1 Data Task: Average Estimation. The average estimation is a typical data task, frequently used in VIS research [2, 74, 84]. We asked participants to estimate the average confidence of the robot after each confidence level (*"Please estimate the average uncertainty (in %) of the robot:"*). Participants entered a numeric value from 0-100 in a text field. This allowed us to assess whether participants were able to correctly distinguish between low, high and 100% confidence.

3.5.2 Behavioral Measures: Decision-Making. We investigated a binary decision-making scenario to assess the participant's strategy and perception of risk in the context of (a) a robot visualising uncertainty and (b) a critical scenario, i.e. pills. Our scenario considered whether participants decided to put a package to "ready for shipment" or to "needs additional testing" after taking into account the robot's conveyed confidence in each pill box. Considering the context of the task, we defined a "correct" decision as the judgement to send the package off to shipment if the robot conveyed a 100% confidence. Otherwise, additional testing would be needed. We based this decision on the fact that we situated the task in a critical scenario with high risk, namely shipping medicine, i.e. pills, that could be defective and potentially may harm humans. We furthermore investigated if participants altered their judgement while making their final decision. We encoded "changing mind" as switching a package from "shipment" to "needs testing" and vice versa after the initial decision has been made, tracked in real-time by the experimenter. This served as an additional indicator of how confident participants were in their decision-making.

3.5.3 Subjective Measures: Trust, Transparency and Performance.

Overall Trust. First, we investigated participants' overall subjective trust in the robot, asking *"Do you trust the robot?"*, inspired by Yang et al. [94]. Participants were able to rate their trust on a bipolar slider, with *I don't trust the robot at all* on the left and *I trust the robot completely* on the right. To avoid biases caused by an anchor as a starting point on a slider, we provided participants with a bipolar slider without an anchor [78].

Multidimensional Trust Questionnaire. We furthermore captured participants perceived trust in the robot as repeated measures using the Multidimensional Trust Questionnaire (MTQ) [73]. The MQT is a validated questionnaire that particularly targets trust in automated systems and agents, while considering various subdimensions of trust to capture trust and trust dynamics in a nuanced manner. We selected the respective items related to transparency (*The way the robot works is clear to me*, *I am well informed how the robot works*, *I understand how the robot works*) and performance (*The robot works safely*, *The robot works well*, *The robot works accurately*) and asked participants to please rate the robot on a 4 point scale from *"Disagree"* to *"Agree"*.

4 RESULTS

Participants age ranged between 18 and 60 (M = 25.53, SD = 7.84); with 50% of the participants having no prior experience with robots and 50% reporting prior experience only with everyday robots, e.g. vacuum robots. Furthermore, participants generally trusted new technology (M = 5.48, SD = 1.13; 7-point Likert scale from (1) *"strongly disagree"* to (7) *"strongly agree"*).

4.1 Quantitative Results

4.1.1 Average Estimation. We summarized participants' performance on the average estimation task in Figure 5. Although the estimation of the robot's confidence was not employed as the primary manipulation check, nor the main focus of this work, we did observe that participants successfully distinguished between low, high and 100% confidence and thus perceived our manipulation as intended. Across both conditions, the GUI and the EMBODIED visualisation type, participants correctly estimated the average of the low confidence condition < high confidence < 100% confidence. Our results show an average estimation of M = 48.28% (SD = 32.96) for the low confidence condition, M = 72.78% (SD = 26.98) for the high confidence condition, and M = 75.61% (SD = 34.75) for the 100% condition for participants experiencing the GUI, see Figure 5 A. Demonstrating a comparable pattern, our findings show an average estimation of M = 54.44% (SD = 28.69) for the low confidence condition, M = 80.28% (SD = 16.49) for the high confidence condition, and M = 87.89% (SD = 13.71) for the 100% confidence condition for participants experiencing the EMBODIED visualisation, see Figure 5 B.

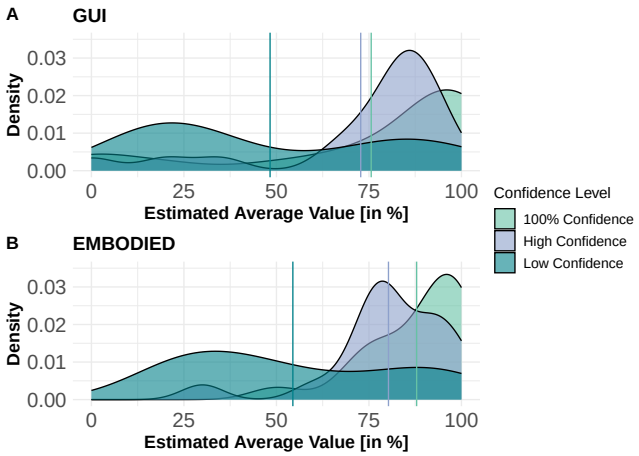


Figure 5: Average estimations shown for each visualisation type and confidence level.

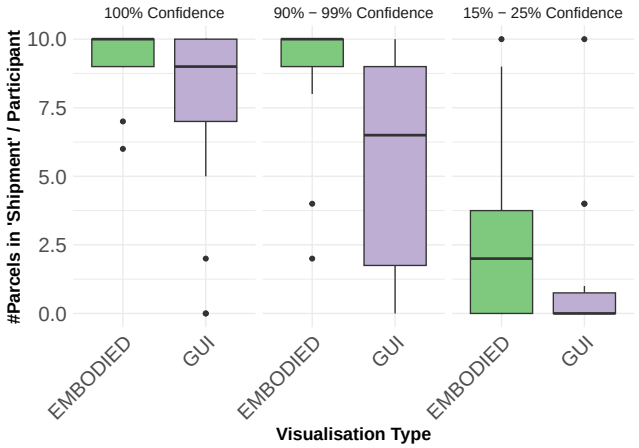
4.1.2 Decision-Making: Shipment. Next, we calculated participants' decision accuracy to get insights on RQ1.1. To do so, we investigated how often participants decided to send packages off to shipment, depending on the experienced type of uncertainty visualisation and confidence level. In the 100% confidence condition, 12 out of 18 participants in the EMBODIED group correctly decided to put all 10 packages into the ready for shipment pile. Contrary to only 7 out of 18 participants out of GUI group, see Figure 6a. We further conducted an analysis on how many packages participants decided to ship depending on the actual probability value the robot conveyed for each pill box and the type of visualisation. For this analysis, we therefore considered the noise packages and examined the probability range of the respective pill box (100% vs. 90%-99% confidence vs. 15%-25% confidence). The results are shown in Figure 6b.

We performed a mixed analysis of covariance (ANCOVA) test after controlling for the dispositional trust in new technology. Our analysis reveals a significant main effect of the type of visualisation on the decision to send packages off to shipment ($F(1, 33) = 11.304$, $p = .002$, $\eta^2 = 0.13$). A post-hoc Tukey pairwise comparison shows that participants who experienced the EMBODIED visualisation (M = 7.0, SE = 1.73) placed significantly more packages to the "ready for shipment" pile compared to participants who experienced

the GUI ($M = 4.7, SE = 1.80$), $p < .005$. The results show a significant main effect of the confidence level on the decision to send packages off to shipment ($F(2, 68) = 62.80$, $p < .000$, $\eta^2 = 0.56$). A post-hoc Tukey pairwise comparison reveals that participants in the 100% confidence condition ($M = 8.38$, $SE = 1.77$) sent significantly more packages off to shipment, compared to the low confidence condition ($M = 1.93$, $SE = 1.77$); $p < .0001$. Additionally, participants in the high confidence ($M = 7.24$, $SE = 1.77$) condition sent significantly more packages off to shipment compared to the low confidence condition; $p < .0001$. There was no significant difference between the 100% and high confidence condition; $p = .160$.



(a) Total count of participants who correctly decided to place all 10 packages to shipment in the 100% confidence condition.



(b) Total count of packages placed to shipment/participant for each confidence range of the respective pill boxes.

Figure 6: Participants' decisions to send packages off to shipment.

We additionally assessed if participants changed their mind during their decision-making process to get insights on RQ1.2. Results are shown in Figure 7. Interestingly, 6 out of 18 participants in the GUI condition changed their mind in the 100% confidence condition compared to 0 out of 18 participants in the EMBODIED condition. In both, the high and low confidence condition, the amount of participants who changed their mind was ≤ 4.

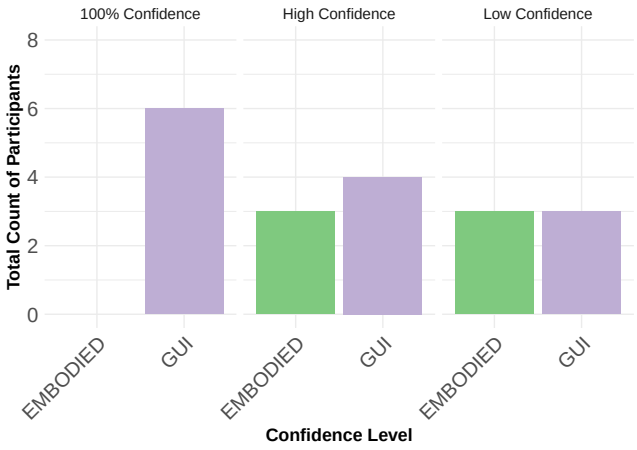


Figure 7: Total count of participants who changed their mind during their decision-making process for each confidence level.

4.1.3 Overall Trust in the Robot. Guided by RQ2, we explored the relationship between visualising uncertainty and trust in autonomous agents. Participants' overall subjective trust in the robot was descriptively high across conditions, see Figure 8a. Notably, a mixed ANCOVA shows no significant main effect of the confidence level ($F(2, 68) = 2.356$, $p = .102$) or the visualisation type ($F(2, 68) = 2.716$, $p = .109$), nor does it show an interaction effect between the type of uncertainty visualisation and the level of confidence the participants experienced ($F(2, 68) = 1.080$, $p = .345$). Interestingly, on a descriptive level, participants in the GUI condition rated their overall trust in the robot the lowest in the 100% confidence condition, shown in Figure 8a.

4.1.4 Trust-related Transparency and Performance. To obtain more nuanced insights regarding trust in robots, particularly in high-risk assisted decision-making scenarios, we examined two subdimensions of the MTQ [73] in addition to our single-item: transparency (RQ2.1) and performance (RQ2.2). The Cronbach's α for the transparency ($\alpha = 0.87 - 0.91$) and the performance ($\alpha = 0.77 - 0.89$) subdimension indicates a strong internal consistency among the items. Interestingly, the results of a mixed ANCOVA show no significant effect between the GUI ($M = 1.68$, $SE = 0.746$) and the EMBODIED visualisation ($M = 1.68$, $SE = 0.714$) on perceived transparency; $F(1, 33) = 0.031$, $p = .862$. This is surprising, considering the items that measure transparency as a subdimension (*The way the robot works is clear to me*, *I am well informed how the robot works*, *I understand how the robot works*). The results are shown in Figure 8b. Further, there was no interaction effect between the type of uncertainty visualisation and the confidence level ($F(2, 68) = 0.765$, $p = .469$). However, we observed a significant effect of confidence level on perceived transparency; $F(2, 68) = 3.895$, $p = .025$, $\eta^2 = 0.02$. A post-hoc Tukey pairwise comparison revealed that participants in the high confidence condition ($M = 1.83$, $SE = 0.719$)

perceived the robot as significantly more transparent compared to the low confidence condition ($M = 1.54$, $SE = 0.719$), $p < 0.05$.

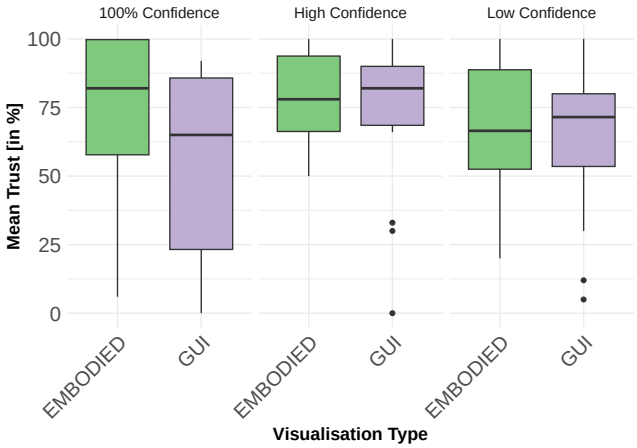
Next, we investigated the subdimension that covers trust related to the perceived performance of the automated agent, in our context, a robotic arm. However, the results of a mixed ANCOVA reveal no significant effect of either the visualisation type ($F(1, 33) = 0.810$, $p = .375$) or confidence level ($F(1.5, 49.48) = 1.020$, $p = .349$), nor an interaction between those two ($F(2, 68) = 1.023$, $p = .365$). Lastly, we investigated the overall score of both subdimensions to infer insights on participant's trust in the robot. Here too, our results reveal no significant effect of either the visualisation type ($F(1, 33) = 0.077$, $p = .784$) or confidence level ($F(2, 68) = 2.939$, $p = .0597$), nor an interaction between those two ($F(1.37, 45.17) = 0.710$, $p = .446$). However, descriptively, our results show that participants in the high confidence condition ($M = 2.67$, $SE = 0.568$) rated their trust higher compared to the 100% confidence condition ($M = 2.53$, $SE = 0.568$) and the low confidence condition ($M = 2.45$, $SE = 0.568$).

4.2 Qualitative Results

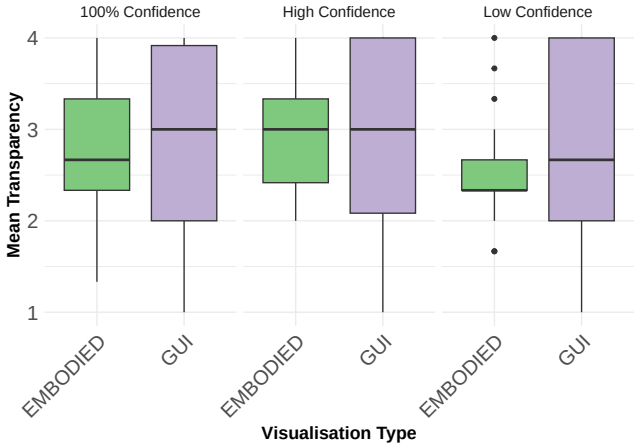
Lastly, we analysed the open-ended questions asked at the end of the experiment. We aimed to gain insights into people's decision-making process, trust and their experience with a robot that is assisting them in a critical task. We followed a deductive thematic analysis approach to systematically code participants' responses [13]. To perform a thematic analysis, we constructed a coding framework rooted in pre-established themes. Each theme was based on past literature and corresponded to our research objectives. First, we gained a holistic understanding of the data by reviewing all the responses provided by the participants. Subsequently, we annotated sections of participants' responses that aligned with our predetermined themes and assigned these quotes to the respective themes.

4.2.1 Perception of Confidence. The majority of the participants who experienced the EMBODIED visualisation interpreted *"the time it took the robot to pick up the pills"* (P12, EMBODIED) as a cue for the robot's confidence. In line with previous research in HRI, the majority of participants identified the robot's waiting time as *hesitation* [61–63]; *"It would hesitate for quite a while whenever it was not confident, and it wouldn't stop at all if it was confident."* (P30, EMBODIED). Besides the robot's waiting time, participants also perceived the robot's travel time as an additional cue for uncertainty; *"Its movements were very smooth when it packed bottles it knew were not defective."* (P2, EMBODIED). Hence, the robot's smoothness, also referred to as motion fluency [88], was recognized. Interestingly, their retrospective perception of the robot's confidence was perceived rather binary (confident or not confident), depending on the presence of these cues; *"It appeared confident to me when it picked up the pill box quickly, because it seemed like it was not hesitating to put it into the parcel."* (P36, EMBODIED).

As expected, the majority of participants in the GUI condition correctly interpreted the graphical display as the robot's confidence. However, their perception of the robot's confidence was more varied compared to participants in the EMBODIED condition, resulting in the following insights.



(a) Participants' overall trust in the robot [in %] for each condition.



(b) Participants' perceived transparency of the robot, using the transparency subdimension of the MTQ [73].

Figure 8: Overall trust and transparency as a trust-related construct.

4.2.2 Risk Threshold and Awareness. The majority of participants in the GUI condition applied a personal risk threshold to make a decision; *"I just kind of set the threshold to be 95%."* (P1, GUI) and *"If it was above 95% I would place it in the ready to be shipped section, otherwise it [...] needs additional testing."* (P31, GUI). Notably, several participants indicated a high risk-awareness of the setting in which they collaboratively executed the task with the robot. Participants elaborated *"I didn't want to risk it"* (P3, GUI) and explained that *"above 90% is still not enough assurance of someone taking a bad pill."* (P21, GUI). Further, this participant stated that *"I wouldn't give these pills to my children"* (P21, GUI). Besides setting a defined threshold, some participants reported that their risk threshold changed over time, depending on the confidence the robot conveyed and the frame of reference they experienced; *"As some boxes were predicted to have around 20%, I moved the 90% confidence boxes to the ready for shipment area."* (P25, GUI). Further, several participants struggled to apply a threshold that could be realistically applied; *"I originally*

planned to assign all of the boxes that did not achieve 100% confidence for additional testing. But I was just not sure if the number can really be achieved in a real situation, so I set a threshold for the last several boxes." (P1, GUI).

4.2.3 Fully Confident vs. Overconfident. Surprisingly, several participants in the GUI condition interpreted the 100% confidence condition as the robot being too confident. This perception made participants feel uneasy and caused them to distrust the robot; *"In the third session [100% Confidence], the robot seemed too confident, for that I can not trust it very well."* (P11, GUI) and *"I think something went wrong with it in trial 2 [100% Confidence] causing it to be overconfident."* (P29, GUI). Additionally, some participants expressed that the 100% confidence condition made them question the robot's reliability; *"When I saw all the packages are all 100% confidence, I am [sic] unsure about the reliability of the robot."* (P7, GUI). Conversely, some participants did believe that a fully confident robot was plausible; *"100% did make me question its reliability, but I still thought it was a plausible result and would still trust the robot's judgement."* (P31, GUI). In the embodied uncertainty visualisation condition, only a few participants experienced a distrust in a fully confident robot; *"The trust changed from positive to negative extremely after doing the second session [100% Confidence] in which I got a sense that the robot didn't work properly as it showed no hesitation on picking and putting all bottles in the boxes."* (P22, EMBODIED).

A few participants considered the robot's *behavioural consistency* as an influencing factor in their assessment of the robot, unveiling contrasting statements. Some participants perceived consistent exhibitions of confidence to be an indicator of poor ability; *"The results must be varied to show me that the robot is really working."* (P1, GUI). In contrast, some participants regarded a lack of consistency as an indication of the robot's reduced confidence; *"Since it did not show consistency, I felt that the robot [sic] behavior is not confident."* (P15, GUI).

4.2.4 A Need for Additional Explanation. Interestingly, some participants in the GUI condition expressed concerns regarding a lack of information and their need for further explanation, since they did not know *"the mechanism behind the confidence"* (P7, GUI). *"I would just wish I had an explanation for what it is checking for and how. How it is coming up with its confidence level."* (P3, GUI). Further, participants explained that the lack of additional information impacted the robot's credibility. *"I need more information to be able to believe the robot's assessment"* (P15, GUI).

4.2.5 Responsibility. Our qualitative analysis also reveals that participants had a different idea of who should be in charge for the final call of the decision-making task. Only a few participants stated that the final judgement should be up to the user; *"Workers can make their own choices about if they believe the pills are ok to be shipped or not."* (P16, EMBODIED). Conversely, several participants indicated that they were willing to default the decision-making to the robot; *"I gave up my decision making capability to the robot's confidence. That's how far it affected my expectations."* (P22, EMBODIED). Further, they expressed that they would have preferred the robot to explicitly tell them which package to ship or test; ' *"Tell me if the pills are ready for shipment or additional packaging is required."* (P9, GUI). Participants also explained that they *"assumed that if it [the robot] was fully unconfident, it would not place that particular pill box in the [shipping] box."* (P32, EMBODIED), which also suggests a tendency to rely on the robot's judgement in the decision-making process, and the willingness to hand over responsibility.

5 DISCUSSION

Both the EMBODIED and GUI visualisation can potentially serve as *communicative actions* [33] to make humans understand what robots do and think and to inform decision-making in a collaborative setting. We discuss a potential trade-off between EMBODIED uncertainty visualisations, that enable users to make intuitive decisions, with the downside of information loss, and GUI visualisations, that allow users to make fully informed decisions, with the pitfall of distrust in robots. To do so, we highlight our insights on how using uncertainty visualisations in robot-assisted decision-making affect people's judgement, their perception of the robot and the risk itself. Further, we outline fundamental implications to inform future design of uncertainty visualisations in robot-assisted decision tasks.

5.1 Visualising Uncertainty affects the Perception of the Robot

The results from our manipulation check and average estimation show that participants across conditions were able to successfully perceive the robot conveying a low, high, and 100% confidence. Moreover, participants in the EMBODIED condition recognized the manipulated *waiting and travel time* as *hesitation* and an indicator for the robot's confidence. These results are in line with previous research on non-verbal cues in HRI and show that people apply known anthropomorphic gestures and characteristics to a robot's behavior [36, 62, 96]. In the GUI condition, participants also correctly interpreted the uncertainty visualisation as the robot's conveyed confidence. Further, our results show that an icon array, a classical approach to visualise risk [60], is an appropriate approach to visualise a robot's confidence. Our qualitative analysis additionally suggests that the graphical display was interpreted as a robot's extension, and not as an artefact separate from the robot.

5.2 Robot-Assisted Decision-Making

We defined decision as a binary variable: Participants were able to either send a package "off to shipment" or place it in the "needs additional testing" area. Our results reveal a general risk-awareness for both uncertainty visualisation conditions, with a clear distinction between high and low risks. Participants in the 100% confidence and the high confidence condition sent significantly more packages off to shipment compared to the low confidence condition. Based on the task setup and the severity of a false-positive, i.e. to ship defective pills, we defined a *correct decision* as the judgement to send a package off to shipment *if* the robot conveys a 100% confidence. Surprisingly, our results show that in the 100% confidence condition, the majority of participants in the EMBODIED condition correctly sent packages off to shipment; compared to only 7 out of the participants in the GUI condition. This is non-trivial, since it indicates that participants in the GUI condition were more willing to override the robot's information suggesting that the pills are not defective with the utmost confidence.

5.2.1 The Curse of Perfection. Interestingly, our qualitative analysis reveals that participants in the GUI condition felt somewhat uncomfortable if the robot was fully confident: the majority of participants in the GUI condition perceived the robot as "over-confident", which made them distrust the robot and question its "reliability." Thus, participants in the GUI condition showed reservations about the robot's abilities despite it exhibiting 100% confidence. We posit that the perceptional differences of a fully confident robot (100% confidence condition) could have occurred based on three reasons: First, the robot in the EMBODIED condition neither *waited* nor *adjusted its speed*, since both trajectory functions take the respective confidence value as the variable that determines its waiting and travel time. As a result, participants interacted with a robot that, for them, displayed an expected default movement, never deviating from it. Second, the graphical display mounted on the robot's end-effector could have resulted in a higher information salience, which might have made participants in the GUI condition more cautious and more engaged with the displayed information. In a recent think-aloud study, Prabhudesai et al. [69] show that "communicating uncertainty about ML predictions forced people to slow down and think analytically about their decision." The authors conclude that visualising uncertainty made people more aware and considerate, which reduced their over-reliance in AI-assisted decision-support systems. The latter is also reflected in the amount of alterations in participants' opinions, shown in Figure 7. If the robot was fully confident, 6 out of 18 participants in the GUI condition changed their mind during their decision-making, compared to 0 participants in the EMBODIED condition. We expect that an alteration of an opinion can be an indicator for an ongoing reflection and hence can be interpreted as being more considerate. Some participants in the GUI condition even specifically asked for additional explanations and wished to understand the robot's functionalities better. Third, participants in the GUI condition could have been risk-aware based on the fact that the robot repetitively conveyed 100% and might have been taken by surprise by the robot's behavioral consistency, compared to participants in the EMBODIED condition who were satisfied with the expected default movement of the robot.

5.2.2 Threshold vs. General Heuristic. Besides peoples' distrust in a robot that conveys 100% confidence, our qualitative results show that the majority of participants in the GUI condition expressed to have applied a personal risk-threshold in the range of 90% - 95% to inform their decision-making. This risk threshold is higher compared to e.g. a recently conducted gluten detection task, in which people applied an average threshold of 75% to trust a scanner [44]. Participants hereby based their personal threshold on the fact that shipping defect pills poses a critical risk. In addition, several participants explained to have changed their risk-threshold over time, which also indicates a more considered approach to form their decision.

Conversely, participants in the EMBODIED condition retrospectively perceived the robot as rather confident or not confident and did not express any thoughts on thresholds. In addition, our quantitative analysis reveals that participants in the EMBODIED condition placed significantly more packages to the "ready for shipment" pile compared to participants in the GUI condition, which we interpret as an indicator for less caution and risk-awareness. The embodied uncertainty visualisation might have encouraged participants to apply a more general and intuitive heuristic to assess the robot's confidence, instead of carefully checking and comparing the robot's speed and waiting time. This assumption is motivated by Buçinca et al. [15], who proposes that people are *intrinsically* not motivated to approach a decision analytically in an AI-assisted task. The authors suggest that people rather develop an overall rule-of-thumb to generally assess the competence of an AI [15]. We postulate that the same theory can be applied to robot-assisted decision-making, if the uncertainty visualisation allows such a shortcut due to a lack of salience.

5.2.3 The Willingness to Hand Over Responsibility. In our setup, the robot was responsible to assess whether pills are defect or not and to visualise its confidence respectively. As a result, participants had to take into account the robot's assessment and recommendation. Surprisingly, our qualitative analysis reveals that several participants were willing to give up their responsibility in the decision-making process and hand over the final judgement to the robot. This might be due to individual differences in risk perception. Huang et al. [37] highlight that individual risk perceptions might affect the extent to which people are willing to follow a robot in an assisted-decision making task. People with an overall low-risk awareness might be more willing to hand over the final judgement. Further, this result is in line with previous research on AI-assisted decision-making that identified over-reliance as a crucial issue and supports the theory that people tend to avoid cognitive effort if another shortcut presents itself [15].

5.3 Trust, Performance and Transparency

Neither the robot's uncertainty visualisation nor the conveyed confidence level had a significant effect on the perceived overall trust in the robot or its perceived performance. A potential explanation could be the fact that the robot conveyed its uncertainty to inform the participants' decision, without explicitly providing a recommendation of whether a pill box *should* be shipped or tested. Hence, the participant holds the agency to decide. Further, we did not offer the opportunity to verify the accuracy of the robot and to factor in human-robot agreement [56]. We acknowledge the latter in our limitations, see Section 5.4.

Furthermore, transparency has been identified as a highly relevant trust construct when interacting with an automated system or agent [70, 73]. A recent paper by Leusmann et al. [54] proposes that a robot needs to be able to convey its own uncertainty in order to support understandability. We assumed that a graphical display would provide a higher salience and context-awareness, based on the symbolism of an icon array [58] and hence would support transparency. However, contrary to our assumptions, there was no difference between the visualisation groups regarding transparency. We argue that this result is a strong indication for the power of embodied visualisations and their intuitive understanding. If we interpret the EMBODIED condition as an anthropomorphic visualisation and the GUI as a technical visualisation, our results are also not confirming recent findings by Roesler [70] on perceived transparency.

Notably, we observe a significant effect of the robot's conveyed confidence level on perceived transparency. Participants in the high confidence condition perceived the robot as significantly more transparent compared to the low confidence condition. Interestingly, the items we assessed to determine the perceived transparency focus on how well people *understood* the robot and its functionality (*I understand how the robot works*) [73]. Hence, our results suggest that participants in the low confidence condition felt as if they understood the robot less. Research in psychology has investigated the perception of probabilities. An early study by Kunreuther et al. [51] shows that low probability-high consequence events are particularly difficult to interpret, which is why these settings require rich context and comparison scenarios. Interestingly, our results indicate that (a) the level of confidence a robot conveys affects the user's *understanding* of the robot and (b) that we need to find ways to enhance users understanding in low-probability scenarios. It is important to highlight that transparency and performance are only two out of many trust-related constructs. Future work should investigate other related constructs such as distrust, reliance and compliance, that have been identified as important trust factors in assisted decision-making [89].

5.3.1 Implications for Robot-assisted Decision-making. Along with the aforementioned findings, we outline fundamental implications and design considerations when visualising uncertainty in a robot-assisted decision-making task.

1. *Uncertainty visualised by a robot affects humans' decision-making.* Our results show that the type of uncertainty visualisation conveyed by a robot has a significant impact on participants' decision-making in a high-stake scenario. Future applications have to carefully consider the type of visualisation and the expected confidence range to best support human decision-making in robot-assisted decision tasks.
2. *Caution is needed when visualising a fully confident robot.* In detail, our study shows that repetitively visualising 100% confidence in a robot-assisted decision-task can be a double-edged sword. Participants in the GUI condition perceived the robot as too confident, which caused them to incorrectly decide to test pill boxes. These non-trivial findings suggest that in cases of repetitive 100% probability values, the robot's confidence should be conveyed using an implicit EMBODIED visualisation. Future research should investigate whether these findings are based on the particular probability value (100%) or based on the repetition itself and hence the behavioural consistency of the robot. Our results also open up the question whether, in case of a 100% confidence, confidence should be visualised at all.
3. *Visualising uncertainty is highly context-dependent.* The aforementioned GUI induced skepticism of a robot conveying 100% confidence can be interpreted as a hindrance or an asset, depending on the application context and goal of the task. A robot using a GUI as a means to convey its confidence enables users to assess a risk in detail, to define or to apply a risk threshold. However, this approach may lead to incorrect decisions caused by overthinking, distrust, or fluctuations in judgment, potentially impacting decision speed. Conversely, our results show that a robot conveying its confidence using an EMBODIED visualisation is perceived as either confident or not confident, with the pitfall of less information granularity and a less detailed risk assessment. We believe that an EMBODIED visualisation could nevertheless be useful for low-stakes scenarios and decisions that require an intuitive understanding of the situation. Further, people with low literacy skills could benefit from such implicit visualisations.
4. *Low probabilities should be explained.* Participants perceived the robot as significantly less transparent in the low confidence condition compared to the high confidence condition. Taking into account our qualitative findings and literature on the perception of probabilities [48, 51], we recommend that low risks displayed by a robot should be further explained and contextualized to support user understanding.
5. *Conveying uncertainty does not affect trust in a robot.* Notably, participants' overall trust in the robot was high across conditions. This finding raises the question of whether visualising a robot's uncertainty could potentially prevent negative effects after e.g. a robot's failure [22] and act as a buffer, cushioning user trust. However, more research is needed to assess the latter in detail.

5.4 Limitations

We acknowledge a number of limitations in our study. First, we situated our study in a high-stake decision-making scenario. Therefore, our findings may not be applicable to more mundane scenarios. However, we were particularly interested in robot-assisted decision-making for critical applications, since robots can be ideally used to be implemented in hazardous environments or to work with hazardous materials. Hence, it is important to understand how humans perceive and assess the robot, and the risks in case of severe consequences. Interestingly, even though we did not assign a cost to "additional testing", participants did not blindly decided to put packages into the "needs additional testing" area and over-reported the safe decision, as previous studies have shown [44]. However, future research should investigate how putting a cost to testing could affect human's decision-behaviour and the perception of the robot.

Furthermore, sending a package to either "shipment" or "testing" represents a binary decision, a very common daily life experience [44]. However, we acknowledge the fact that there are more complex decision types, like tactical decision-making [35], and that our results cannot be generalized to other decision categories. Moreover, we acknowledge that our research does not cover the entire range of probability values, but focuses on three distinct categories (low, high and 100%) based on McCaffery et al. [60]. Addressing uncertainty values outside our categories opens up new possibilities for future research. To increase the internal validity of our study, we further conducted our experiment as a lab study instead of a pharmaceutical warehouse. Nevertheless, we aimed to enhance the transferability of our study by designing our setup as realistically as possible. For instance, we prepared the working station to resemble a packing station. We further provided participants with 30 different packages and separate pill boxes, to ensure that they understand that the robot is assessing different and not yet detected pill boxes during each consecutive trial.

We did not find significant difference in participants' trust and perceived performance of the robot. This could be based on the fact that we did not manipulate the robot's competence or error frequency, which has proven to affect people's perceptions negatively [82]. In our study setup, the robot conveyed its confidence in its pill detection, since this study's focus is on the impact of uncertainty visualisations and confidence levels. Future work should aim to address the relationship between trust formation and failure of a robot that conveys its confidence in a robot-assisted decision-making. We believe that visualising a robot's confidence could potentially serve as a failure justification [49], and improve trust and error-prone robot-assisted decision-making.

6 CONCLUSION

We conducted a lab study with 36 participants who collaborated with a robot arm in an assisted decision-making task. Our main objectives were twofold: First, to investigate how a robot that conveys its uncertainty affects the human's decision-making process. Second, to explore how visualising a robot's uncertainty affects the human's perception of the robot itself. We further situated the study in a high-stake scenario. Our findings reveal that the type of visualisation significantly affects users' judgement and decision-behaviour. We show that the robot's conveyed confidence level significantly affects a user's perceived transparency of the robot. Besides, we shed light on a potential trade-off between making an intuitive or an informed decision in a robot-assisted decision-making scenario and how visualising a fully confident robot can be a double-edged sword. Based on our findings, we identify new avenues for future research and provide implications for uncertainty visualisations in critical robot-assisted decision-making tasks. With this study, we take a step closer towards better understanding how data visualisation can be employed in collaborative setting between autonomous embodied agents and humans to support human decision-making.

ACKNOWLEDGMENTS

We wish to acknowledge the contribution of the Melbourne Research Scholarship funded by the University of Melbourne and the Rowden White Scholarship established with a donation from Sir Alfred Edward Rowden White. This research is partially supported by the Australian Research Council Discovery Early Career Research Award (Grant No. DE210100858). Further, we thank Jiahe Pan, Muhammad Bilal and Antony Chacon Salas for their support with the technical implementation.

REFERENCES

[1] FRANKA EMIKA GmbH 2023. *The new FRANKA RESEARCH 3 - The platform of choice for cutting edge AI & Robotics research.* FRANKA EMIKA GmbH. Retrieved September 06, 2023 from https://www.franka.de/research/

[2] Danielle Albers, Michael Correll, and Michael Gleicher. 2014. Task-Driven Evaluation of Aggregation in Time Series Visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA, 551–560. https://doi.org/10.1145/2556288.2557200

[3] Jessica S. Ancker, Yalini Senathirajah, Rita Kukafka, and Justin B. Starren. 2006. Design Features of Graphs in Health Risk Communication: A Systematic Review. *Journal of the American Medical Informatics Association* 13, 6 (11 2006), 608–618. https://doi.org/10.1197/jamia.M2115

[4] Adriana Arcia, Michael E Bales, William Brown III, Manuel C Co Jr, Melinda Gilmore, Young Ji Lee, Chin S Park, Jennifer Prey, Mark Velez, Janet Woollen, et al. 2013. Method for the development of data visualizations for community members with varying levels of health literacy. In *AMIA Annual Symposium Proceedings*, Vol. 2013. American Medical Informatics Association, 51.

[5] Adriana Arcia, Niurka Suero-Tejeda, Michael E Bales, Jacqueline A Merrill, Sunmoo Yoon, Janet Woollen, and Suzanne Bakken. 2015. Sometimes more is more: iterative participatory design of infographics for engagement of community members with varying levels of health literacy. *Journal of the American Medical Informatics Association* 23, 1 (07 2015), 174–183. https://doi.org/10.1093/jamia/ocv079

[6] S. Sandra Bae, Clement Zheng, Mary Etta West, Ellen Yi-Luen Do, Samuel Huron, and Danielle Albers Szafir. 2022. Making Data Tangible: A Cross-Disciplinary Design Space for Data Physicalization. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, New York, NY, USA, Article 81, 18 pages. https://doi.org/10.1145/3491102.3501939

[7] Bernard W. Balleine. 2007. The Neural Basis of Choice and Decision Making. *Journal of Neuroscience* 27, 31 (2007), 8159–8160. https://doi.org/10.1523/JNEUROSCI.1939-07.2007

[8] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[9] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. ACM, New York, NY, USA, 401–413. https://doi.org/10.1145/3461702.3462571

[10] Melissa Bica, Julie L. Demuth, James E. Dykes, and Leysia Palen. 2019. Communicating Hurricane Risks: Multi-Method Examination of Risk Imagery Diffusion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300545

[11] Michelle Borkin, Krzysztof Gajos, Amanda Peters, Dimitrios Mitsouras, Simone Melchionna, Frank Rybicki, Charles Feldman, and Hanspeter Pfister. 2011. Evaluation of Artery Visualizations for Heart Disease Diagnosis. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2479–2488. https://doi.org/10.1109/TVCG.2011.192

[12] Nadia Boukhelifa and David John Duke. 2009. Uncertainty Visualization: Why Might It Fail?. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. ACM, New York, NY, USA, 4051–4056. https://doi.org/10.1145/1520340.1520616

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[14] Ken Brodlie, Rodolfo Allendes Osorio, and Adriano Lopes. 2012. A review of uncertainty in data visualization. *Expanding the frontiers of visual analytics and visualization* (2012), 81–109.

[15] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[16] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[17] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 471 (nov 2022), 23 pages. https://doi.org/10.1145/3555572

[18] Felix Carros, Johanna Meurer, Diana Löffler, David Unbehaun, Sarah Matthies, Inga Koch, Rainer Wieching, Dave Randall, Marc Hassenzahl, and Volker Wulf. 2020. Exploring Human-Robot Interaction with the Elderly: Results from a Ten-Week Case Study in a Care Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376402

[19] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[20] Alan M. MacEachren Christoph Kinkeldey and Jochen Schiewe. 2014. How to Assess Visual Communication of Uncertainty? A Systematic Review of Geospatial Uncertainty Visualisation User Studies. *The Cartographic Journal* 51, 4 (2014), 372–386. https://doi.org/10.1179/1743277414Y.0000000099

[21] Jae Chung and Steven Wark. 2016. Visualising uncertainty for decision support. *Defence Science and Technology Group* (2016).

[22] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. 2018. Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 507–513.

[23] Mohammad Daradkeh, Clare Churcher, and Alan McKinnon. 2013. Supporting Informed Decision-Making under Uncertainty and Risk through Interactive Visualisation. In *Proceedings of the Fourteenth Australasian User Interface Conference - Volume 139* (Melbourne, Australia) *(AUIC '13)*. Australian Computer Society, Inc., AUS, 23–32.

[24] Eric Deng, Bilge Mutlu, and Maja J Mataric. 2019. Embodiment in Socially Interactive Robots. *Foundations and Trends® in Robotics* 7, 4 (2019), 251–356. https://doi.org/10.1561/2300000056

[25] George Denison. 2023. *How much should you pay research participants?* Prolific. Retrieved August 14, 2023 from https://www.prolific.co/blog/how-much-should-you-pay-research-participants

[26] Ruta Desai, Fraser Anderson, Justin Matejka, Stelian Coros, James McCann, George Fitzmaurice, and Tovi Grossman. 2019. Geppetto: Enabling Semantic Design of Expressive Robot Behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300599

[27] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (2007), 175–191. https://doi.org/10.3758/BF03193146

[28] S Larry Goldenberg, Guy Nir, and Septimiu E Salcudean. 2019. A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology* 16, 7 (2019), 391–403. https://doi.org/10.1038/s41585-019-0193-3

[29] Miriam Greis, Jessica Hullman, Michael Correll, Matthew Kay, and Orit Shaer. 2017. Designing for Uncertainty in HCI: When Does Uncertainty Help?. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. ACM, New York, NY, USA, 593–600. https://doi.org/10.1145/3027063.3027091

[30] Jian Guo, Xiaodong Zhang, Haiming Shen, Haibo Xu, and Liangliang Li. 2020. Design on Control System of Bionic Crawling Inspection Robot. In *Proceedings of the 2019 2nd International Conference on Control and Robot Technology* (Jeju, Republic of Korea) *(ICCRT 2019)*. ACM, New York, NY, USA, 24–29. https://doi.org/10.1145/3387304.3387312

[31] Zixuan Guo and Tomoo Inoue. 2019. Using a Conversational Agent to Facilitate Non-Native Speaker's Active Participation in Conversation. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3313075

[32] Sarah T. Hawley, Brian Zikmund-Fisher, Peter Ubel, Aleksandra Jancovic, Todd Lucas, and Angela Fagerlin. 2008. The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient Education and Counseling* 73, 3 (2008), 448–455. https://doi.org/10.1016/j.pec.2008.07.023 4th International Conference on Shared Decision Making.

[33] Thomas Hellström and Suna Bensch. 2018. Understandable robots - What, Why, and How. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 110–123. https://doi.org/doi:10.1515/pjbr-2018-0009

[34] Alejandra C. Hernández, Clara Gómez, Jonathan Crespo, and Ramón Barber. 2017. Adding Uncertainty to an Object Detection System for Mobile Robots. In *2017 6th International Conference on Space Mission Challenges for Information Technology (SMC-IT)*. 7–12.

[35] Carl-Johan Hoel, Katherine Driggs-Campbell, Krister Wolff, Leo Laine, and Mykel J. Kochenderfer. 2020. Combining Planning and Deep Reinforcement Learning in Tactical Decision Making for Autonomous Driving. *IEEE Transactions on Intelligent Vehicles* 5, 2 (2020), 294–305.

[36] Julian Hough and David Schlangen. 2017. It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. ACM, New York, NY, USA, 274–282. https://doi.org/10.1145/2909824.3020214

[37] Hanjing Huang, Pei-Luen Patrick Rau, and Liang Ma. 2021. Will you listen to a robot? Effects of robot ability, task complexity, and risk on human decision-making. *Advanced Robotics* 35, 19 (2021), 1156–1166. https://doi.org/10.1080/01691864.2021.1974940

[38] Jessica Hullman. 2016. Why Evaluating Uncertainty Visualization is Error Prone. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (Baltimore, MD, USA) *(BELIV '16)*. ACM, New York, NY, USA, 143–151.

[39] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2019. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 903–913. https://doi.org/10.1109/TVCG.2018.2864889

[40] Anil Kumar Inkulu, MVA Raju Bahubalendruni, and Ashok Dara. 2022. Challenges and opportunities in human robot collaboration context of Industry 4.0-a state of the art review. *Industrial Robot: the international journal of robotics research and application* 49, 2 (2022), 226–239.

[41] Yvonne Jansen, Pierre Dragicevic, Petra Isenberg, Jason Alexander, Abhijit Karnik, Johan Kildal, Sriram Subramanian, and Kasper Hornbæk. 2015. Opportunities and Challenges for Data Physicalization. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 3227–3236. https://doi.org/10.1145/2702123.2702180

[42] Luke Jefferson and Richard Harvey. 2007. An Interface to Support Color Blind Computer Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. ACM, New York, NY, USA, 1535–1538. https://doi.org/10.1145/1240624.1240855

[43] Feng Jiang, Su Lu, Xiang Yao, Xiaodong Yue, and Wing tung Au. 2014. Up or Down? How Culture and Color Affect Judgments. *Journal of Behavioral Decision Making* 27, 3 (2014), 226–234. https://doi.org/10.1002/bdm.1800

[44] Weiwei Jiang, Zhanna Sarsenbayeva, Niels van Berkel, Chaofan Wang, Difeng Yu, Jing Wei, Jorge Goncalves, and Vassilis Kostakos. 2021. User Trust in Assisted Decision-Making Using Miniaturized Near-Infrared Spectroscopy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 153, 16 pages. https://doi.org/10.1145/3411764.3445710

[45] Susan Joslyn and Jared LeClerc. 2013. Decisions With Uncertainty: The Glass Half Full. *Current Directions in Psychological Science* 22, 4 (2013), 308–315. https://doi.org/10.1177/0963721413481473

[46] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (Ish) is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5092–5103. https://doi.org/10.1145/2858036.2858558

[47] Matthew Kay, Dan Morris, mc schraefel, and Julie A. Kientz. 2013. There's No Such Thing as Gaining a Pound: Reconsidering the Bathroom Scale User Interface. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. ACM, New York, NY, USA, 401–410. https://doi.org/10.1145/2493432.2493456

[48] Mel W. Khaw, Luminita Stevens, and Michael Woodford. 2021. Individual differences in the perception of probability. *PLOS Computational Biology* 17, 4 (04 2021), 1–25. https://doi.org/10.1371/journal.pcbi.1008871

[49] Ross A Knepper, Stefanie Tellex, Adrian Li, Nicholas Roy, and Daniela Rus. 2015. Recovering from failure by asking for help. *Autonomous Robots* 39 (2015), 347–362. https://doi.org/10.1007/s10514-015-9460-1

[50] Michelle Korporaal, Ian T. Ruginski, and Sara Irina Fabrikant. 2020. Effects of Uncertainty Visualization on Map-Based Decision Making Under Time Pressure. *Frontiers in Computer Science* 2 (2020). https://doi.org/10.3389/fcomp.2020.00032

[51] Howard Kunreuther, Nathan Novemsky, and Daniel Kahneman. 2001. Making low probabilities useful. *Journal of risk and uncertainty* 23 (2001), 103–120.

[52] Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems* 16, 10 (2015), 1.

[53] Gregory Lemasurier, Gal Bejerano, Victoria Albanese, Jenna Parrillo, Holly A. Yanco, Nicholas Amerson, Rebecca Hetrick, and Elizabeth Phillips. 2021. Methods for Expressing Robot Intent for Human–Robot Collaboration in Shared Workspaces. *J. Hum.-Robot Interact.* 10, 4, Article 40 (sep 2021), 27 pages. https://doi.org/10.1145/3472223

[54] Jan Leusmann, Chao Wang, Michael Gienger, Albrecht Schmidt, and Sven Mayer. 2023. Understanding the Uncertainty Loop of Human-Robot Interaction. *arXiv preprint arXiv:2303.07889* (2023).

[55] Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. 2020. A General Framework for Uncertainty Estimation in Deep Learning. *IEEE Robotics and Automation Letters* 5, 2 (2020), 3153–3160. https://doi.org/10.1109/LRA.2020.2974682

[56] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562

[57] Bing L. Luk, David S. Cooke, Stuart Galt, Arthur A. Collie, and Sheng Chen. 2005. Intelligent legged climbing service robot for remote maintenance applications in hazardous environments. *Robotics and Autonomous Systems* 53, 2 (2005), 142–152. https://doi.org/10.1016/j.robot.2005.06.004

[58] Alan M. MacEachren, Robert E. Roth, James O'Brien, Bonan Li, Derek Swingley, and Mark Gahegan. 2012. Visual Semiotics & Uncertainty Visualization: An Empirical Study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2496–2505. https://doi.org/10.1109/TVCG.2012.279

[59] Steven Macenski, Tully Foote, Brian Gerkey, Chris Lalancette, and William Woodall. 2022. Robot Operating System 2: Design, architecture, and uses in the wild. *Science Robotics* 7, 66 (2022), eabm6074. https://doi.org/10.1126/scirobotics.abm6074

[60] Kirsten J. McCaffery, Ann Dixon, Andrew Hayen, Jesse Jansen, Sian Smith, and Judy M. Simpson. 2012. The Influence of Graphic Display Format on the Interpretations of Quantitative Risk Information among Adults with Lower Education

and Literacy: A Randomized Experimental Study. *Medical Decision Making* 32, 4 (2012), 532–544. https://doi.org/10.1177/0272989X11424926 PMID: 22074912.

[61] Ajung Moon, Maneezhay Hashmi, H. F. Machiel Van Der Loos, Elizabeth A. Croft, and Aude Billard. 2021. Design of Hesitation Gestures for Nonverbal Human-Robot Negotiation of Conflicts. *J. Hum.-Robot Interact.* 10, 3, Article 24 (jul 2021), 25 pages. https://doi.org/10.1145/3418302

[62] AJung Moon, Chris A. C. Parker, Elizabeth A. Croft, and H. F. Machiel Van der Loos. 2011. Did you see it hesitate? - Empirically grounded design of hesitation trajectories for collaborative robots. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1994–1999. https://doi.org/10.1109/IROS.2011.6094605

[63] AJung Moon, Chris A. C. Parker, Elizabeth A. Croft, and H. F. Machiel Van der Loos. 2013. Design and Impact of Hesitation Gestures during Human-Robot Resource Conflicts. *J. Hum.-Robot Interact.* 2, 3 (sep 2013), 18–40. https://doi.org/10.5898/JHRI.2.3.Moon

[64] Lace Padilla, Matthew Kay, and Jessica Hullman. 2020. Uncertainty visualization. (2020).

[65] Lace M Padilla, Ian T Ruginski, and Sarah H Creem-Regehr. 2017. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive research: principles and implications* 2 (2017), 1–16.

[66] Max Pascher, Uwe Gruenefeld, Stefan Schneegass, and Jens Gerken. 2023. How to Communicate Robot Motion Intent: A Scoping Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, New York, NY, USA, Article 409, 17 pages. https://doi.org/10.1145/3544548.3580857

[67] Ameya Patil, Gaëlle Richer, Christopher Jermaine, Dominik Moritz, and Jean-Daniel Fekete. 2023. Studying Early Decision Making with Progressive Bar Charts. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 407–417. https://doi.org/10.1109/TVCG.2022.3209426

[68] Aaron Powers and Sara Kiesler. 2006. The Advisor Robot: Tracing People's Mental Model from a Robot's Physical Attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (Salt Lake City, Utah, USA) *(HRI '06)*. ACM, New York, NY, USA, 218–225. https://doi.org/10.1145/1121241.1121280

[69] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-Makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. ACM, New York, NY, USA, 379–396. https://doi.org/10.1145/3581641.3584033

[70] Eileen Roesler. 2023. Anthropomorphic framing and failure comprehensibility influence different facets of trust towards industrial robots. *Frontiers in Robotics and AI* 10 (2023). https://doi.org/10.3389/frobt.2023.1235017

[71] E. Roesler, D. Manzey, and L. Onnasch. 2021. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics* 6, 58 (2021), eabj5425. https://doi.org/10.1126/scirobotics.abj5425

[72] Eileen Roesler, Dietrich Manzey, and Linda Onnasch. 2023. Embodiment Matters in Social HRI Research: Effectiveness of Anthropomorphism on Subjective and Objective Outcomes. *J. Hum.-Robot Interact.* 12, 1, Article 7 (feb 2023), 9 pages. https://doi.org/10.1145/3555812

[73] Eileen Roesler*, Tobias Rieger*, and Dietrich Manzey. 2022. Trust towards Human vs. Automated Agents: Using a Multidimensional Trust Questionnaire to Assess The Role of Performance, Utility, Purpose, and Transparency. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 66, 1 (2022), 2047–2051. https://doi.org/10.1177/1071181322661065

[74] Abhraneel Sarma, Shunan Guo, Jane Hoffswell, Ryan Rossi, Fan Du, Eunyee Koh, and Matthew Kay. 2023. Evaluating the Use of Uncertainty Visualisations for Imputations of Data Missing At Random in Scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 602–612. https://doi.org/10.1109/TVCG.2022.3209348

[75] Marilyn M. Schapira, Ann B. Nattinger, and Colleen A. McHorney. 2001. Frequency or Probability? A Qualitative Study of Risk Communication Formats Used in Health Care. *Medical Decision Making* 21, 6 (2001), 459–467. https://doi.org/10.1177/0272989X0102100604 PMID: 11760103.

[76] Jonathan Schwabish. 2021. *4. COMPARING CATEGORIES*. Columbia University Press, New York Chichester, West Sussex, 67–132. https://doi.org/doi:10.7312/schw19310-005

[77] Sarah Schömbs, Jorge Goncalves, and Wafa Johal. 2024. Exploring Data Agency and Autonomous Agents as Embodied Data Visualizations. arXiv:2402.04598 [cs.HC]

[78] Ron Sellers. 2013. How sliders bias survey data. *MRA's Alert* 53, 3 (2013), 56–57.

[79] Samuel Silva, Beatriz Sousa Santos, and Joaquim Madeira. 2011. Using color in visualization: A survey. *Computers & Graphics* 35, 2 (2011), 320–333. Virtual Reality in Brazil Visual Computing in Biology and Medicine Semantic 3D media and content Cultural Heritage.

[80] Meredith Skeels, Bongshin Lee, Greg Smith, and George Robertson. 2008. Revealing Uncertainty for Information Visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Napoli, Italy) *(AVI '08)*. ACM, New York, NY, USA, 376–379.

[81] Robert Soden, Laura Devendorf, Richmond Y. Wong, Lydia B. Chilton, Ann Light, and Yoko Akama. 2020. Embracing Uncertainty in HCI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3375177

[82] Maia Stiber and Chien-Ming Huang. 2021. Not All Errors Are Created Equal: Exploring Human Responses to Robot Errors with Varying Severity. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion)*. ACM, New York, NY, USA, 97–101. https://doi.org/10.1145/3395035.3425245

[83] Megan Strait, Cody Canning, and Matthias Scheutz. 2014. Let Me Tell You! Investigating the Effects of Robot Communication Strategies in Advice-Giving Situations Based on Robot Appearance, Interaction Modality and Distance. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld, Germany) *(HRI '14)*. ACM, New York, NY, USA, 479–486. https://doi.org/10.1145/2559636.2559670

[84] Daniel Szafir and Danielle Albers Szafir. 2021. Connecting Human-Robot Interaction and Data Visualization. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. ACM, New York, NY, USA, 281–292.

[85] Christoph Tauchert, Neda Mesbah, et al. 2019. Following the Robot? Investigating Users' Utilization of Advice from Robo-Advisors.. In *ICIS*.

[86] Cristen Torrey, Susan R. Fussell, and Sara Kiesler. 2013. How a robot should give advice. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 275–282. https://doi.org/10.1109/HRI.2013.6483599

[87] James Trevelyan, William R. Hamel, and Sung-Chul Kang. 2016. *Robotics in Hazardous Applications*. Springer International Publishing, Cham, 1521–1548. https://doi.org/10.1007/978-3-319-32552-1_58

[88] Rik van den Brule, Ron Dotsch, Gijsbert Bijlstra, Daniel Wigboldus, and Pim Haselager. 2014. Do Robot Performance and Behavioral Style affect Human Trust?: A Multi-Method Approach. *International Journal of Social Robotics* 6 (03 2014), 519–531. https://doi.org/10.1007/s12369-014-0231-5

[89] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (oct 2021), 39 pages. https://doi.org/10.1145/3476068

[90] Joshua Wainer, David J. Feil-Seifer, Dylan A. Shell, and Maja J. Mataric. 2007. Embodiment and Human-Robot Interaction: A Task-Based Perspective. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*. 872–877.

[91] Astrid Weiss, Ann-Kathrin Wortmeier, and Bettina Kubicek. 2021. Cobots in Industry 4.0: A Roadmap for Future Practice Studies on Human–Robot Collaboration. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 335–345. https://doi.org/10.1109/THMS.2021.3092684

[92] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzębski, Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. 2020. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging* 39, 4 (2020), 1184–1194. https://doi.org/10.1109/TMI.2019.2945514

[93] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. *CoRR* abs/1904.09612 (2019). http://arxiv.org/abs/1904.09612

[94] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. ACM, New York, NY, USA, 408–416. https://doi.org/10.1145/2909824.3020230

[95] Lixuan Zhang, Iryna Pentina, and Yuhong Fan. 2021. Who do you choose? Comparing perceptions of human vs robo-advisor in the context of financial services. *Journal of Services Marketing* 35, 5 (2021), 634–646.

[96] Allan Zhou, Dylan Hadfield-Menell, Anusha Nagabandi, and Anca D. Dragan. 2017. Expressive Robot Motion Timing. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. ACM, New York, NY, USA, 22–31.

[97] Tian Zhou, Jackie S. Cha, Glebys Gonzalez, Juan P. Wachs, Chandru P. Sundaram, and Denny Yu. 2020. Multimodal Physiological Signals for Workload Prediction in Robot-Assisted Surgery. *J. Hum.-Robot Interact.* 9, 2, Article 12 (jan 2020), 26 pages. https://doi.org/10.1145/3368589

[98] Brian J. Zikmund-Fisher, Holly O. Witteman, Mark Dickson, Andrea Fuhrel-Forbis, Valerie C. Kahn, Nicole L. Exe, Melissa Valerio, Lisa G. Holtzman, Laura D. Scherer, and Angela Fagerlin. 2014. Blocks, Ovals, or People? Icon Type Affects Risk Perceptions and Recall of Pictographs. *Medical Decision Making* 34, 4 (2014), 443–453. https://doi.org/10.1177/0272989X13511706 PMID: 24246564.

6.3 Chapter Reflection

Addressing RQ1, this chapter draws on information visualisation and human-robot interaction research to inform the design of uncertainty communication, reinforcing user-facing risk communication as an interdisciplinary design challenge, now extended to the uncertainty component of risk. This chapter asked how uncertainty communication can be designed in a human-centred way, and how the robot's affordances can be leveraged to achieve this, investigated here through hesitation gesture as an embodied anthropomorphic uncertainty cue, alongside icon arrays as an information-visualisation-informed alternative.

Non-trivially, both uncertainty representations investigated in this chapter proved capable of communicating uncertainty in a comprehensive manner, though with nuances that enrich our understanding of user-facing risk communication design and that are worth dissecting. Specifically, this chapter surfaces the tension between granularity and heuristic understanding of uncertainty as a fundamental design challenge. Whereas icon arrays afford precise, instance-based uncertainty representations that support deliberate risk assessments and the application of individual uncertainty thresholds, hesitation gestures are interpreted more intuitively and heuristically, with less scrutiny. This reveals an important caveat to anthropomorphism as a design mechanism for human-centred uncertainty communication.

Addressing RQ2, this tension manifests most clearly in instances of no uncertainty, when the robot expressed 100% confidence. In the icon array condition, this elicited perceptions of system overconfidence, caused individuals to change decisions more frequently as an indicator of cognitive deliberation, and manifested in lower behavioural trust in the robot's assessment. This is a double-edged finding. In practice, as AI systems rarely perform perfectly, cautious deliberation is arguably desirable, particularly in high-stakes contexts where a false negative tends to be preferable to a false positive. Such deliberation, however, comes at a cost in lower-stakes contexts, both cognitively and in terms of decision efficiency. This positions hesitation gesture as a suitable uncertainty cue, as it elicits intuitive, heuristic understanding by mapping uncertainty onto a known human-human interaction script, rendering uncertainty information more accessible without inducing undue scrutiny. Quantitatively, the two uncertainty representations showed no significant differences in system trustworthiness perceptions, which we interpret as a strength of anthropomorphic uncertainty cues, in that, a hesitation gesture does not produce a perceptual downfall relative to the more established icon array. System transparency perceptions were further influenced by the robot's level of uncertainty, with the high-uncertainty condition perceived as less transparent than the low-uncertainty condition. This finding raises the question of how uncertainty representations can be enriched to support transparency in high-uncertainty instances and suggests the need for a more adaptive approach to uncertainty communication.

Addressing RQ4, this chapter re-emphasises human-centredness as the overarching design desideratum and positions anthropomorphism as a potentially valuable, but context-dependent, mechanism for uncertainty communication in robot-assisted decision-making. The findings reveal an important design trade-off: the suitability of anthropomorphism must be weighed against the level of granularity in uncertainty communication required by the deployment context and the collaboration objectives. This chapter suggests that hesitation gestures may be more suitable where intuitive, heuristic comprehension suffices, and icon arrays may be more appropriate where precise, instance-based estimates and deliberate risk assessment are required.

This page is intentionally left blank.

Chapter 7

Discussion and Future Directions

This thesis set out to explore the design of in-situ risk communication in human-AI interaction and its perceptual effects on end-users. On the quest to understand how the design of user-facing risk communication across AI systems with varying affordances can be improved (see Chapter 1.1), risk perception served as the primary lens through which we reflected on *effective* risk communication design, with system trustworthiness providing a further dimension through which our findings could be contextualised within XAI research.

In the remainder of this thesis, RQ1 and RQ2, concerned with *design* and *perceptual effects* respectively, are discussed in a synthesised manner throughout, as design decisions and perceptual outcomes are tensioned against one another and cannot be meaningfully examined in isolation, each illuminating different frictions, trade-offs, and sometimes competing design considerations. RQ3, addressing *individual* and *contextual factors*, is examined in a dedicated section (Section 7.5), and RQ4 surfaces through the implications and recommendations woven across the discussion, crystallised in the *propositions* we derive from each section. By doing so, the following discussion takes a walk across the explorative landscape of risk communication design in human-AI interaction. We begin with steps toward human-centred risk communication and the role of framing, before turning to anthropomorphism and risk embodiment. We then examine the contextual and individual factors that shape how risk information is perceived and processed, and position risk communication as a transparency cue and cognitive forcing function. Lastly, we discuss the ethical and societal considerations that follow from our work, and outline directions for future research.

7.1 The Message and the Medium: Designing Risk Communication across AI Systems

The work conducted throughout this thesis demonstrates that designing risk communication for human-AI interaction requires attention to two pillars of risk communication: the *risk message* and the *medium* through which the risk information is communicated. These pillars shape *what* should be communicated and *how* it should be communicated to the user [24, 25].

In Chapter 2.1, we identify risk as consisting of two core components - uncertainty and consequence. Accordingly, a distinctive aspect of this thesis is its explicit focus on both *uncertainty* and *consequence* as core components of the *risk message*, building on decades of risk research. By doing so, our work extends the Explainable AI (XAI) literature, which predominantly foregrounds uncertainty as the central communication strategy, while the consequences associated with AI systems remain largely unaddressed (see Chapter 2.2.2). Thus far, risk consequences frequently surface implicitly through the application context rather than through deliberate communication design. In Chapter 4, we identified this consequence agnosticism as a critical gap

in human-AI interaction. We further observed that existing approaches to uncertainty communication, predominantly driven by ML and human-AI interaction research, often fail to adopt a risk-centric lens when designing the format of uncertainty communication.

In risk literature, the risk message and the medium have often been considered separately, as distinct elements of the risk communication pipeline [24, 25]. We argue that in-situ risk communication challenges this separation, as it blurs the lines between message and medium. In in-situ risk communication, the risk information must be conveyed at the moment of the interaction and embedded within the very system that often poses the risk itself. Consequently, the AI system itself can be both the conduit and the subject of the risk message it conveys (see Chapter 3, 4, 6). To examine how risk messages can be 'delivered' in-situ, this thesis introduces the concept of *communicative affordances*. We use the concept of affordance throughout this thesis in a narrower sense than its broader application in HCI research, where it typically refers to what actions are available and possible [135]. Across our work, we conceptualise communicative affordances as delineations of the design space for risk communication, that is, the channels and encoding possibilities available to a system to communicate risk information and to a user to perceive it. This resonates with Gibson [65]'s original notion, first published in 1979, in that affordances describe what an environment or a system "offers" an agent.

This thesis surfaces that, precisely because in-situ risk communication is embedded within the very AI system through which the interaction takes place, these communicative affordances are shaped by the nature of that system. Whether the system is a *narrow classifier*, a *general-purpose LLM*, or an *embodied robot* determines the interface and interaction modalities available for communicating risk, and therefore the forms of risk encoding that are possible and appropriate within the interaction. Our work further shows that the communicative affordances do not merely carry the risk message. They can reshape the meaning of the encoding itself. In Article III, the anthropomorphic morphology of the robot platform altered the interpretive frame through which colour was understood. Green may conventionally communicate safety, correctness, or low risk in an interface, but when applied to an anthropomorphic system, it may instead draw on associations related to sickness or physical distress.

We observed that the nature of the AI system and the purpose of the interaction further shape which component of the risk message, uncertainty or consequence, can be represented meaningfully and provides the most relevant information for the user. In narrow AI contexts such as AI- or robot-assisted decision-making investigated in Chapter 3 and Chapter 6, uncertainty is quantifiable due to the nature of the AI system itself, and thus uncertainty represents a relevant concrete metric, akin to AI accuracy and AI confidence, crucial to the interaction purpose of decision-making. In more open-ended interaction contexts with LLM systems such as those investigated in Chapter 4, uncertainty is not quantifiable in the same way and thus may provide less meaningful information when communicated in isolation. Further, the consequences of risk can be diffuse, as the system itself is less task-specific and instead functions as a general-purpose technology. We believe that this breadth of possible consequence is more difficult for users to anticipate, particularly in comparison to interactions with a system built for a specific purpose, such as a classifier, where the possible consequences are comparatively bounded and more easily inferred. We argue that this relative opacity of consequence in general-purpose interaction warrants a shift in attention towards consequence as the crucial information cue that must be foregrounded. However, we propose that this does not mean that one component should be communicated to the exclusion of the other. In Chapter 4, for instance, the consequence-explicit framing of the AI risk disclaimer deliberately encodes innate uncertainty through linguistic hedging (through using terminology 'may'). Rather, the AI type and interaction context determine which component is *foregrounded as the primary risk communication ingredient*, and how the other is carried alongside it.

We propose that this thesis further demonstrates that risk communication design in human-AI interaction is inherently an interdisciplinary research challenge. At its foundation, it requires building on risk perception research and judgement under uncertainty literature, which surface the fundamental challenges in how people process and respond to risk information, including its cognitive and affective mechanisms. A crucial

contribution of this thesis is that it keeps this body of work at its theoretical core, and additionally draws on information visualisation, human-computer interaction, and human-robot interaction research to inform the design of user-facing risk communication in human-AI interaction, with framing theory informing Articles I and II, information visualisation principles guiding Article III and IV, and HRI research grounding the embodiment work (Article III-IV). We posit that this interdisciplinary approach is not only a strength of the work but an imperative condition for advancing user-facing risk communication in human-AI interaction.

7.2 Taking a Step toward Human-centred Risk Communication in Human-AI Interaction

On the path of designing effective risk communication (RQ1), we explored in-situ risk communication across varying communicative affordances, i.e. risk communication as a *text-based intervention* and *risk embodiment*. Within a given communicative affordance, specific design mechanisms determine how the risk message can be encoded and presented to the user. In Chapter 3 and Chapter 4, we utilised *framing* as the design mechanism for risk communication in a lightweight linguistic intervention suited to the respective text-based AI interaction contexts, as deployed in existing AI disclaimer practice across industry and pushed forward by policy-driven disclosure standards [52]. The use of framing as a design mechanism is grounded in a well-established body of risk research demonstrating that how a risk is described, its content and format, shapes people's risk judgements and attitudes [200]. Framing effects have been shown to influence risk-averse and risk-seeking tendencies in medical decision-making contexts [6, 127], and framing has been proposed as a means to counteract cognitive biases in risk communication [68]. In Chapter 3, we applied framing to the uncertainty-explicit component by adopting a negative frame valence that foregrounded system fallibility rather than accuracy. This shift in frame valence contributed to more appropriate perceptions of system trustworthiness, greater cognitive deliberation, and increased AI scrutiny. In Chapter 4, we applied framing to the consequence-explicit component, shifting the framing approach from *frame valence* to the *frame of reference* through which a consequence is communicated, systematically comparing system-level, user-level, and systemic perspectives, and revealing significant framing effects on risk perception (RQ2). While we did not systematically compare the frame valence in Chapter 4, the empirical findings across both works support the suitability of risk literature informed framing as a design mechanism for risk communication in human-AI interaction. Our work resonates with and extend a growing body of recent work in human-AI interaction that has begun to look more closely at how AI limitations and uncertainty can be communicated to users, shifting away from simple confidence or accuracy values in percentage format toward frequency-based representations [32], to disclosing AI limitations more explicitly [155], to more elaborate AI disclosure statements [16], and to calls for risk research informed approaches in the design of interpretable AI explanations [76]. The findings of this thesis contribute to and advance this trajectory.

Proposition 1

We propose *framing* as an accessible, lightweight design mechanism for text-based risk communication in human-AI interaction, instantiated as *frame valence* for the uncertainty component and *frame of reference* for the consequence component.

The empirical findings in Chapter 4, demonstrating that a human-interaction framing is the most effective condition in raising risk perceptions, are consequential beyond selecting a communicative default for consequence-explicit AI risk disclaimers. They surface *human-centredness* as a *core design objective for user-facing risk communication* in human-AI interaction, which informs RQ1 and RQ2. We observed that a communication framed in a way that emphasises how a risk manifests *for the user* is more affectively reso-

nant, perceived as more severe, and more personally relevant. Our empirical findings in Chapter 4 specifically point toward the pitfalls of systemic AI risk communication, suggesting psychological distance [197] as an underlying mechanism that drives its ineffectiveness, a phenomenon well documented in climate risk communication [40, 112, 124, 185]. We posit that it is therefore crucial for the design of in-situ user-facing risk communication to position AI risks as personally and experientially consequential to the user, rather than as an abstract property of the system or a diffuse societal concern, which has direct implications for both system designers and policy makers.

If human-centredness by design is the core design objective of risk communication in human-AI interaction, framing is one mechanism to achieve it, but it is bounded by what linguistic cues can accomplish. A more fundamental question emerging from our work in Chapter 4 is how risk communication can be made more human-centred through other communicative affordances provided by systems. Risk research points to affect induction as a key mechanism here. Elicited affect is the strongest predictor of perceived personal relevance [210], and risk communication practices have accordingly investigated how to induce affect through fear appeals, visual imagery, and risk narratives [33, 60, 77, 92, 131, 205]. At the same time, previous work in risk research has discussed the interplay between affect induction and the availability heuristic [92]. Effective risk communication invites individuals to envision a risk, or to draw on past experiences, in order to support its imaginability, so that it can be *felt* rather than solely cognitively processed [92, 183, 206]. Effective human-centred risk communication must therefore both elicit affect and support the mental construction of risk to make it imaginable and experienceable for the user.

Proposition 2

Human-centredness is not only the foundational distinction of HCI from fields such as software engineering [81], but a design desideratum for risk communication in human-AI interaction. We demonstrate that users may readily accept that AI systems are fallible conceptually, without internalising that this fallibility is personally relevant to them as users of the system. A risk communication design that is too abstract or systemic reinforces this distance. We propose that the design of risk communication in human-AI interaction must be *human-centred*, i.e. it must make the risk personally and experientially consequential to the user, in order for end-users to internalise the risk communicated.

7.3 Humanising Risk through Anthropomorphism

What does it mean to make risk communication human-centred by design? If we consider human-centredness by design in its most literal sense, anthropomorphism emerges as its most direct expression. Chapters 5 and 6 explore anthropomorphism as a design mechanism afforded through physical embodiment, drawing on human-robot interaction research establishing their role in shaping how people relate to and interpret agent behaviour (RQ1). Anthropomorphism plays a central role in human-robot and human-AI interaction, as it serves as a means to facilitate people's understanding of system capabilities [82, 96, 101], influences people's perceptions [156], and how individuals interact with the system, as it allows us to transfer known human-human scripts to interactions with artificial systems in order to make sense of them. Consider an approaching robot designed with a pair of eyes. The seemingly simple design decision to give a robot a set of eyes invites a cascade of inference, i.e. eyes enable sight, sight implies perception, perception suggests the robot will detect my presence, and that belief will shape how I move toward it. Anthropomorphism is reciprocal by nature, as a human perceptual tendency to attribute humanlike qualities to inanimate objects [216], and as a deliberate design strategy to activate or reinforce that perceptual tendency [156]. Our work demonstrates that anthropomorphism by design can function as a risk information encoding system, where anthropomorphic cues serve as the information-carrying signal to "humanise risks", and, by doing so, serve as a design mecha-

nism to encode risk consequences (Chapter 5) and uncertainty (Chapter 6). Rather than introducing entirely novel encoding systems, risk embodiment draws on familiar signals from human-human interaction to activate existing cognitive scripts and associations. Chapter 5 demonstrates this anthropomorphic encoding for consequence communication through the concept of *risk mirroring*, where the humanlikeness of the agent invites users to internalise risks by projecting how a risk consequence may manifest for themselves, and by doing so, activating the associative and affective processes central to personal risk perception [60, 205, 206]. We are confident that similar design mechanism may be suitable beyond embodiment, extending to graphical interface design.

Guided by RQ2, our work reveals that the concept of *risk mirroring* is effective precisely because it is *not* a one-to-one mapping of user similarities, as conducted in the VR space using doppelgänger to communicate everyday risk [2]. In Chapter 5, the risk mirroring was 'human enough' to invite self-projection without being so similar to the user that it triggered overidentification or rejection. This balance is particularly consequential in the context of risk communication, given that risk research has consistently documented an optimistic bias [66]. People tend to intellectually acknowledge a risk as real and severe, but perceive themselves as less susceptible than others, believing risks apply to people in general but not to them specifically. We observed a similar tendency in Chapter 4, in that people acknowledge the systemic risks of AI, but perceived it as more relevant *to others* than themselves. This form of *risk immunity* [66] is well established across domains, from health risk to environmental hazards, and our empirical findings (Chapter 4) indicate that AI risks may be subject to a similar tendency. It is worth noting that, beyond self-projection, Chapter 5 further revealed that risk mirroring elicited empathetic concern for others who may experience the risk, suggesting that anthropomorphic design mechanisms may offer a means to counteract the tendency observed in Chapter 4, where a systemic framing of AI risk consequences elicited significantly lower perceived severity and affective concern compared to framings that grounded the consequence in the human-AI interaction itself, with qualitative data further revealing that participants perceived systemic AI risks as more distant and less personally relevant. This suggests that anthropomorphic risk communication may be a promising design mechanism to bridge the perceived distance of systemic AI risks.

Proposition 3

Risk mirroring through an anthropomorphic agent that is recognisably human but not a direct reflection of the user may offer a path to overcoming risk immunity [66] – sufficiently proximate to trigger self-referencing and self-projection, but sufficiently distant to mitigate the defensive rejection underlying the optimistic bias, the tendency to perceive oneself as less at risk than others. At the same time, this suggests that there may be a bound to the degree of human-centredness by design, a ceiling beyond which further risk mirroring may cease to strengthen personal relevance and instead may undermine it.

Chapter 6 applies anthropomorphism as a design mechanism to uncertainty communication, using hesitation gestures [82, 130] as a known analogy from human-human interaction to encode probabilistic information in a way that activate existing cognitive scripts, an by doing so, offering a more intuitive and accessible path to numerical information (RQ1). This approach was also inspired by research in human-robot interaction, in which a robot's incapabilities were encoded through expressive motion behaviours that resembled what we characterise as anthropomorphic straining cues [101]. We argue that the non-significant differences of system trustworthiness observed in Chapter 6 suggest that an anthropomorphic uncertainty encoding performs comparably well to icon array visualisation grounded in established uncertainty InfoVis research [3, 137, 227], supporting its suitability as a design mechanism for uncertainty-explicit risk communication.

A notable limitation of anthropomorphism, however, lies in the trade-off between information granularity and comprehensibility, as observed in Chapters 5 and 6. In Chapter 5, this tension emerged in consequence

communication: conveying a meaningful and emotionally resonant risk consequence through the robot's anthropomorphic expressive features competed with preserving informational precision. The anthropomorphic encoding prioritised the affective essence of the message, but did so at the cost of accurate and granular data representation by collapsing the information content. We posit that this finding raises important design questions around how much precision can be sacrificed without compromising the communicative integrity of the representation. Chapter 6 exposes the same underlying constraint from the perspective of uncertainty communication. Although the anthropomorphic encoding supported an intuitive, heuristic understanding of uncertainty as a system characteristic, it did not afford granular uncertainty *reasoning*, thereby collapsing and thus limiting the resolution at which users could assess and reason about the risk.

Crucially, Chapter 5 contextualises the physical agent as a conduit for risk information, with the design objective of increasing perceived risk. Chapter 6, by contrast, situates risk communication within an AI-assisted decision-making context, in which users are required to draw inferences from uncertainty-explicit information in order to reach an informed decision. This latter objective necessitates a design orientation towards accurate risk assessment rather than affective salience alone. We emphasise that the tension that surfaces in these two chapters (Chapter 5- 6) maps directly onto a foundational distinction between designing for risk *assessment* and risk *perception*. Whereas risk assessment prioritises the accurate representation of risk information through precision and granularity across its uncertainty and consequence components, risk perception concerns the subjective interpretation and affective response elicited by risk communication.

Proposition 4

Anthropomorphic risk encoding draws on cognitive scripts from human-human interaction to make risks understandable and to heighten risk perception, but in doing so, it trades granularity for comprehensibility. We argue that the communicative objective of the risk communication must govern how this trade-off is resolved. Where it is designed to support a decision against a defined risk threshold, precision must be preserved. Where it aims at a holistic understanding of risk or to heighten risk perceptions, heuristic accessibility may take precedence. This maps onto a broader distinction in risk research between what is designed to enable *risk assessment* and what is designed to *shape risk perception* [221], a distinction that is critical for the design of human-centred risk communication in human-AI interaction.

Another design consideration that emerged across our work in Chapters 4 and 5 concerns the balance between affective resonance and avoidance (RQ1, RQ2). Human-centred risk communication necessarily draws on affect to achieve personal relevance, but affect induction carries risks of its own. Chapter 5 shows that anthropomorphic risk mirroring heightens vividness, negative affective responses, including fear, worry, and concern, and the perceived severity of the risk. Yet, it also points to a fine line between affectively resonant risk communication and communication experienced as distressing, provocative, or excessively fear-inducing, potentially eliciting avoidance rather than calibrated concern. Such avoidance can paradoxically reduce personal risk perception and cause individuals to disengage, a phenomenon previously observed in the communication of smoking-related health risks [28]. This tension is not confined to risk embodiment. In Chapter 4, a systemic AI risk framing was perceived as dystopian and doom-laden, producing a comparable dynamic of disengagement by signalling the scale of the AI risk. Drawing on these empirical insights, we suggest that affect-inducing cues in risk communication for human-AI interaction must be calibrated to motivate concern without escalating into fear or distress that prompts avoidance and undermines the intended communicative purpose.

7.4 Why XAI Needs Risk - Risk Communication as a Cognitive Forcing Function

AI systems increasingly support, augment, and take over decisions, tasks, and reasoning in everyday life, raising concerns related to cognitive offloading [64], the reinforcement of cognitive biases through AI systems [69], and overreliance (e.g. [29, 103, 141, 226]). Recent theoretical work has conceptualised AI as a cognitive extension of human cognition, coined System 0 thinking [37, 38], with more provocative formulations going so far as to describe human cognition in AI interaction as a form of cognitive surrender [169, 222].

XAI approaches have sought to address these concerns by making systems more *transparent* as a means of supporting appropriate system trust. Traditionally, explanations in human-AI interaction have been designed to provide access to model internals, for instance through feature contribution or importance [213], to support people's understanding of the AI model. Uncertainty communication has similarly been positioned as a lightweight explanation approach [29, 152, 203, 225]. These strategies, however, frequently fall short in supporting appropriate system trust [213] and users tend to not engage with explanations sufficiently to support critical appraisal [128]. More recent work in human-AI interaction has accordingly identified the need to support active critical thinking in human-AI interaction [29, 31], giving rise to paradigms such as Evaluative AI [128], in which the AI presents decision candidates that a user can narrow down and then evaluate with the support of an AI, which provides evidence for or against the candidate. However, we argue that these approaches still rest on the implicit assumption that an explanation or the presence of candidatures alone provides sufficient incentive for users to engage critically with the systems and its outputs. But why would it? In everyday life, people rarely consult ingredient lists, read product manuals, or examine the chemical composition of cleaning supplies before use. This is not to suggest that XAI approaches lack merit; there are clearly contexts in which access to model internals is both appropriate and valuable. Rather, we argue that opacity may not always be the central obstacle. XAI has predominantly asked what users can know about the system and to what degree 'the black-box' should be opened. We suggest that an equally important question is what information will motivate users to think carefully about the system and its outputs. We argue that risk information may be what provides that motivation, as it offers a crucial, decision- and system-relevant information and makes system limitations consequential to the user, and by doing so, providing a reason to allocate cognitive effort.

Another stream in XAI research moves towards cognitive forcing functions, which describes an approach designed to elicit cognitive deliberation and appeal to type 2 thinking [29]. Approaches to achieve such cognitive forcing often entail the withholding of an AI output to prompt deliberation from the user to attenuate blind acceptance, for instance, by delaying an AI output [142] or by asking the user to first provide a decision prior to the AI output [29]. We position risk communication as an intervention that has the potential to operate in a manner akin to a cognitive forcing function, but without necessarily imposing the same degree of procedural friction. Rather than withholding the AI output or requiring an additional user action, we propose that risk communication can introduce informational and motivational friction by foregrounding the potential of system fallibility and the associated consequences.

Across our empirical work, we observed that risk communication as an intervention may operate at two levels (RQ2): it emerges as a *transparency cue* that shapes how users perceive system fallibility (Chapters 3 and 4), and as a mechanism that elicits active deliberation at the point of interaction (Chapters 3 and 6). Chapter 3 shows that risk communication (uncertainty and consequence explicit) functions as a *transparency cue* that significantly enhances perceived system transparency, particularly following erroneous AI behaviour, and does so across decision-stakes. It also prompted users to instil vigilance and to proactively anticipate AI fallibility, hinging toward a re-calibration of a user's mental model of the system and its failability, and to scrutinise AI outputs more deliberately. Importantly, it did so without eroding task performance or AI utility perceptions, which reduces concerns related to whether foregrounding AI risk pre-emptively penalises a

system, or re-enforces seeded doubt when the AI systems errs to a degree that leaves system trustworthiness unrecoverable or induce algorithmic aversion [143]. Chapter 4 reinforces this notion, in that the presence of a consequence-explicit AI risk disclaimer shifted system trustworthiness and risk perceptions, and that disclaimers framed through the lens of system capabilities and human-interaction were interpreted as transparency cues. Chapter 6 extends our argument further, showing that uncertainty communication, informed by risk and information visualisation research, elicited cognitive deliberation, as indicated by changes in users' decision-making behaviour and the application of risk thresholds, which points to an active engagement in risk assessment beyond risk perception. We therefore position risk communication as a potential cognitive forcing function, a mechanism that supports system transparency and triggers deliberate appraisal.

Our results further suggest that risk communication can contribute to the desired criteria of explanations identified by Wang and Yin [212], namely supporting users' understanding of system capabilities and limitations, increasing uncertainty awareness, and improving trust calibration - provided that the communication is designed through a risk-theoretical lens. We note, however, that this potential is contingent on the context in which risk communication is deployed. Consistent with the context in which Evaluative AI has been advocated [128], our findings suggest that risk communication as a cognitive forcing function is most appropriate in high-stakes contexts, where the deliberative costs it imposes are warranted by the potential consequences (Chapter 3 and Chapter 6). We re-emphasise, further, the discussion in Section 7.3, where we argued that cognitive deliberation may be attenuated if risk communication design prioritises heuristic processing over deliberate appraisal.

Proposition 5

We propose that risk communication in human-AI interaction can function as a *cognitive forcing function*, an intervention that elicits deliberate appraisal of AI output and fosters more appropriate system trustworthiness perceptions. Ideally, risk communication is both uncertainty and consequence explicit; more fundamentally, however, it must be designed through a risk-theoretical lens for either component to realise this potential.

7.5 One Size Does NOT Fit All - Contextual and Individual Influencing Factors

The empirical findings across this thesis demonstrate that users' perceptions of risk communication are shaped not only by how risk is communicated, but also by contextual and individual factors. Guided by RQ3, the following discussion reviews our findings, considers their importance for human–AI interaction, and situates them within the broader risk communication literature. In doing so, we highlight the need to design risk communication in ways that are appropriate to the *interaction context* and effective for the *individual user*.

Chapter 3 systematically investigates decision-stakes as an important contextual factor for human-AI interaction. Decision-stakes present the *situational risk* associated with the context in which the human-AI interaction is embedded [189], and are considered crucial given that the fallibility of AI systems is particularly consequential in deployment contexts or tasks that carry high situational risks [78, 100]. For example, an inaccurate music recommendation carries less consequential weight compared to an inaccurate health treatment decision. Despite its importance, research on decision-stakes has been scarce, not least because it is difficult to operationalise situational risk systematically and ethically [78, 83]. Prior work in human-AI interaction has frequently examined stakes in silos, focusing either on high-stakes [32, 76, 90] or low-stakes contexts [96, 140], and where stakes have been compared, findings on the role of situational risk in AI-assisted decision-making remained largely inconsistent [7, 55, 78, 141, 225]. Chapter 3 contributes to resolving this inconsistency, providing an important step forward in human-AI interaction research. We show that decision-

stakes significantly moderate the effects of perception calibration strategies on system trustworthiness. Critically, we argue that this surfaces decision-stakes as a determinant not only of user perception but of the *appropriateness* of the calibration strategy itself, and in doing so, also reveals crucial drawbacks of AI accuracy communication as an established trust calibration approach [29, 152, 225]. Our empirical findings in Chapter 3 allow us to position risk communication as a strategy without such potentially detrimental pitfalls under high decision-stakes, with a directional trend toward more appropriate calibration under high-stakes. However, in low-stakes contexts, it also imposed cognitive costs where additional deliberation may not have been necessary. In line, Chapter 6 further suggests that anthropomorphically encoded risk information supports more intuitive, heuristic-driven uncertainty assessment, which may be more suitable in lower-stakes contexts, whereas granular uncertainty representations appear better suited to high-stakes contexts where deliberate appraisal is warranted. In summary, our findings suggest that risk communication should be designed and deployed in relation to the associated stakes of the decision rather than applied uniformly across deployment contexts, and suggests risk communication as a suitable mechanism for high-risk AI systems, as defined by the EU AI Acts [53]. We call for closer collaboration between HCI and policy research to ensure that transparency obligations are grounded in empirical findings on how end users actually process and respond to risk information in specific contexts.

Individual differences represent a further important determinant of how risk information is perceived and processed. A plethora of work has established that individual factors shape risk perception, from personality traits and gender to broader values and beliefs [25, 178], and must be considered when designing risk communication. Motivated by the increasingly political climate surrounding AI and prior risk research establishing political orientation as a significant predictor of risk perception [39], Chapter 4 surfaces *political orientation* as an important individual lens through which users appraise AI risk information, and by extension, the AI system associated with it. From our work, we observe political orientation both as a significant main effect on risk perception of AI systems accompanied with AI risk disclaimers and, more importantly, as a significant moderator of the framing of consequence-explicit AI risk disclaimers on risk perceptions, which revealed crucial insights for the design of risk communication, as discussed in Chapter 7.2. Our findings provide a nuanced picture of the role of political orientation in AI risk perception, partially contradicting traditional accounts of political orientation and attitudes toward new technologies [39], while supporting more recent findings on political orientations and AI specifically [144, 174]. Given the increasingly politicised landscape of AI, these findings are consequential for understanding how individual values shape the reception of user-facing risk communication. The qualitative findings deepen the picture further, as they point to underlying value-oriented evaluations of AI risk beyond political orientation, revealing a divide in community-minded and egalitarian orientations versus individualistic orientations. Given that political orientation correlates with distinct moral foundations [8, 73], our work positions moral foundations theory [75] as a promising framework to examine individual factors influencing AI risk perception at a more granular level, and the design of effective risk communication respectively. Chapter 5, albeit not a primary focus of the empirical work, surfaces individuals' *lived experience* as an individual lens through which risks were appraised. We observed that participants drew on past encounters and personal histories when interpreting the risk communicated through 'risk mirroring', contemplating about people they knew who had been affected by similar risks and re-evaluating prior experiences in light of the risk information presented. Interestingly, lived experiences either amplify risk perception or, conversely, produce desensitisation effects. Our findings contribute to prior work showing that personal experience with a specific risk [25], such as prior exposure to natural hazards [92], leads to increased risk perception.

Looking at RQ3 from a broader perspective, our findings on contextual and individual factors surface an important tension in the design of risk communication. When an AI system is developed for a specific use case, its stakes are often defined by the domain and interaction context, which makes the selection of an appropriate communication strategy relatively straightforward. General-purpose AI systems present a different challenge, as the same system may be used across low- and high-stakes domains. We argue that, in the latter

case, a single fixed intervention may be insufficient, which suggests the need for more adaptive approaches to risk communication and, potentially, more adaptive policy requirements. A similar distinction emerges from our findings on individual differences, as we posit that a risk communication design that is less contingent on individual differences may be more suitable when risk communication is designed for broad and heterogeneous public audiences. However, when risk communication is designed purposefully in-situ to a *specific* end user interacting with an AI system, the design objective may instead be to provide the communication that is most effective for that particular individual. This creates a tension between broadly applicable and individually targeted risk communication. We introduce these considerations here and examine their implications more comprehensively in Chapter 7.8.3, illuminating two important avenues of future research on adaptive risk communication policies.

Our work contributes to the HCI community by positioning contextual and individual factors not as peripheral sources of noise, but as constitutive dimensions of risk communication design. In doing so, our work further specifies the question of *how* risk communication should be designed by decomposing it into *under which conditions* and *for whom* it should be communicated. This calls for risk communication to be evaluated as a situated intervention rather than as a fixed interface element. Our findings also carry implications for AI policy and regulation. Transparency obligations should not assume that standardised risk information will be equally meaningful or effective across systems, contexts, and user groups. For general-purpose AI systems in particular, policies may need to account for the context of use and the level of situational risk it carries. At the same time, adapting risk communication to individual users raises important HCI concerns around privacy, fairness, autonomy, and manipulation, which we reflect upon further in Section 7.7. Our findings on political orientation should therefore not be interpreted as a pathway to political profiling or targeted persuasion, but as grounds for evaluating user-facing risk communication across heterogeneous audiences.

Proposition 6

Our work points to an important consideration for user-facing, in-situ risk communication in human-AI interaction: risk information should not be designed in a vacuum, but in relation to decision-stakes as an important *contextual factor* (Chapter 3 and Chapter 6). Nor should it be designed as a one-size-fits-all intervention, as Chapters 4–5 demonstrate the need to account for the end user and the *individual lens* through which risk information is appraised. Effective risk communication must therefore be designed with respect to the interaction context and the individual user, rather than applied as a uniform intervention across settings and audiences.

7.6 Risk Communication as a Cue: From Relevance to Utilisation

Positioning our findings within the Trustworthiness Assessment Model (TrAM) proposed by Schlicker et al. [165] offers a theoretically grounded lens through which the function of risk communication in human-AI interaction can be examined. We believe that the TrAM is of particular relevance to this thesis for three reasons. First, it focuses on *perceptions* of system trustworthiness, which aligns with the focus of the empirical work presented across Chapters 3, 4, and 6. Second, it conceptualises trustworthiness perceptions as being formed through *cues*, understood as *pieces of information* that systems provide and users draw on when assessing a system. This provides a direct conceptual connection to risk communication, which can itself be understood as a form of information provided to support users' assessments. Third, the TrAM situates the design and communication of such cues within the broader objective of *designing trustworthy AI systems*. The central proposition of the TrAM is that a system's trustworthiness, as a latent construct, is not observable in itself but must be inferred by users through these cues. According to the model, the accuracy of this inference depends on four conditions: the *availability* and *relevance* of cues at the system level, and their *detection* and

utilisation at the user level.

Our findings suggest that risk communication may be well positioned to support more appropriate trustworthiness assessments, as it has the potential to operate across the four conditions proposed by the TrAM. However, we do not claim to have empirically tested these conditions or to provide systematic evidence for them. System trustworthiness was not consistently measured across the studies presented in this thesis, and the TrAM did not guide their design or analysis, as it was developed concurrently with this work. We therefore use the TrAM as a post-hoc conceptual lens through which to reflect on the role that risk communication may play in trustworthiness assessments. The arguments that follow should be understood as theoretical interpretations and propositions derived from patterns observed across our work, rather than as a direct empirical validation of the model.

With respect to cue *relevance*, the TrAM proposes that a cue is relevant to the extent that it is indicative of a system's actual trustworthiness. We propose that risk information may satisfy this criterion due to its constitutive components, uncertainty and consequence. The probabilistic nature of AI systems represents a highly relevant system characteristic, and consequence, though not a direct property of the system itself, communicates a potential outcome of relying on it - whether it is a decision, recommendation, or action. We therefore argue that risk communication may serve as a highly relevant cue in the TrAM sense. With respect to cue *availability*, the TrAM posits that cues must be exposed to the user. That dimension emphasises that system designers need to actively surface the cues that are actually relevant, a design question in essence. Viewed through this lens, the work presented in this thesis illustrates several ways in which risk information can be made available across AI systems with different communicative affordances, ranging from lightweight a priori textual interventions in Chapter 3, to AI risk disclaimers in Chapter 4, and situated risk communication embodied in robot platforms in Chapters 5 and 6.

On the user's side, cue *detection* describes the fact that available cues must be *recognised* by an end user. The TrAM proposes that detection, beyond system design per se, depends on human factors such as attention, emotion, and situation awareness, which guide what a user notices in the first place. Our findings can be interpreted as suggesting that well-designed risk communication may support detection by engaging affective processing and rendering information personally relevant (see Chapter 4 and Chapter 5). It is important to note that affective responses were measured here as part of our risk perception outcomes rather than as a direct measure of trustworthiness, so these findings speak to detection in the broader sense proposed by the TrAM. We nevertheless consider this meaningful, since it allows us to propose that risks designed as affectively-inducing and personally relevant may be more likely to draw attention and engage cognitive deliberation than abstract model parameters (see Chapter 7.2 and Chapter 7.4).

Cue *utilisation* is where we believe risk communication may offer its most crucial contribution. The TrAM describes *utilisation* as the process by which a detected cue is interpreted and weighted in the trustworthiness assessment, and proposes that this process depends on, for instance, a user's domain knowledge, AI literacy, and susceptibility to cognitive biases. This is where a user's cue processing is most vulnerable to distortion. Many factors, such as numeracy, may lead to inaccurate interpretation of probabilistic information, and cognitive biases may skew the weighting of cues. In Chapter 4, we observed that the presence of a consequence-explicit AI risk disclaimer decreased trustworthiness perceptions relative to no disclaimer, which suggests that risk communication as a cue was incorporated into users' trustworthiness assessments. In Chapter 3, we observed that AI accuracy information led to an inaccurate utilisation of the information cue (inflated trustworthiness under high stakes), while risk communication avoided this distortion, which indicates that risk information as a cue supported more appropriate utilisation. Our findings further suggest that political orientation as an individual factor may shape how risk information is interpreted and weighted. In Chapter 4, this moderation manifested at the level of risk perception rather than at the level of system trustworthiness, a consideration we discuss in detail below. We argue that risk communication - if designed through a risk theoretical lens and with human-centredness by design in mind - can support appropriate cue

utilisation, as an intervention that translates system uncertainty and consequence into representational forms that are accessible and interpretable by users.

Proposition 7

We posit that user-facing, in-situ risk communication may help reduce the gulf between actual and perceived system trustworthiness. On the system side, risk information can be designed as a cue that is both relevant and available. On the user's side, it has the potential to support cue detection and more appropriate utilisation. Notably so, these propositions emerge from a post-hoc interpretation of patterns across the thesis and require direct empirical investigation [165, 166].

The observed findings in Chapter 4 can also be interpreted in relation to the *individual standards* dimension proposed by the TrAM [165, 166], which suggests that users hold individual standards that determine what makes a system trustworthy for them personally, described by the authors as an internal checklist against which information cues are weighted. We do not consider our findings to constitute a direct empirical test of this dimension. It is important to be transparent that the individual differences observed in Chapter 4 manifested at the level of risk perception rather than system trustworthiness, the primary construct the TrAM addresses. Political orientation investigated in Chapter 4 did not moderate the effects of risk communication framing on system trustworthiness, but the perceived severity of the AI consequence. Nevertheless, the finding that the same risk cue was appraised differently depending on a person's political orientation is consistent with the broader idea that individuals apply different standards when interpreting and weighting information. That this moderation surfaced in risk perception rather than trustworthiness leads to another important discussion.

The finding that risk consequence framing did not shift system trustworthiness perceptions in Chapter 4 may reflect the fact that the comparative manipulation, that is, the frame of reference through which consequences were communicated, did not entail a systematic value alignment comparison akin to moral reframing [58], where an alignment between information and a person's moral orientations increases persuasion. Individual standards may therefore have been insufficiently triggered to produce effects beyond risk perception. This interpretation is necessarily tentative, and warrants future investigation by the HCI community.

More importantly so, Chapter 4 showed that the presence of an AI risk disclaimer shifted both system trustworthiness and risk perceptions, whereas the framing of that disclaimer, and its moderation by political orientation, shifted risk perceptions alone. We believe that this points to system trustworthiness and risk perception as *complementary* outcome measures. This observation speaks to a critique recently raised by Kahr and Bernstein [89], who argue that trust has become the default construct in human-AI interaction research to evaluate transparent and trustworthy AI, despite unresolved questions around what it represents, how it should be operationalised, and whether existing measures capture it reliably. The authors further emphasise that researchers may need to rethink what constitutes a "successful" human-AI interaction beyond trust. Our findings offer a concrete illustration of this concern. Had Chapter 4 measured solely system trustworthiness, both the framing effect and the political orientation moderation would have gone undetected, since neither manifested beyond risk perception. We therefore propose that research on AI explanations and interventions should extend their investigation to include risk perceptions as a crucial construct in XAI, since this may offer a more complete and nuanced picture of their perceptual effects and and underlying cognitive mechanisms. We call for complementary constructs to trust, and position risk perception as one such construct.

7.7 Societal and Ethical Considerations

> *"What has been reduced substantially in the past century is the time from first use to widespread integration into our social system. The techniques for societal diffusion of a new technology and its subsequent exploitation are now highly developed."*
>
> Starr (1969)

This thesis examines the design and understanding of in-situ risk communication in human-AI interaction, a topic that is inextricably linked to real-world risks and contexts characterised by uncertainty and vulnerability. While our empirical work operates at the level of the *individual* interaction, we argue that its implications extend well beyond. AI systems are adopted at scale, and risks that manifest at the individual level do not remain there; they cascade and diffuse across society in ways that warrant careful reflection. We therefore consider it imperative to shed light on the broader ethical and societal implications of this work. In the following, we address the normative question of what appropriate risk perception in human-AI interaction ought to achieve, the ethical tensions inherent in risk communication design, the potential for manipulation and adversarial misuse, and the question of who carries the burden of AI risk.

In Chapter 4, we discuss that the *objective* of risk communication is, at its core, a normative question, and therefore raises crucial considerations around how to define the *effectiveness* of risk communication in the context of human-AI interaction. Risk communication is not solely concerned with the mere objective to transfer information between sender and receiver, to complete gaps or correct discrepancies in an individual's mental model about a risk [9], but often persuasive in nature. The motivation for this, figuratively speaking, *sculpting* of risk perception through risk communication design, i.e. to shift its perceptual magnitude and direction, is commonly justified by the assumption that individuals may lack the necessary resources, are subject to constraints or cognitive biases that prevent them from forming appropriate perceptions and deriving respective protective actions [190]. Risk communication therefore moves beyond the technical question of *how* information can be transmitted accurately. It also raises the normative questions of *what* should be communicated and *how* it should be communicated to produce a particular perceptual or behavioural *response*, which determines its design.

The question of *what* should be communicated, i.e. which risk needs to be communicated in the first place, is a relatively straightforward one, with advancing efforts in formulating AI risk taxonomies from policy and regulation works that can offer guidance for human-AI researchers (e.g. [176, 202, 217]), and which informed our work (Chapter 4). The question of whether risk communication should be designed to *produce* a particular *response*, however, is more nuanced and requires careful reflection. In this thesis, we have implicitly treated an *increase* in risk perception as a design objective and as a measure of effectiveness. This objective may conflict with traditional HCI objectives, in that interventions that increase risk salience and risk perception may reduce cognitive deference to an AI system, or lead users to disengage from it altogether. This stands in contrast to the conventional assumption in HCI, which foregrounds seamless interaction between human and computing system, and to success metrics prevalent in human-AI interaction research and industry, which tend to prioritise AI engagement and adoption.

We do not hold the position that algorithmic aversion is a desirable outcome per se, which we note in Chapter 3, and which motivated the initial investigation into risk communication. We do, however, argue that the 'success' metrics frequently applied in HCI and human-AI interaction research warrant careful reconsideration in light of downstream risks. This means walking a tightrope between risk perception as a step toward mitigating AI-related harm, and preserving the benefits that AI systems can offer — a tension discussed in Chapter 3 in relation to the trade-off between perceptions of AI utility and risk. The findings from

Chapter 3 suggest that this balance is, indeed, desirable and achievable. This consideration reflects concerns around over-reliance (e.g. [13, 29, 103, 226]), cognitive offloading [37, 38], and cognitive surrender [169, 222], and echoes current efforts in XAI advocating for *calibrating appropriate* trust in human-AI interaction, in that perception of system trustworthiness should match system capabilities [165, 166]. Notably so, *appropriateness* itself presupposes some gold standard of a 'right level' of risk perception, which, we believe, entails functions of normative risk-benefit trade-offs. In the context of AI risk, this raises important downstream questions. What does appropriateness mean in the context of risk perception? And who defines that standard? It may be set top-down by policy or regulation, or bottom-up by individuals who hold their own risk-taking propensities and values, which reflects that the question of risk communication outcome is not solely a scientific and intellectual debate, but influenced by potentially divergent stakeholder interests and priorities.

In Chapter 4, we examine political orientation as an important moderator and explore its design implications for risk communication. In the wake of this, we discuss its ethical considerations and potential adverse impacts related to privacy concerns and user-aware and adaptive AI risk communication in ways that may downplay risks or amplify less relevant concerns for the end user. The discussion illuminates ethical concerns that necessitate re-emphasis and contextual broadening across this thesis. We argue that these concerns are not confined to the specific case of political orientation, but surface in any inference-based risk communication design, where inferences are decoded to inform and manipulate the cue relevance and utilisation [165, 166] of risk information, and could therefore be used to steer perception accordingly. These inferences can be drawn from dispositional user characteristics, or stem from the human-AI interaction itself, e.g. from the situational risk as a contextual factor (see Chapter 3, where we posit to adapt the risk communication strategy to the decision-stakes to avoid unwarranted cognitive costs in low stakes or incentivise cognitive deliberation in high stakes), or from the conversational content (as in Chapter 3, where risk communication was designed to be content-aware per default to presents AI risks relevant to the specific human-AI interaction). These considerations reveal intricate ethical concerns around consent, privacy, and the potential for manipulation in inference-based risk communication design in human-AI interaction. We believe that these concerns are constitutive of risk communication design, as they hinge upon the innate relationship of risk communication with persuasion. Risk communication is therefore a double-edged sword. While the findings of this thesis can be drawn upon to design effective risk communication to mitigate real-world AI harm, the same results can be exploited, in that risk communication can be engineered to nudge individuals in a certain direction, to downplay risks or render individuals less perceptive to them, touching upon (in itself) an important AI risks around malicious actors and use [176, 217]. We therefore urge human-AI interaction researchers and AI policymakers to treat the ethical dimensions and potential adverse effects of risk communication design as a core consideration.

AI systems are reshaping the way people live, interact, seek information, and make decisions at a scale that is historically unprecedented. The societal diffusion of technologies has never occurred at this speed, and we therefore argue that the boundary between risks arising at the system level and the systemic risks that emerge from their aggregated and cascading effects is becoming increasingly difficult to draw. Non-factual information generated by AI aggregates to an erosion of information quality across society, harmful misrepresentations accumulate to re-enforcing discrimination [218]. Finding a balance between the risks and benefits in the face of rapid technological advancement has long been recognised as a serious dilemma for society [62], and the scale of AI development and deployment makes this dilemma more urgent. We therefore urge the research community to recognise that failing to establish an understanding of how to communicate in-situ risks in human-AI interaction, and overlooking the need to integrate risk theory into existing intervention approaches, may have systemic implications.

We further note that, while we narrowed our lens to individuals using AI systems, AI risks are not confined to a one-to-one interaction. We argue that it is therefore essential to ask *who carries the burden* and *who bears the social costs of AI risks*. Risk may not manifest for the user directly interacting with the system, but

for third parties affected by the user's AI-mediated decisions or outputs. On the individual level, an end-user may evaluate a given risk as irrelevant to their own value system and therefore perceive it as acceptable, yet the application of the AI-generated output may produce harm for others. This distributional dynamic is exemplified in Chapter 4, in which an AI risk disclaimer warned the user about potential discriminatory language, while the consequences of that risk may extend to the individuals affected by the AI-generated output. Notably so, an AI risk may, on an individual level, be deemed acceptable. However, it carries social costs that are borne by those *outside* the human-AI interaction. This raises a broader question of fairness and risk distribution beyond individual risk perception. Our findings derived from the empirical investigation examined in Chapter 4 made us realise that simply falling back to communicating systemic risk consequences fails to elicit desired risk responses. We therefore encourage HCI researchers and AI designers to consider risk as a distributional concern that extends across those affected by AI-mediated decisions. We further emphasise that individual-level risk communication cannot be solely relied upon as a mechanism for mitigating distributional harms. We also call on policymakers to develop regulatory frameworks that explicitly account for the relationship between individual-level human-AI interaction and its aggregated societal consequences.

This discussion alludes to another interesting consideration, which connects to one of the most fundamental debates in risk perception research dating back to 1969, namely that between voluntary and involuntary risk [188]. Whereas a voluntary risk describes a circumstance in which a decision or an action emerges from an individual's own risk-benefit trade-off, involuntary risks are imposed, for instance by governing bodies that define what is considered an acceptable risk. How do involuntary risks take shape in the context of AI risks? What risks related to AI are imposed on individuals, and who is the source of such imposition? In the context of AI, involuntary risk may extend beyond regulatory imposition, or risks imposed by AI providers. For instance, in the race toward AI adoption observable across society today, end users may face a form of involuntary risk exposure driven by broader social, institutional, and economic pressure toward technological adoption and performance in a capitalist system.

7.8 Future Directions and Limitations

The empirical findings across Chapters 3 to 6 and the synthesised discussion guided by our research questions give rise to a landscape of future research avenues that we outline here as a prospective agenda for human-AI interaction research. These directions span longitudinal investigations, behavioural outcomes, adaptive policy, bi-directional risk communication, and overlooked determinants, and we posit that each represents a meaningful pathway toward a deeper understanding of risk communication (design) in human-AI interaction. We believe that this thesis can provide a stepping stone toward a future in which the understanding and design of risk communication are not an afterthought in human-AI interaction, but a core concern.

7.8.1 Longitudinal Studies and Real-world Settings

Experimental decisions always involve trade-offs, one of which is choosing the methodology and experimental setting most suitable for addressing a research question. Across this thesis, we conducted user studies to gain empirical insights both online through crowdsourcing platforms (Chapters 3 and 4) and in laboratory environments (Chapters 5 and 6). We note that those types of investigations, while providing many benefits such as a controlled environment, also have limitations. We discuss these limitations in Chapter 1.2 and demonstrate how we counteracted limitations such as novelty effects and concerns about ecological validity.

Importantly, in the context of risk communication, the interactions across this thesis were confined to short-term temporal windows that, in many cases, reflect solely a snapshot of human-AI interaction. However, we emphasise that user interactions with AI systems, and therefore potential touchpoints with user-facing

risk communication, increasingly move toward habitual integration into everyday life and span across a variety of settings, with AI systems adopted as general-purpose technologies. We therefore propose the need for two complementary methodological approaches, i.e. a) longitudinal studies, accompanied with b) field evaluations, following more in-the-wild approaches to draw implications on long-term effects of in-situ risk communication embedded in real-world settings, potentially even integrated into real-world deployments. While this would increase ecological validity, it would also allow future research to investigate important research questions that emerged from our work (Chapter 3), and which have been raised in the human-automation literature [78], and which we re-emphasise as important future avenues for human-AI interaction researchers. Among them the question of the degree to which risk communication can be *sustained* over repeated long-term interactions and AI use. Prolonged exposure to risk information may produce habituation effects, whereby the communicative signal loses salience or relevance over time [78]. This, in turn, may foster a broader pattern of risk normalisation, a process in which increased awareness of risk does not convert into heightened risk perception, which has been documented in the context of climate change [119]. Longitudinal studies would allow future work to assess whether and how habituation to risk information develops over the course of long-term human-AI interaction, and therefore gain insights into the temporal robustness of in-situ user facing risk communication. Field deployments, on the other hand, would allow future research to e.g. investigate situations in which risk communication competes with real-world AI benefits. The approach of longitudinal studies combined with field deployments has gained increasing attention and application over the last decade in the context of mobile HCI [98], for instance, related to privacy concerns [120], and, we presume, is crucial for future pathways of in-situ AI risk communication, especially concerning the use of AI systems such as AI-assisted support tools (Chapter 3), LLMs (Chapter 4), but also collaborative robot platforms in industrial contexts (Chapter 6).

7.8.2 Behavioural Outcomes of Risk Communication - Beyond Reliance

While our work predominantly focuses on perceptual outcomes of risk communication in human-AI interaction, we strongly emphasise that we acknowledge the importance of reliance research. Importantly so, our findings give us confidence that risk communication is a promising mechanism for eliciting appropriate behavioural outcomes in human-AI interaction. We encourage future work to shift dependent measures toward said behaviours. Our work suggests that risk communication can lead to AI scrutiny, as participants noted that they would be tempted to more acutely double-check AI outputs, or verify them using other sources (Chapters 3, 4, 6). Further, the risk communication examined in Chapter 5 was interpreted as a cue for protective mitigation action by some participants. These observations open a plethora of questions that we believe are crucial for future research, and at the forefront of it: what behavioural outcomes result from risk communication in human-AI interaction?

We propose that reliance, as it is currently understood in XAI research, may be too narrow a lens through which to examine these outcomes, and we invite future research to advance a more comprehensive understanding of behavioural responses to AI risk communication through the lens of risk theory. In risk and human factors research, typical behavioural outcomes would span across *responding* to an alarm as well as *preparing* for an adverse event. To progress in this realm, we believe future research working on XAI and trustworthy AI needs to draw on existing frameworks provided by risk research, among them Protective Motivation Theory (PMT) [159] (see Floyd et al. [63] for a meta-analysis of literature on PMT and adjacent frameworks), which may also serve to develop risk models for human-AI interaction. For instance, PMT proposes that in addition to a risk appraisal, individuals also carry out a so-termed *coping appraisal* that draws on concepts such as response efficacy and self-efficacy [150], albeit we would like to note that this framework falls short in that it postulates primarily intentions of behaviour. In line with this, we believe that research can draw on the theory of "inoculation" from psychology, which has been adopted specifically in misinformation research (see e.g. [115, 172, 195]). We believe that these frameworks will serve to provide relevant concepts and

guidance in human-AI interaction. We encourage future research to dissect more closely how behavioural outcomes of risk communication could materialise to mitigate AI risks. We assume that an overarching structure could be *preventive* strategies, as well as *reactive* outcomes. As an example, preventive strategies may manifest in prompting behaviour, for instance setting constraints on AI use proactively in response to risk communication, whereas reactive outcomes may involve the verification of AI output posterior to the interaction, which more traditionally would fall under typical reliance measures. How people select between these strategies, and what determines whether a risk communication leads to preventive or reactive action remain open questions.

7.8.3 Two Pathways into Adaptive Policies for Risk Communication

We re-emphasise future work considerations that emerged across this thesis and that we began to address specifically under RQ3 (see Section 7.5), concerning the investigation of *adaptive* risk communication. Drawing on our findings, we propose two avenues for future research on adaptive risk communication policies.

First, adaptation may occur *within* the risk communication strategy itself, shaping the communication according to *individual differences* inferred from user characteristics during human-AI interaction. In Chapter 4, we address this form of adaptation extensively in relation to an individual's political orientation. Inference-based adaptation may also concern the *domain* encoding the *risk type* being communicated [176], based on information inferred from the content of the human-AI interaction or its deployment context. In the study design of Chapter 4, we aligned the communicated consequence with the content of the respective conversation between the user and the LLM to ensure that the risk communication was meaningful within that interaction. To illustrate, if a privacy risk is inferred from the content of a conversation with an LLM, the risk communication must be aligned accordingly rather than communicating, for example, a risk of discrimination. Chapter 5 further suggests that *lived experience* and *risk familiarity* may serve as relevant characteristics on which such inference-based adaptation could be grounded. We further envision adaptation to extend to the anthropomorphic features of risk mirroring, for instance through feature similarity between the user and the agent, such as gender or age, to foster self-projection as discussed in Section 7.3, or through the inclusion of self-referent cues in the risk communication design [50], given that perceived similarity has been shown to play an important role in risk communication through virtual avatars [95]. We propose that inference-based adaptation is therefore concerned with *personalising* the content, format, and structure of the risk communication strategy itself. It remains for future HCI research to investigate which inference characteristics are most relevant, how such adaptation can be realised in practice, and how the communicative affordances of a given AI system may constrain or enable the forms of personalised risk communication that are available and meaningful within a given interaction context.

Second, adaptation may occur at the level of the *intervention selection*. We envision adaptive policies akin to those investigated by Buçinca et al. [31], in which optimisation concerns the appropriate selection of an intervention from a pool of calibration approaches, positioning risk communication as a *complementary* strategy alongside other established approaches rather than as a remedy applicable across all tasks, contexts, AI systems, and audiences. As demonstrated in Chapter 3, such adaptation would serve to leverage the complementary strengths of different XAI intervention approaches. Buçinca et al. [31] investigated adaptive policies using reinforcement learning, with intervention selection depending on factors such as an individual's need for cognition, domain knowledge, and the objective of the human-AI interaction. Our findings in Chapter 3 contribute to this work by demonstrating the potential drawbacks and strengths of AI accuracy information and risk communication across different conditions of decision-stakes. Our findings suggest that decision-stakes could be a crucial contextual factor guiding the selection of adaptive policies. Chapter 6 further surfaced considerations around *thresholds* for risk communication, raising questions for future work about *when* risk communication should be displayed. In particular, explicitly communicating uncertainty in instances where

no uncertainty existed eroded system trust (Chapter 6). This finding invites a broader reflection on the potential pitfalls of presenting risk communication when no risk is present. Can risk simply be determined and risk communication switched on or off accordingly? The inherently probabilistic nature of AI systems makes the determination of such communication thresholds difficult. For consequence-explicit risk communication, we posit that withholding risk communication may be more justifiable in instances where no risk consequences are 'detected' according to, for instance, the AI risk domain taxonomy [175, 176]. Notably, these questions carry normative weight and require further investigation, as choosing when to display risk communication is not merely a question of technical accuracy, but may require value judgments about what level of risk is acceptable or whose interests should be prioritised. Importantly, Buçinca et al. [31] further demonstrate that the objective for which such a policy is optimised, whether learning or decision accuracy, is a crucial consideration in adaptive policy design. In Chapter 5, we surface a related distinction between the objectives of risk assessment and risk perception, elaborated in Section 7.3, which suggests that future research on adaptive policies should carefully consider what objective risk communication is optimised for, which may fundamentally shape the selection and the design of the communication strategy itself.

Overall, we propose that adaptive approaches may serve to prevent potential habituation effects [78] of static risk communication, in which risk communication may become background noise over repeated interaction rather than a meaningful signal that shifts perceptions. Relatedly, it also raises research questions foundational to human factors research, which demonstrates cognitive effects such as alert fatigue, suggesting that an increased exposure of risk communication in human-AI interaction, similarly to alerts in human factors research, may create cognitive overload [4]. Further, static risk communication may create a sense of false alarm and therefore may elicit effects akin to the "crying wolf effect" [27, 219], in which a 'real' risk is not perceived as such because the user has built resilience to risk signals due to suboptimal false alarm rates. These are crucial questions for the design and perceptual effects of risk communication in human-AI interaction, and it remains for future work in HCI to shed light on them.

7.8.4 Risk Communication as a Bi-Directional Process

Inference-based policies for risk communication already hinge upon a set of potential future directions that emerge naturally from our investigation and from the nature of inference itself - that of *bi-directionality*. The work presented in this thesis presented risk communication as a primarily directional on-way communication, in line with traditional information theory [168], in that a sender transmits information to a receiver. However, communication is innately a two way street, and risk communication in human-AI interaction could be understood as a *dialogue* between AI system and end-user. This conceptualisation would adhere to the often collaborative nature of human-AI interaction, and align with increasingly agentic systems that are perceived as *agents* actively participating in a common goal, characterised by increasing agency and autonomy [35].

In a way, inference-based risk communication may already be understood as an implicit form of bi-directionality, in that the system draws on cues provided by the end-user to adapt its communication. This highlights an interesting distinction between adaptive risk communication built on implicit versus explicit cues, and raises the question of what a more explicitly bi-directional risk communication process could afford. If future research assumes risk communication as a bi-directional process, than this would open new pathways to questions related to, for instance, the fundamental act of *grounding* [196] to establish shared mental models of the risk itself [22]. Bi-directional risk communication could provide a more advanced understanding of a user's mental models of risks, in that bi-directional risk communication may serve to resolve existing misconceptions about a risk, understand "peripheral beliefs" that end-users may hold, or "indiscriminate beliefs" characterised by imprecision [22]. The act of grounding is further shaped by the communication *medium*, which determines the technique to achieve grounding [41], opening questions on the design of bi-directional risk communication in human-AI interaction. We suggest that future research could investi-

gate how risk grounding could be achieved, whether through natural language capabilities of AI systems [1, 129], or through more traditional HCI signifier-affordance-feedback mechanisms. Beyond mental models, bi-directional risk communication could also serve to establish an understanding of the user's risk-taking propensity, which Stuck et al. [189] proposes to play a crucial part in the context of trust in human-AI interaction, and which our work has not directly addressed. We believe these considerations offer exciting new research directions, driven by the unique characteristics of human-AI interaction, that will inform risk communication theory, HCI research and system designers.

7.8.5 Overlooked Determinants of Risk Perception in Human-AI Interaction

While we carefully selected and grounded the variables examined across this thesis, risk research has long identified a plethora of determinants of risk perception [25] that remain uninvestigated in the context of human-AI interaction. We introduce a selection of these here and encourage future work to examine their role in shaping how users perceive and respond to risk communication in human-AI interaction. Risk research has emphasised worldviews, cultural orientations, and moral foundations as important determinants of risk perception [8, 46, 66, 74]. In Chapter 4, we examined political orientation as one such value-laden lens, and we posit that moral foundations offer a more granular lens, capturing individualising intuitions related to care, fairness, and proportionality that may more precisely predict how individuals appraise AI-related risks. Cross-cultural differences in risk perception are similarly well-established [66] and yet to be investigated in the context of risk communication in human-AI interaction.

In Chapter 4, perceived risk controllability emerged as an important factor in participants' risk appraisal, consistent with its well-established role in risk research [210]. We observed that participants felt they could engage in protective action only when the risk felt somewhat controllable, contrary to systemic AI risks that felt outside of the user's control. Two constructs closely related to controllability that we propose are particularly relevant in the context of human-AI interaction are self-efficacy [11, 91], which describes an individual's belief in the effectiveness of their own action and is considered central to how individuals evaluate risk coping strategies [150], and locus of control [160], which reflects whether an outcome is perceived as contingent on one's own behaviour or on external factors. Including these constructs in future investigations may serve to understand the sweet spot between preserving controllability to a degree that neither undermines the communicated risk nor creates a sense of helplessness or avoidance. Several insights across our empirical work touched on these constructs, and we encourage future work to investigate them more systematically as determinants of AI risk perception. Similarly, responsibility attribution emerged as a recurring theme across our empirical work in Chapter 4 and Chapter 6, which we believe warrants more systematic future investigation. In Chapter 4, a human-interaction framed consequence-explicit AI risk disclaimer shifted responsibility attribution toward the user, positioning them as accountable for scrutinising AI outputs and actively mitigating risk. Notably, in Chapter 6, some participants indicated a willingness to hand over responsibility to the robot, a pattern we posit may reflect the consequence-agnosticism in that study, where risk communication focused solely on the uncertainty component. We assume that consequence explicitness, if designed in a human-centred way, may therefore be a relevant mechanism for user-centric responsibility attribution in human-AI interaction. However, we encourage future research to examine responsibility attribution as a response to risk communication more directly and systematically.

Finally, across this thesis we included AI literacy and dispositional trust in technology as important covariates, consistent with prior work [72, 117, 134, 141]. We encourage future work to investigate the different dimensions of AI literacy more closely, as outlined in work by Carolus et al. [34], alongside familiarity with AI systems, as users increasingly accumulate everyday experience and conceptual understanding with AI systems, and by doing so, potentially develop more stable mental models that may shape their perception of and response to risk communication in human-AI interaction. We further propose that the effectiveness of risk

communication as a cognitive forcing function may depend on individual differences in need for cognition [31] and thus the degree to which users are willing to engage in deliberative processing and appraisal of risk information, which warrants further examination.

Chapter 8

Conclusion

This thesis investigates the design and the perceptual effects of risk communication cues in human-AI interaction across AI systems with varying communicative affordances, and it examines the contextual and individual factors influencing these effects, in order to derive recommendations for human-AI interaction. Throughout, we treat risk as a function of *uncertainty* and *consequence*, and we distinguish between cues which communicate *relational risk* (Chapters 3, 4, 6), with the AI system as the primary causal entity, and *non-relational risk* (Chapter 5), which may not stem directly from the system itself, with a focus on the former.

We first examine framing as an established design approach to risk communication, in the form of a lightweight text-based intervention, by investigating the *frame of valence* (Chapter 3) and the *frame of reference* (Chapter 4). From this work, we derive *human-centredness* as a design desideratum, i.e. risk cues need to communicate risk in a way that is personally and experientially consequential for users. In order to pursue human-centredness further, we then move toward *anthropomorphism* as a design mechanism to encode risk information across embodied affordances, operationalised as consequence encoded through risk mirroring (Chapter 5) and uncertainty encoded through a hesitation gesture (Chapter 6). Across the text-based interventions, we show that risk communication can support more appropriate perceptions of system trustworthiness, that its mere presence can heighten risk perceptions, and that it holds potential as a cognitive forcing function. The framing of risk matters, in that a human-centred frame elicits severity perceptions and personal relevance, whereas systemic AI risk communication carries pitfalls. We further show that there is no one-size-fits-all remedy, as the appropriateness of a cue is highly determined by the stakes associated with the risk and by the political lens and lived experiences through which the risk information is appraised. Across the work centred on robots as embodied agents, an anthropomorphic encoding of risk information proves promising as a means to make risk communication more human-centred, but also reveals a trade-off between the holistic perception of risk and the risk assessment qualities of a cue, and we suggest that affect-inducing cues must be calibrated to motivate concern without undermining the intended communicative purpose.

From these results, we distil a set of propositions for user-facing risk communication in human-AI interaction across communicative affordances, and we reflect on its normative questions and systemic implications. Looking forward, we argue that effort is needed to conduct longitudinal studies embedded in real-world settings and to investigate the behavioural outcomes of risk communication through the lens of risk theory, beyond currently established reliance measures, where we propose preventive and reactive measures. We further propose to investigate two pathways for adaptive risk communication, at the level of the cue itself, i.e. its format and content, and at the level of intervention selection, and to consider risk communication as a bi-directional process. This thesis positions risk communication as an interdisciplinary challenge, in which the inclusion of risk theory is imperative for user-facing risk communication in human-AI interaction, and its findings inform both researchers and policymakers.

Bibliography

[1] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, Aug. 2022. URL http://arxiv.org/abs/2204.01691. arXiv:2204.01691 [cs.RO].

[2] S. J. G. Ahn, J. Fox, and J. M. Hahm. Using Virtual Doppelgängers to Increase Personal Relevance of Health Risk Communication. In T. Bickmore, S. Marsella, and C. Sidner, editors, *Intelligent Virtual Agents*, pages 1–12, Cham, 2014. Springer International Publishing. ISBN 978-3-319-09767-1. doi: 10.1007/978-3-319-09767-1_1.

[3] J. S. Ancker, Y. Senathirajah, R. Kukafka, and J. B. Starren. Design Features of Graphs in Health Risk Communication: A Systematic Review. *Journal of the American Medical Informatics Association*, 13(6): 608–618, Nov. 2006. ISSN 1067-5027. doi: 10.1197/jamia.M2115. URL https://doi.org/10.1197/jamia.M2115.

[4] J. S. Ancker, A. Edwards, S. Nosal, D. Hauser, E. Mauer, and R. Kaushal. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making*, 17(1):36, Dec. 2017. ISSN 1472-6947. doi: 10.1186/s12911-017-0430-8. URL https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0430-8.

[5] Anthropic. Claude. https://claude.ai, 2026.

[6] K. Armstrong, J. S. Schwartz, G. Fitzgerald, M. Putt, and P. A. Ubel. Effect of Framing as Gain versus Loss on Understanding and Hypothetical Treatment Choices: Survival and Mortality Curves. *Medical Decision Making*, 22(1):76–83, 2002. doi: 10.1177/0272989X0202200108. URL https://doi.org/10.1177/0272989X0202200108. _eprint: https://doi.org/10.1177/0272989X0202200108.

[7] M. Ashoori and J. D. Weisz. In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes, Dec. 2019. URL http://arxiv.org/abs/1912.02675. arXiv:1912.02675 [cs].

[8] M. Atari, J. Haidt, J. Graham, S. Koleva, S. T. Stevens, and M. Dehghani. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157–1188, Nov. 2023. ISSN 1939-1315, 0022-3514. doi: 10.1037/pspp0000470. URL https://doi.apa.org/doi/10.1037/pspp0000470.

[9] C. J. Atman, A. Bostrom, B. Fischhoff, and M. G. Morgan. Designing Risk Communications: Completing and Correcting Mental Models of Hazardous Processes, Part I. *Risk Analysis*, 14(5):779–788, Oct. 1994. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.1539-6924.1994.tb00289.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.1994.tb00289.x.

[10] T. Aven, Y. Ben-Haim, H. B. Andersen, T. Cox, E. López Droguett, M. Greenberg, S. Guikema, W. Kröger, O. Renn, K. M. Thompson, and E. Zio. Society for Risk Analysis Glossary. Technical report, Society for Risk Analysis, Aug. 2018. URL https://www.sra.org/wp-content/uploads/2020/04/SRA-Glossary-FINAL.pdf.

[11] A. Bandura and N. E. Adams. Analysis of self-efficacy theory of behavioral change. *Cognitive Therapy and Research*, 1(4):287–310, Dec. 1977. ISSN 0147-5916, 1573-2819. doi: 10.1007/BF01663995. URL http://link.springer.com/10.1007/BF01663995.

[12] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:2–11, Oct. 2019. ISSN 2769-1349, 2769-1330. doi: 10.1609/hcomp.v7i1.5285. URL https://ojs.aaai.org/index.php/HCOMP/article/view/5285.

[13] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445717. URL https://dl.acm.org/doi/10.1145/3411764.3445717.

[14] D. A. A. Baranger, M. C. Finsaas, B. L. Goldstein, C. E. Vize, D. R. Lynam, and T. M. Olino. Tutorial: Power analyses for interaction effects in cross-sectional regressions. *Advances in Methods and Practices in Psychological Science*, 6(3), 2023. doi: 10.1177/25152459231187531.

[15] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, L. Nachman, R. Chunara, M. Srikumar, A. Weller, and A. Xiang. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, Virtual Event USA, July 2021. ACM. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462571. URL https://dl.acm.org/doi/10.1145/3461702.3462571.

[16] J. Y. Bo, S. Wan, and A. Anderson. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23, Yokohama Japan, Apr. 2025. ACM. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3714097. URL https://dl.acm.org/doi/10.1145/3706598.3714097.

[17] E. P. Bogucka, M. Constantinides, S. Šćepanović, and D. Quercia. Impact Assessment Card: Communicating Risks and Benefits of AI Uses. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–42, Oct. 2025. ISSN 2573-0142. doi: 10.1145/3757482. URL https://dl.acm.org/doi/10.1145/3757482.

[18] N. Boonprakong, X. Chen, C. Davey, B. Tag, and T. Dingler. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases when Facing Different Opinions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, Hamburg Germany, Apr. 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3580917. URL https://dl.acm.org/doi/10.1145/3544548.3580917.

[19] N. Boonprakong, B. Tag, J. Goncalves, and T. Dingler. How Do HCI Researchers Study Cognitive Biases? A Scoping Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20, Yokohama Japan, Apr. 2025. ACM. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3713450. URL https://dl.acm.org/doi/10.1145/3706598.3713450.

[20] A. Bostrom, B. Fischhoff, and M. G. Morgan. Characterizing Mental Models of Hazardous Processes: A Methodology and an Application to Radon. *Journal of Social Issues*, 48(4):85–100, Jan. 1992. ISSN 0022-4537, 1540-4560. doi: 10.1111/j.1540-4560.1992.tb01946.x. URL https://spssi.onlinelibrary.wiley.com/doi/10.1111/j.1540-4560.1992.tb01946.x.

[21] A. Bostrom, C. J. Atman, B. Fischhoff, and M. G. Morgan. Evaluating Risk Communications: Completing and Correcting Mental Models of Hazardous Processes, Part II. *Risk Analysis*, 14(5):789–798, Oct. 1994. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.1539-6924.1994.tb00290.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.1994.tb00290.x.

[22] A. Bostrom, M. G. Morgan, B. Fischhoff, and D. Read. What Do People Know About Global Climate Change? 1. Mental Models. *Risk Analysis*, 14(6):959–970, Dec. 1994. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.1539-6924.1994.tb00065.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.1994.tb00065.x.

[23] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, Jan. 2006. ISSN 1478-0887. doi: 10.1191/1478088706qp063oa. URL https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

[24] G. M. Breakwell. Risk communication: fators affecting impact. *British Medical Bulletin*, 56(1):110–120, Jan. 2000. ISSN 0007-1420. doi: 10.1258/0007142001902824. URL https://doi.org/10.1258/0007142001902824.

[25] G. M. Breakwell. *The Psychology of Risk*. Cambridge University Press, 2007.

[26] S. B. Brennen, S. Lake, A. J. Lazard, and A. Reid. In Disclaimers We Trust: The Effectiveness of State-Required AI Disclaimers on Political Ads. Policy Brief, Center on Technology Policy, New York University, New York, NY, 2024.

[27] S. Breznitz. False Alarms: Their Effects on Fear and Adjustment. *Issues in Mental Health Nursing*, 7 (1-4):335–348, Jan. 1985. ISSN 0161-2840, 1096-4673. doi: 10.3109/01612848509009460. URL http://www.tandfonline.com/doi/full/10.3109/01612848509009460.

[28] S. L. Brown and E. Z. Smith. The inhibitory effect of a distressing anti-smoking message on risk perceptions in smokers. *Psychology & Health*, 22(3):255–268, Apr. 2007. ISSN 0887-0446. doi: 10.1080/14768320600843127. URL https://doi.org/10.1080/14768320600843127. _eprint: https://doi.org/10.1080/14768320600843127.

[29] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, Apr. 2021. ISSN 2573-0142. doi: 10.1145/3449287. URL https://dl.acm.org/doi/10.1145/3449287.

[30] Z. Buçinca, C. M. Pham, M. Jakesch, M. T. Ribeiro, A. Olteanu, and S. Amershi. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms, June 2023. URL http://arxiv.org/abs/2306.03280. arXiv:2306.03280 [cs].

[31] Z. Buçinca, S. Swaroop, A. E. Paluch, S. A. Murphy, and K. Z. Gajos. Towards Optimizing Human-Centric Objectives in AI-Assisted Decision-Making With Offline Reinforcement Learning, Apr. 2024. URL http://arxiv.org/abs/2403.05911. arXiv:2403.05911 [cs].

[32] S. Cao, A. Liu, and C.-M. Huang. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–32, Apr. 2024. ISSN 2573-0142. doi: 10.1145/3637318. URL https://dl.acm.org/doi/10.1145/3637318.

[33] R. N. Carey and K. M. Sarma. Threat appeals in health communication: messages that elicit fear and enhance perceived efficacy positively impact on young male drivers. *BMC Public Health*, 16(1): 645, July 2016. ISSN 1471-2458. doi: 10.1186/s12889-016-3227-2. URL https://doi.org/10.1186/s12889-016-3227-2.

[34] A. Carolus, M. J. Koch, S. Straka, M. E. Latoschik, and C. Wienrich. MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans*, 1(2):100014, Aug. 2023. ISSN 29498821. doi: 10.1016/j.chbah.2023.100014. URL https://linkinghub.elsevier.com/retrieve/pii/S2949882123000142.

[35] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, and T. Maharaj. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, Chicago IL USA, June 2023. ACM. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3594033. URL https://dl.acm.org/doi/10.1145/3593013.3594033.

[36] A. Chan, C. Ezell, M. Kaufmann, K. Wei, L. Hammond, H. Bradley, E. Bluemke, N. Rajkumar, D. Krueger, N. Kolt, L. Heim, and M. Anderljung. Visibility into AI Agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 958–973, Rio de Janeiro Brazil, June 2024. ACM. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658948. URL https://dl.acm.org/doi/10.1145/3630106.3658948.

[37] M. Chiriatti, M. Ganapini, E. Panai, M. Ubiali, and G. Riva. The case for human–AI interaction as system 0 thinking. *Nature Human Behaviour*, 8(10):1829–1830, Oct. 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-01995-5. URL https://www.nature.com/articles/s41562-024-01995-5.

[38] M. Chiriatti, M. Bergamaschi Ganapini, E. Panai, B. K. Wiederhold, and G. Riva. System 0: Transforming Artificial Intelligence into a Cognitive Extension. *Cyberpsychology, Behavior, and Social Networking*, 28(7):534–542, July 2025. ISSN 2152-2715, 2152-2723. doi: 10.1089/cyber.2025.0201. URL https://journals.sagepub.com/doi/10.1089/cyber.2025.0201.

[39] B. L. Choma, Y. Hanoch, M. Gummerum, and G. Hodson. Relations between risk perceptions and socio-political ideology are domain- and ideology- dependent. *Personality and Individual Differences*, 54(1):29–34, Jan. 2013. ISSN 0191-8869. doi: 10.1016/j.paid.2012.07.028. URL https://linkinghub.elsevier.com/retrieve/pii/S0191886912003662.

[40] H. Chu and J. Z. Yang. Risk or Efficacy? How Psychological Distance Influences Climate Change Engagement. *Risk Analysis*, 40(4):758–770, Apr. 2020. ISSN 0272-4332, 1539-6924. doi: 10.1111/risa.13446. URL https://onlinelibrary.wiley.com/doi/10.1111/risa.13446.

[41] H. H. Clark and S. E. Brennan. Grounding in communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, editors, *Perspectives on socially shared cognition.*, pages 127–149. American Psychological Association, Washington, 1991. ISBN 978-1-55798-121-9. doi: 10.1037/10096-006. URL https://content.apa.org/books/10096-006.

[42] M. C. Claudy, M. Parkinson, and K. Aquino. Why should innovators care about morality? Political ideology, moral foundations, and the acceptance of technological innovations. *Technological Forecasting and Social Change*, 203:123384, June 2024. ISSN 00401625. doi: 10.1016/j.techfore.2024.123384. URL https://linkinghub.elsevier.com/retrieve/pii/S004016252400180X.

[43] M. Coeckelbergh. Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 3(4):1341–1350, Nov. 2023. ISSN 2730-5953, 2730-5961. doi: 10.1007/s43681-022-00239-4. URL https://link.springer.com/10.1007/s43681-022-00239-4.

[44] M. Coeckelbergh and H. S. Sætra. Climate change and the political pathways of AI: The technocracy-democracy dilemma in light of artificial intelligence and human agency. *Technology in Society*, 75: 102406, Nov. 2023. ISSN 0160791X. doi: 10.1016/j.techsoc.2023.102406. URL https://linkinghub.elsevier.com/retrieve/pii/S0160791X23002117.

[45] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva. Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 507–513, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[46] K. Dake. Orienting Dispositions in the Perception of Risk: An Analysis of Contemporary Worldviews and Cultural Biases. *Journal of Cross-Cultural Psychology*, 22(1):61–82, 1991. doi: 10.1177/0022022191221006. URL https://doi.org/10.1177/0022022191221006. _eprint: https://doi.org/10.1177/0022022191221006.

[47] J. De Miguel Velázquez, S. Šćepanović, A. Gvirtz, and D. Quercia. Decoding Real-World Artificial Intelligence Incidents. *Computer*, 57(11):71–81, Nov. 2024. ISSN 1558-0814. doi: 10.1109/MC.2024.3432492. URL https://ieeexplore.ieee.org/document/10718657/.

[48] L. Derczynski, H. R. Kirk, V. Balachandran, S. Kumar, Y. Tsvetkov, M. R. Leiser, and S. Mohammad. Assessing Language Model Deployment with Risk Cards, Mar. 2023. URL http://arxiv.org/abs/2303.18190. arXiv:2303.18190 [cs].

[49] A. D. Dragan, K. C. Lee, and S. S. Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308, Tokyo, Japan, Mar. 2013. IEEE. ISBN 978-1-4673-3101-2 978-1-4673-3099-2 978-1-4673-3100-5. doi: 10.1109/HRI.2013.6483603. URL http://ieeexplore.ieee.org/document/6483603/.

[50] S. Dunlop, M. Wakefield, and Y. Kashima. Can You Feel It? Negative Emotion, Risk, and Narrative in Health Communication. *Media Psychology*, 11(1):52–75, Mar. 2008. ISSN 1521-3269, 1532-785X. doi: 10.1080/15213260701853112. URL http://www.tandfonline.com/doi/abs/10.1080/15213260701853112.

[51] H. Elder, C. Canfield, D. B. Shank, T. Rieger, and C. Hines. Knowing When to Pass: The Effect of AI Reliability in Risky Decision Contexts. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 66(2):348–362, Feb. 2024. ISSN 0018-7208, 1547-8181. doi: 10.1177/00187208221100691. URL https://journals.sagepub.com/doi/10.1177/00187208221100691.

[52] European Parliament and Council. Regulation (eu) 2024/1689 of the european parliament and of the council on artificial intelligence, article 50, 2024. URL https://artificialintelligenceact.eu/article/50/.

[53] European Parliament and Council. Regulation (eu) 2024/1689 of the european parliament and of the council on artificial intelligence, chapter iii: High-risk ai system, 2024. URL https://artificialintelligenceact.eu/chapter/3/.

[54] J. S. B. T. Evans and K. E. Stanovich. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8(3):223–241, May 2013. ISSN 1745-6916, 1745-6924. doi: 10.1177/1745691612460685. URL https://journals.sagepub.com/doi/10.1177/1745691612460685.

[55] H. Fahnenstich, T. Rieger, and E. Roesler. Trusting under risk – comparing human to AI decision support agents. *Computers in Human Behavior*, 153:108107, Apr. 2024. ISSN 07475632. doi: 10.1016/j.chb.2023.108107. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563223004582.

[56] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, 2007. ISSN 1554-351X. doi: 10.3758/BF03193146.

[57] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, May 2007. ISSN 1554-3528. doi: 10.3758/BF03193146. URL https://doi.org/10.3758/BF03193146.

[58] M. Feinberg and R. Willer. Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, 13(12):e12501, Dec. 2019. ISSN 1751-9004, 1751-9004. doi: 10.1111/spc3.12501. URL https://compass.onlinelibrary.wiley.com/doi/10.1111/spc3.12501.

[59] E. Ferrara. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 7(1):549–569, Apr. 2024. ISSN 2432-2717, 2432-2725. doi: 10.1007/s42001-024-00250-1. URL https://link.springer.com/10.1007/s42001-024-00250-1.

[60] M. L. Finucane, A. Alhakami, P. Slovic, and S. M. Johnson. The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, 13(1):1–17, Jan. 2000. ISSN 0894-3257, 1099-0771. doi: 10.1002/(SICI)1099-0771(200001/03)13:1⟨1::AID-BDM333⟩3.0.CO;2-S. URL https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S.

[61] B. Fischhoff. Risk Perception and Communication Unplugged: Twenty Years of Process[1]. *Risk Analysis*, 15(2):137–145, Apr. 1995. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.1539-6924.1995.tb00308.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.1995.tb00308.x.

[62] B. Fischhoff, P. Slovic, S. Lichtenstein, S. Read, and B. Combs. How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences*, 9(2):127–152, Apr. 1978. ISSN 0032-2687, 1573-0891. doi: 10.1007/BF00143739. URL http://link.springer.com/10.1007/BF00143739.

[63] D. L. Floyd, S. Prentice-Dunn, and R. W. Rogers. A Meta-Analysis of Research on Protection Motivation Theory. *Journal of Applied Social Psychology*, 30(2):407–429, Feb. 2000. ISSN 0021-9029, 1559-1816. doi: 10.1111/j.1559-1816.2000.tb02323.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.2000.tb02323.x.

[64] M. Gerlich. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1):6, Jan. 2025. ISSN 2075-4698. doi: 10.3390/soc15010006. URL https://www.mdpi.com/2075-4698/15/1/6. Number: 1.

[65] J. J. Gibson. The Theory of Affordances. In *The Ecological Approach to Visual Perception*, pages 119–135. Psychology Press, 1 edition, 2015. doi: 10.4324/9781315740218-10.

[66] E. Gierlach, B. E. Belsher, and L. E. Beutler. Cross-Cultural Differences in Risk Perceptions of Disasters. *Risk Analysis*, 30(10):1539–1549, 2010. ISSN 1539-6924. doi: 10.1111/j.1539-6924.2010.01451.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.2010.01451.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.2010.01451.x.

[67] G. Gigerenzer. How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases". *European Review of Social Psychology*, 2(1):83–115, Jan. 1991. ISSN 1046-3283, 1479-277X. doi: 10.1080/14792779143000033. URL http://www.tandfonline.com/doi/abs/10.1080/14792779143000033.

[68] G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin. Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest*, 8(2):53–96, 2007. doi: 10.1111/j.1539-6053.2008.00033.x. URL https://doi.org/10.1111/j.1539-6053.2008.00033.x. _eprint: https://doi.org/10.1111/j.1539-6053.2008.00033.x.

[69] M. Glickman and T. Sharot. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2):345–359, Dec. 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-02077-2. URL https://www.nature.com/articles/s41562-024-02077-2.

[70] J. Gong, Y. Zhang, Z. Yang, Y. Huang, J. Feng, and W. Zhang. The framing effect in medical decision-making: a review of the literature. *Psychology, Health & Medicine*, 18(6):645–653, Dec. 2013. ISSN 1354-8506, 1465-3966. doi: 10.1080/13548506.2013.766352. URL https://www.tandfonline.com/doi/full/10.1080/13548506.2013.766352.

[71] K. Gorkovenko, D. J. Burnett, J. K. Thorp, D. Richards, and D. Murray-Rust. Exploring The Future of Data-Driven Product Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Honolulu HI USA, Apr. 2020. ACM. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376560. URL https://dl.acm.org/doi/10.1145/3313831.3376560.

[72] J. Govers, E. Velloso, V. Kostakos, and J. Goncalves. Ai-driven mediation strategies for audience depolarisation in online debates. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 979-8-4007-0330-0/24/05. doi: 10.1145/3613904.3642322. URL https://doi.org/10.1145/3613904.3642322.

[73] J. Graham, J. Haidt, and B. A. Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046, May 2009. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0015141. URL https://doi.apa.org/doi/10.1037/a0015141.

[74] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto. Moral Foundations Theory. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier, 2013. ISBN 978-0-12-407236-7. doi: 10.1016/B978-0-12-407236-7.00002-4. URL https://linkinghub.elsevier.com/retrieve/pii/B9780124072367000024.

[75] J. Haidt and C. Joseph. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66, Sept. 2004. ISSN 0011-5266, 1548-6192. doi: 10.1162/0011526042365555. URL https://direct.mit.edu/daed/article/133/4/55-66/27470.

[76] G. He, S. Buijsman, and U. Gadiraju. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–29, Sept. 2023. ISSN 2573-0142. doi: 10.1145/3610067. URL https://dl.acm.org/doi/10.1145/3610067.

[77] L. Hendrickx, C. Vlek, and H. Oppewal. Relative importance of scenario information and frequency information in the judgment of risk. *Acta Psychologica*, 72(1):41–63, Sept. 1989. ISSN 00016918. doi: 10.1016/0001-6918(89)90050-4. URL https://linkinghub.elsevier.com/retrieve/pii/0001691889900504.

[78] S. Hoesterey and L. Onnasch. The effect of risk on trust attitude and trust behavior in interaction with information and decision automation. *Cognition, Technology & Work*, 25(1):15–29, Feb. 2023. ISSN 1435-5558, 1435-5566. doi: 10.1007/s10111-022-00718-y. URL https://link.springer.com/10.1007/s10111-022-00718-y.

[79] K. A. Hoff and M. Bashir. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(3):407–434, May 2015. ISSN 0018-7208, 1547-8181. doi: 10.1177/0018720814547570. URL https://journals.sagepub.com/doi/10.1177/0018720814547570.

[80] M. Hoffmann and H. Frase. Adding Structure to AI Harm. Technical report, Center for Security and Emerging Technology, July 2023. URL https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/.

[81] K. Hornbæk, P. O. Kristensson, and A. Oulasvirta. *Introduction to Human-Computer Interaction*. Oxford University Press, United Kingdom, 2025. ISBN 978-0-19-286454-3. doi: 10.1093/oso/9780192864543.001.0001.

[82] J. Hough and D. Schlangen. It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 274–282, Vienna Austria, Mar. 2017. ACM. ISBN 978-1-4503-4336-7. doi: 10.1145/2909824.3020214. URL https://dl.acm.org/doi/10.1145/2909824.3020214.

[83] S. Hösterey and L. Onnasch. Manipulating situational risk in human-automation research – Validation of a new experimental paradigm in virtual reality. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 65(1):1109–1113, Sept. 2021. ISSN 1071-1813, 2169-5067. doi: 10.1177/1071181321651161. URL https://journals.sagepub.com/doi/10.1177/1071181321651161.

[84] International Organization for Standardization. ISO 31000:2018 Risk Management – Guidelines, Term 3.6, 2018. URL https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en:term:3.6.

[85] J.-Y. Jian, A. M. Bisantz, and C. G. Drury. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, Mar. 2000. ISSN 1088-6362. doi: 10.1207/S15327566IJCE0401_04. URL http://www.tandfonline.com/doi/abs/10.1207/S15327566IJCE0401_04.

[86] J. T. Jost, B. A. Nosek, and S. D. Gosling. Ideology: Its Resurgence in Social, Personality, and Political Psychology. *Perspectives on Psychological Science*, 3(2):126–136, Mar. 2008. ISSN 1745-6916. doi: 10.1111/j.1745-6916.2008.00070.x. URL https://doi.org/10.1111/j.1745-6916.2008.00070.x.

[87] D. Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, NY, 2011.

[88] D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, 1982. ISBN 0-521-24064-6.

[89] P. Kahr and A. Bernstein. Enough With Trust! Why We Must Move Beyond a Convenient but Insufficient Concept in AI-Supported Decision-Making Research. In *Proceedings of the 1st International Conference on Human-Computer Interaction in the Alps*, pages 109–113, Ascona Switzerland, Mar. 2026. ACM. ISBN 979-8-4007-1948-6. doi: 10.1145/3780045.3780056. URL https://dl.acm.org/doi/10.1145/3780045.3780056.

[90] P. K. Kahr, G. Rooks, C. Snijders, and M. C. Willemsen. The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 609–622, Greenville SC USA, Mar. 2024. ACM. ISBN 979-8-4007-0508-3. doi: 10.1145/3640543.3645167. URL https://dl.acm.org/doi/10.1145/3640543.3645167.

[91] H. Kallmen. Manifest anxiety, general self-efficacy and locus of control as determinants of personal and general risk perception. *Journal of Risk Research*, 3(2):111–120, Apr. 2000. ISSN 1366-9877, 1466-4461. doi: 10.1080/136698700376626. URL http://www.tandfonline.com/doi/abs/10.1080/136698700376626.

[92] C. Keller, M. Siegrist, and H. Gutscher. The Role of the Affect and Availability Heuristics in Risk Communication. *Risk Analysis*, 26(3):631–639, June 2006. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.1539-6924.2006.00773.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.2006.00773.x.

[93] H. Kennedy and R. L. Hill. The Feeling of Numbers: Emotions in Everyday Engagements with Data and Their Visualisation. *Sociology*, 52(4):830–848, Aug. 2018. ISSN 0038-0385, 1469-8684. doi: 10.1177/0038038516674675. URL https://journals.sagepub.com/doi/10.1177/0038038516674675.

[94] E. Kerzner, S. Goodwin, J. Dykes, S. Jones, and M. Meyer. A Framework for Creative Visualization-Opportunities Workshops. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):748–758, Jan. 2019. ISSN 1941-0506. doi: 10.1109/TVCG.2018.2865241. URL https://ieeexplore.ieee.org/abstract/document/8440830. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[95] H.-K. Kim and S.-H. Kim. Understanding emotional bond between the creator and the avatar: Change in behavioral intentions to engage in alcohol-related traffic risk behaviors. *Computers in Human Behavior*, 62:186–200, Sept. 2016. ISSN 07475632. doi: 10.1016/j.chb.2016.03.092. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563216302631.

[96] S. S. Y. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, and J. W. Vaughan. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 822–835, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658941. URL https://dl.acm.org/doi/10.1145/3630106.3658941.

[97] T. Kim and H. Song. "I Believe AI Can Learn from the Error. Or Can It Not?": The Effects of Implicit Theories on Trust Repair of the Intelligent Agent. *International Journal of Social Robotics*, 15(1):115–128, Jan. 2023. ISSN 1875-4791, 1875-4805. doi: 10.1007/s12369-022-00951-5. URL https://link.springer.com/10.1007/s12369-022-00951-5.

[98] J. Kjeldskov and J. Paay. A longitudinal review of Mobile HCI research methods. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pages 69–78, San Francisco California USA, Sept. 2012. ACM. ISBN 978-1-4503-1105-2. doi: 10.1145/2371574.2371586. URL https://dl.acm.org/doi/10.1145/2371574.2371586.

[99] F. H. Knight. *Risk, uncertainty and profit*, volume 31. Houghton Mifflin, 1921.

[100] S. C. Kohn, E. J. De Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology*, 12:604977, Oct. 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.604977. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2021.604977/full.

[101] M. Kwon, S. H. Huang, and A. D. Dragan. Expressing Robot Incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 87–95, Chicago IL USA, Feb. 2018. ACM. ISBN 978-1-4503-4953-6. doi: 10.1145/3171221.3171276. URL https://dl.acm.org/doi/10.1145/3171221.3171276.

[102] M. Körber. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, and Y. Fujita, editors, *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, pages 13–30, Cham, 2019. Springer International Publishing. ISBN 978-3-319-96074-6. doi: 10.1007/978-3-319-96074-6_2.

[103] V. Lai and C. Tan. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38, Atlanta GA USA, Jan. 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287590. URL https://dl.acm.org/doi/10.1145/3287560.3287590.

[104] X. Lan, Y. Shi, Y. Wu, X. Jiao, and N. Cao. Kineticharts: Augmenting Affective Expressiveness of Charts in Data Stories with Animation Design. *IEEE Transactions on Visualization and Computer Graphics*, 28 (1):933–943, Jan. 2022. ISSN 1077-2626, 1941-0506, 2160-9306. doi: 10.1109/TVCG.2021.3114775. URL https://ieeexplore.ieee.org/document/9552240/.

[105] X. Lan, Y. Wu, Y. Shi, Q. Chen, and N. Cao. Negative Emotions, Positive Outcomes? Exploring the Communication of Negativity in Serious Data Stories. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, New Orleans LA USA, Apr. 2022. ACM. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3517530. URL https://dl.acm.org/doi/10.1145/3491102.3517530.

[106] X. Lan, Y. Wu, and N. Cao. Affective Visualization Design: Leveraging the Emotional Impact of Data. In *IEEE Transactions on Visualization and Computer Graphics*, pages 1–11, 2023. doi: 10.1109/TVCG.2023.3327385. URL https://ieeexplore.ieee.org/document/10301796/.

[107] N. Lankton, D. McKnight, and J. Tripp. Technology, Humanness, and Trust: Rethinking Trust in Technology. *Journal of the Association for Information Systems*, 16(10), Oct. 2015. ISSN 1536-9323. doi: 10.17705/1jais.00411. URL https://aisel.aisnet.org/jais/vol16/iss10/1.

[108] N. K. Lankton, D. H. McKnight, and J. Tripp. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10):1, 2015.

[109] H. D. Lasswell. The Structure and Function of Communication in Society. In L. Bryson, editor, *The Communication of Ideas*, pages 84–99. The Institute for Religious and Social Studies, New York, 1948.

[110] J. Lee and N. Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, Oct. 1992. ISSN 0014-0139, 1366-5847. doi: 10.1080/00140139208967392. URL https://www.tandfonline.com/doi/full/10.1080/00140139208967392.

[111] J. D. Lee and K. A. See. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1):50–80, 2004. ISSN 0018-7208. doi: 10.1518/hfes.46.1.50_30392. URL https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392. _eprint: https://journals.sagepub.com/doi/pdf/10.1518/hfes.46.1.50_30392.

[112] A. A. Leiserowitz. American Risk Perceptions: Is Climate Change Dangerous? *Risk Analysis*, 25(6): 1433–1442, Dec. 2005. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.1540-6261.2005.00690.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2005.00690.x.

[113] G. LeMasurier, A. Gautam, Z. Han, J. W. Crandall, and H. A. Yanco. Reactive or Proactive? How Robots Should Explain Failures. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 413–422, Boulder CO USA, Mar. 2024. ACM. ISBN 979-8-4007-0322-5. doi: 10.1145/3610977.3634963. URL https://dl.acm.org/doi/10.1145/3610977.3634963.

[114] S. F. LeRoy and L. D. Singell. Knight on Risk and Uncertainty. *Journal of Political Economy*, 95(2): 394–406, Apr. 1987. ISSN 00223808. URL https://research.ebsco.com/linkprocessor/plink?id=d5bec9e8-38a5-3c8a-a1f5-887095721911.

[115] S. Lewandowsky and S. Van Der Linden. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 32(2):348–384, July 2021. ISSN 1046-3283, 1479-277X. doi: 10.1080/10463283.2021.1876983. URL https://www.tandfonline.com/doi/full/10.1080/10463283.2021.1876983.

[116] G. F. Loewenstein, E. U. Weber, C. K. Hsee, and N. E. Welch. Risk as feelings. *Psychological Bulletin*, 127 (2):267–286, 2001. doi: 10.1037/0033-2909.127.2.267.

[117] D. Long and B. Magerko. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16, Honolulu HI USA, Apr. 2020. ACM. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376727. URL https://dl.acm.org/doi/10.1145/3313831.3376727.

[118] D. M. Luke and B. Gawronski. Political Ideology and Moral Dilemma Judgments: An Analysis Using the CNI Model. *Personality and Social Psychology Bulletin*, 47(10):1520–1531, Oct. 2021. ISSN 0146-1672, 1552-7433. doi: 10.1177/0146167220987990. URL https://journals.sagepub.com/doi/10.1177/0146167220987990.

[119] S. Luís, C.-M. Vauclair, and M. L. Lima. Raising awareness of climate change causes? Cross-national evidence for the normalization of societal risk perception of climate change. *Environmental Science & Policy*, 80:74–81, Feb. 2018. ISSN 14629011. doi: 10.1016/j.envsci.2017.11.015. URL https://linkinghub.elsevier.com/retrieve/pii/S146290111730881X.

[120] Y. Ma, S. Zhang, D. Yang, Z. Sarsenbayeva, J. Knibbe, and J. Goncalves. Raising Awareness of Location Information Vulnerabilities in Social Media Photos using LLMs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Yokohama Japan, Apr. 2025. ACM. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3714074. URL https://dl.acm.org/doi/10.1145/3706598.3714074.

[121] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall. Robot operating system 2: Design, architecture, and uses in the wild. *Science Robotics*, 7(66):eabm6074, 2022. doi: 10.1126/scirobotics.abm6074. URL https://www.science.org/doi/abs/10.1126/scirobotics.abm6074.

[122] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An Integrative Model Of Organizational Trust. *Academy of Management Review*, July 1995. doi: 10.5465/amr.1995.9508080335. URL https://journals.aom.org/doi/abs/10.5465/amr.1995.9508080335.

[123] K. J. McCaffery, A. Dixon, A. Hayen, J. Jansen, S. Smith, and J. M. Simpson. The Influence of Graphic Display Format on the Interpretations of Quantitative Risk Information among Adults with Lower Education and Literacy: A Randomized Experimental Study. *Medical Decision Making*, 32(4):532–544,

July 2012. ISSN 0272-989X. doi: 10.1177/0272989X11424926. URL https://doi.org/10.1177/0272989X11424926.

[124] R. I. McDonald, H. Y. Chai, and B. R. Newell. Personal experience and the 'psychological distance' of climate change: An integrative review. *Journal of Environmental Psychology*, 44:109–118, Dec. 2015. ISSN 02724944. doi: 10.1016/j.jenvp.2015.10.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0272494415300396.

[125] S. McGregor. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):15458–15463, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i17.17817. URL https://ojs.aaai.org/index.php/AAAI/article/view/17817. Number: 17.

[126] S. McGregor, K. Paeth, and K. Lam. Indexing AI Risks with Incidents, Issues, and Variants, Nov. 2022. URL http://arxiv.org/abs/2211.10384. arXiv:2211.10384 [cs.CY].

[127] B. J. McNeil, S. G. Pauker, H. C. Sox, and A. Tversky. On the Elicitation of Preferences for Alternative Therapies. *New England Journal of Medicine*, 306(21):1259–1262, 1982. doi: 10.1056/NEJM198205273062103. URL https://www.nejm.org/doi/full/10.1056/NEJM198205273062103. _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJM198205273062103.

[128] T. Miller. Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI. In *2023 ACM Conference on Fairness Accountability and Transparency*, pages 333–342, Chicago IL USA, June 2023. ACM. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3594001. URL https://dl.acm.org/doi/10.1145/3593013.3594001.

[129] D. K. Misra, J. Sung, K. Lee, and A. Saxena. Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300, Jan. 2016. ISSN 0278-3649, 1741-3176. doi: 10.1177/0278364915602060. URL https://journals.sagepub.com/doi/10.1177/0278364915602060.

[130] A. Moon, C. A. C. Parker, E. A. Croft, and H. F. M. Van der Loos. Did you see it hesitate? - Empirically grounded design of hesitation trajectories for collaborative robots. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1994–1999, Sept. 2011. doi: 10.1109/IROS.2011.6094605. URL https://ieeexplore.ieee.org/abstract/document/6094605.

[131] R. E. Morss, C. L. Cuite, J. L. Demuth, W. K. Hallman, and R. L. Shwom. Is storm surge scary? The influence of hazard, impact, and fear-based messages and individual differences on responses to hurricane risks in the USA. *International Journal of Disaster Risk Reduction*, 30:44–58, Sept. 2018. ISSN 2212-4209. doi: 10.1016/j.ijdrr.2018.01.023. URL https://www.sciencedirect.com/science/article/pii/S2212420918300876.

[132] S. Mylius, P. Slattery, A. K. Saeri, and N. Thompson. MIT AI Incident Tracker, 2025. URL https://airisk.mit.edu/ai-incident-tracker. Published: MIT AI Risk Initiative. Available at airisk.mit.edu/ai-incident-tracker. Dataset retrieved 2 October 2025.

[133] B. Nesset, M. Romeo, G. Rajendran, and H. Hastie. Robot Broken Promise? Repair strategies for mitigating loss of trust for repeated failures. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1389–1395, Busan, Korea, Republic of, Aug. 2023. IEEE. ISBN 979-8-3503-3670-2. doi: 10.1109/RO-MAN57019.2023.10309558. URL https://ieeexplore.ieee.org/document/10309558/.

[134] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2:100041, 2021. ISSN 2666920X. doi: 10.1016/j.caeai.2021.100041. URL https://linkinghub.elsevier.com/retrieve/pii/S2666920X21000357.

[135] D. A. Norman. *The Psychology of Everyday Things*. Basic Books, New York, 1988. ISBN 0-465-06709-3.

[136] OpenAI. ChatGPT. https://chatgpt.com/, 2026.

[137] L. M. Padilla, I. T. Ruginski, and S. H. Creem-Regehr. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications*, 2(1): 40, Oct. 2017. ISSN 2365-7464. doi: 10.1186/s41235-017-0076-1. URL https://doi.org/10.1186/s41235-017-0076-1.

[138] A. Papenmeier, G. Englebienne, and C. Seifert. How model accuracy and explanation fidelity influence user trust, July 2019. URL http://arxiv.org/abs/1907.12652. arXiv:1907.12652 [cs].

[139] R. Parasuraman and V. Riley. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2):230–253, June 1997. ISSN 0018-7208, 1547-8181. doi: 10.1518/001872097778543886. URL https://journals.sagepub.com/doi/10.1518/001872097778543886.

[140] S. Pareek, E. Velloso, and J. Goncalves. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 546–561, Rio de Janeiro Brazil, June 2024. ACM. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658924. URL https://dl.acm.org/doi/10.1145/3630106.3658924.

[141] S. Pareek, S. Schömbs, E. Velloso, and J. Goncalves. "It's Not the AI's Fault Because It Relies Purely on Data": How Causal Attributions of AI Decisions Shape Trust in AI Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–18, New York, NY, USA, Apr. 2025. Association for Computing Machinery. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3713468. URL https://dl.acm.org/doi/10.1145/3706598.3713468.

[142] J. S. Park, R. Barber, A. Kirlik, and K. Karahalios. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–15, Nov. 2019. ISSN 2573-0142. doi: 10.1145/3359204. URL https://dl.acm.org/doi/10.1145/3359204.

[143] S. Passi and M. Vorvoreanu. Overreliance on AI: Literature Review. Technical Report MSR-TR-2022-12, Microsoft, June 2022. URL https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/.

[144] I. Paul, S. Mohanty, M. Wadhwa, and J. Parker. Swipe right: When and why conservatives are more accepting of AI recommendations. *Journal of Consumer Psychology*, page jcpy.1461, May 2025. ISSN 1057-7408, 1532-7663. doi: 10.1002/jcpy.1461. URL https://myscp.onlinelibrary.wiley.com/doi/10.1002/jcpy.1461.

[145] L. Perkins, J. E. Miller, A. Hashemi, and G. Burns. Designing for Human-Centered Systems: Situational Risk as a Factor of Trust in Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54(25):2130–2134, Sept. 2010. ISSN 1071-1813. doi: 10.1177/154193121005402502. URL https://doi.org/10.1177/154193121005402502.

[146] S. A. C. Perrig, N. Scharowski, and F. Brühlmann. Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. In *Extended Abstracts of the 2023 CHI Conference*

on Human Factors in Computing Systems, pages 1–7, Hamburg Germany, Apr. 2023. ACM. ISBN 978-1-4503-9422-2. doi: 10.1145/3544549.3585808. URL https://dl.acm.org/doi/10.1145/3544549.3585808.

[147] E. Peters, N. F. Dieckmann, D. Västfjäll, C. K. Mertz, P. Slovic, and J. H. Hibbard. Bringing meaning to numbers: The impact of evaluative categories on decisions. *Journal of Experimental Psychology: Applied*, 15(3):213–227, 2009. ISSN 1939-2192, 1076-898X. doi: 10.1037/a0016978. URL https://doi.apa.org/doi/10.1037/a0016978.

[148] E. Peters, P. S. Hart, and L. Fraenkel. Informing Patients: The Influence of Numeracy, Framing, and Format of Side Effect Information on Risk Perceptions. *Medical Decision Making*, 31(3):432–436, May 2011. ISSN 0272-989X, 1552-681X. doi: 10.1177/0272989X10391672. URL https://journals.sagepub.com/doi/10.1177/0272989X10391672.

[149] S. Prabhudesai, L. Yang, S. Asthana, X. Huan, Q. V. Liao, and N. Banovic. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 379–396, New York, NY, USA, Mar. 2023. Association for Computing Machinery. ISBN 979-8-4007-0106-1. doi: 10.1145/3581641.3584033. URL https://doi.org/10.1145/3581641.3584033.

[150] S. Prentice-Dunn and R. W. Rogers. Protection Motivation Theory and preventive health: beyond the Health Belief Model. *Health Education Research*, 1(3):153–161, 1986. ISSN 0268-1153, 1465-3648. doi: 10.1093/her/1.3.153. URL https://academic.oup.com/her/article-lookup/doi/10.1093/her/1.3.153.

[151] P. S. B. Rao, S. Šćepanović, K. Zhou, E. P. Bogucka, and D. Quercia. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–26, Yokohama Japan, Apr. 2025. ACM. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3713979. URL https://dl.acm.org/doi/10.1145/3706598.3713979.

[152] A. Rechkemmer and M. Yin. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, New Orleans LA USA, Apr. 2022. ACM. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3501967. URL https://dl.acm.org/doi/10.1145/3491102.3501967.

[153] L. Reilly and H. Gold. A 'rogue employee' was behind Grok's unprompted 'white genocide' mentions. *CNN Business*, May 2025. URL https://edition.cnn.com/2025/05/16/business/a-rogue-employee-was-behind-groks-unprompted-white-genocide-mentions. Online article.

[154] T. Rieger, L. Kugler, D. Manzey, and E. Roesler. The (Im)perfect Automation Schema: Who Is Trusted More, Automated or Human Decision Support? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 66(8):1995–2007, Aug. 2024. ISSN 0018-7208, 1547-8181. doi: 10.1177/00187208231197347. URL https://journals.sagepub.com/doi/10.1177/00187208231197347.

[155] T. Rieger, H. Schindler, L. Onnasch, and E. Roesler. Explaining AI weaknesses improves human–AI performance in a dynamic control task. *International Journal of Human-Computer Studies*, 199:103505, May 2025. ISSN 10715819. doi: 10.1016/j.ijhcs.2025.103505. URL https://linkinghub.elsevier.com/retrieve/pii/S107158192500062X.

[156] E. Roesler, D. Manzey, and L. Onnasch. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, 6(58):eabj5425, Sept. 2021. doi: 10.1126/scirobotics.abj5425. URL https://www.science.org/doi/abs/10.1126/scirobotics.abj5425.

[157] E. Roesler*, T. Rieger*, and D. Manzey. Trust towards Human vs. Automated Agents: Using a Multidimensional Trust Questionnaire to Assess The Role of Performance, Utility, Purpose, and Transparency. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1):2047–2051, Sept. 2022. ISSN 1071-1813, 2169-5067. doi: 10.1177/1071181322661065. URL https://journals.sagepub.com/doi/10.1177/1071181322661065.

[158] E. Roesler, D. Manzey, and L. Onnasch. Embodiment Matters in Social HRI Research: Effectiveness of Anthropomorphism on Subjective and Objective Outcomes. *ACM Transactions on Human-Robot Interaction*, 12(1):7:1–7:9, Feb. 2023. doi: 10.1145/3555812. URL https://dl.acm.org/doi/10.1145/3555812.

[159] R. W. Rogers. A Protection Motivation Theory of Fear Appeals and Attitude Change1. *The Journal of Psychology*, 91(1):93–114, Sept. 1975. ISSN 0022-3980, 1940-1019. doi: 10.1080/00223980.1975.9915803. URL https://www.tandfonline.com/doi/full/10.1080/00223980.1975.9915803.

[160] J. B. Rotter. Internal versus external control of reinforcement: A case history of a variable. *American Psychologist*, 45(4):489–493, 1990. ISSN 0003-066X.

[161] R. A. Ruiter, C. Abraham, and G. Kok. Scary warnings and rational precautions: A review of the psychology of fear appeals. *Psychology & Health*, 16(6):613–630, Nov. 2001. ISSN 0887-0446. doi: 10.1080/08870440108405863. URL https://doi.org/10.1080/08870440108405863. _eprint: https://doi.org/10.1080/08870440108405863.

[162] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, Harlow, United Kingdom, 4th, global edition, 2022. ISBN 978-1-292-40113-3.

[163] N. Scharowski, S. A. C. Perrig, N. Von Felten, L. F. Aeschbach, K. Opwis, P. Wintersberger, and F. Brühlmann. To Trust or Distrust AI: A Questionnaire Validation Study. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 361–374, Athens Greece, June 2025. ACM. ISBN 979-8-4007-1482-5. doi: 10.1145/3715275.3732025. URL https://dl.acm.org/doi/10.1145/3715275.3732025.

[164] O. Schilke and M. Reimann. The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188:104405, May 2025. ISSN 07495978. doi: 10.1016/j.obhdp.2025.104405. URL https://linkinghub.elsevier.com/retrieve/pii/S0749597825000172.

[165] N. Schlicker, K. Baum, A. Uhde, S. Sterz, M. C. Hirsch, and M. Langer. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *Computers in Human Behavior*, 170:108671, Sept. 2025. ISSN 07475632. doi: 10.1016/j.chb.2025.108671. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563225001189.

[166] N. Schlicker, F. Lechner, K. Wehrle, B. Greulich, M. C. Hirsch, and M. Langer. Trustworthy enough? Examining trustworthiness assessments of large language model-based medical agents. *Technology, Mind, and Behavior*, 6(2):144–172, 2025. ISSN 2689-0208. doi: 10.1037/tmb0000164. URL https://doi.apa.org/doi/10.1037/tmb0000164.

[167] S. Schömbs, J. Klein, and E. Roesler. Feeling with a robot—the role of anthropomorphism by design and the tendency to anthropomorphize in human-robot interaction. *Frontiers in Robotics and AI*, 10, June 2023. ISSN 2296-9144. doi: 10.3389/frobt.2023.1149601. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2023.1149601/full.

[168] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[169] S. D. Shaw and G. Nave. Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender. Jan. 2026. doi: 10.31234/osf.io/yk25n_v1. URL https://ssrn.com/abstract=6097646.

[170] R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, J. Gallegos, A. Smart, E. Garcia, and G. Virk. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741, Montr\'{e}al QC Canada, Aug. 2023. ACM. ISBN 979-8-4007-0231-0. doi: 10.1145/3600211.3604673. URL https://dl.acm.org/doi/10.1145/3600211.3604673.

[171] H. Shen, T. Knearem, R. Ghosh, K. Alkiek, K. Krishna, Y. Liu, Z. Ma, S. Petridis, Y.-H. Peng, L. Qiwei, S. Rakshit, C. Si, Y. Xie, J. P. Bigham, F. Bentley, J. Chai, Z. Lipton, Q. Mei, R. Mihalcea, M. Terry, D. Yang, M. R. Morris, P. Resnick, and D. Jurgens. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions, Aug. 2024. URL http://arxiv.org/abs/2406.09264. arXiv:2406.09264 [cs].

[172] D. Shin and F. Akhtar. Algorithmic Inoculation Against Misinformation: How to Build Cognitive Immunity Against Misinformation. *Journal of Broadcasting & Electronic Media*, 68(2):153–175, Mar. 2024. ISSN 0883-8151, 1550-6878. doi: 10.1080/08838151.2024.2323712. URL https://www.tandfonline.com/doi/full/10.1080/08838151.2024.2323712.

[173] S. B. Sial, M. B. Sial, Y. Ayaz, S. I. A. Shah, and A. Zivanovic. Interaction of robot with humans by communicating simulated emotional states through expressive movements. *Intelligent Service Robotics*, 9(3):231–255, July 2016. ISSN 1861-2776, 1861-2784. doi: 10.1007/s11370-016-0199-0. URL http://link.springer.com/10.1007/s11370-016-0199-0.

[174] T. Sippy, H. Margetts, R. Modhvadia, O. Bruff, J. Perez, and A. Dutta. BSA 42 | Artificial Intelligence: How Political Orientation Shapes Attitudes to AI. Report 42, National Centre for Social Research, London, UK, July 2025. URL https://natcen.ac.uk/publications/bsa-42-artificial-intelligence. Series: British Social Attitudes.

[175] P. Slattery, A. K. Saeri, E. A. C. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, and N. Thompson. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence, Apr. 2025. URL http://arxiv.org/abs/2408.12622. arXiv:2408.12622 [cs].

[176] P. Slattery, A. K. Saeri, E. A. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, and N. Thompson. The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, 7(5):101517, May 2026. ISSN 26663899. doi: 10.1016/j.patter.2026.101517. URL https://linkinghub.elsevier.com/retrieve/pii/S2666389926000267.

[177] P. Slovic. Perception of Risk. *Science*, 236(4799):280–285, Apr. 1987. doi: 10.1126/science.3563507. URL https://www.science.org/doi/abs/10.1126/science.3563507.

[178] P. Slovic. *The Perception of Risk*. Routledge, 1 edition, 2000. doi: 10.4324/9781315661773. URL https://doi.org/10.4324/9781315661773.

[179] P. Slovic, B. Fischhoff, and S. Lichtenstein. Facts versus Fears: Understanding Perceived Risk. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgment under Uncertainty: Heuristics and Biases*, pages 463–492. Cambridge University Press, Cambridge, 1982. ISBN 0-521-24064-6.

[180] P. Slovic, B. Fischhoff, and S. Lichtenstein. Why Study Risk Perception? *Risk Analysis*, 2(2):83–93, June 1982. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.1539-6924.1982.tb01369.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.1982.tb01369.x.

[181] P. Slovic, B. Fischhoff, and S. Lichtenstein. The Psychometric Study of Risk Perception. In V. T. Covello, J. Menkes, and J. Mumpower, editors, *Risk Evaluation and Management*, pages 3–24. Springer US, Boston, MA, 1986. ISBN 978-1-4613-2103-3. doi: 10.1007/978-1-4613-2103-3_1. URL https://doi.org/10.1007/978-1-4613-2103-3_1.

[182] P. Slovic, J. Monahan, and D. G. MacGregor. Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior*, 24(3):271–296, June 2000. ISSN 1573-661X, 0147-7307. doi: 10.1023/A:1005595519944. URL https://doi.apa.org/doi/10.1023/A:1005595519944.

[183] P. Slovic, M. L. Finucane, E. Peters, and D. G. MacGregor. Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality. *Risk Analysis*, 24(2):311–322, Apr. 2004. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.0272-4332.2004.00433.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.0272-4332.2004.00433.x.

[184] C. V. Smedegaard. Reframing the Role of Novelty within Social HRI: from Noise to Information. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 411–420, Daegu, Korea (South), Mar. 2019. IEEE. ISBN 978-1-5386-8555-6. doi: 10.1109/HRI.2019.8673219. URL https://ieeexplore.ieee.org/document/8673219/.

[185] A. Spence, W. Poortinga, and N. Pidgeon. The Psychological Distance of Climate Change. *Risk Analysis*, 32(6):957–972, June 2012. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.1539-6924.2011.01695.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.2011.01695.x.

[186] K. E. Stanovich and M. E. Toplak. Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*, 11(1):3–13, June 2012. ISSN 1860-1839. doi: 10.1007/s11299-011-0093-6. URL https://doi.org/10.1007/s11299-011-0093-6.

[187] K. E. Stanovich and R. F. West. Advancing the rationality debate. *Behavioral and Brain Sciences*, 23(5):701–717, Oct. 2000. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X00623439. URL https://www.cambridge.org/core/product/identifier/S0140525X00623439/type/journal_article.

[188] C. Starr. Social Benefit versus Technological Risk: What is our society willing to pay for safety? *Science*, 165(3899):1232–1238, Sept. 1969. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.165.3899.1232. URL https://www.science.org/doi/10.1126/science.165.3899.1232.

[189] R. E. Stuck, B. J. Tomlinson, and B. N. Walker. The importance of incorporating risk into human-automation trust. *Theoretical Issues in Ergonomics Science*, 23(4):500–516, July 2022. ISSN 1463-922X, 1464-536X. doi: 10.1080/1463922X.2021.1975170. URL https://www.tandfonline.com/doi/full/10.1080/1463922X.2021.1975170.

[190] J. Sutton, S. C. Vos, M. M. Wood, and M. Turner. Designing Effective Tsunami Messages: Examining the Role of Short Messages and Fear in Warning Response. *Weather, Climate, and Society*, 10(1):75–87, Jan. 2018. ISSN 1948-8327, 1948-8335. doi: 10.1175/WCAS-D-17-0032.1. URL https://journals.ametsoc.org/doi/10.1175/WCAS-D-17-0032.1.

[191] D. Szafir and D. A. Szafir. Connecting Human-Robot Interaction and Data Visualization. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21, pages 281–292, New York, NY, USA, Mar. 2021. Association for Computing Machinery. ISBN 978-1-4503-8289-2. doi: 10.1145/3434073.3444683. URL https://dl.acm.org/doi/10.1145/3434073.3444683.

[192] D. R. Thomas. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2):237–246, June 2006. ISSN 1098-2140, 1557-0878. doi: 10.1177/1098214005283748. URL https://journals.sagepub.com/doi/10.1177/1098214005283748.

[193] B. Toff and F. M. Simon. "Or They Could Just Not Use It?": The Dilemma of AI Disclosure for Audience Trust in News. *The International Journal of Press/Politics*, 30(4):881–903, Oct. 2025. ISSN 1940-1612, 1940-1620. doi: 10.1177/19401612241308697. URL https://journals.sagepub.com/doi/10.1177/19401612241308697.

[194] S. Tolmeijer, U. Gadiraju, R. Ghantasala, A. Gupta, and A. Bernstein. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 77–87, Utrecht Netherlands, June 2021. ACM. ISBN 978-1-4503-8366-0. doi: 10.1145/3450613.3456817. URL https://dl.acm.org/doi/10.1145/3450613.3456817.

[195] C. S. Traberg, J. Roozenbeek, and S. Van Der Linden. Psychological Inoculation against Misinformation: Current Evidence and Future Directions. *The ANNALS of the American Academy of Political and Social Science*, 700(1):136–151, Mar. 2022. ISSN 0002-7162, 1552-3349. doi: 10.1177/00027162221087936. URL https://journals.sagepub.com/doi/10.1177/00027162221087936.

[196] D. R. Traum. Computational Models of Grounding in Collaborative Systems. In *AAAI Technical Report FS-99-03*, pages 124–131, Menlo Park, CA, 1999. AAAI.

[197] Y. Trope and N. Liberman. Construal-level theory of psychological distance. *Psychological Review*, 117 (2):440–463, 2010. ISSN 1939-1471, 0033-295X. doi: 10.1037/a0018963. URL https://doi.apa.org/doi/10.1037/a0018963.

[198] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124. URL https://www.science.org/doi/abs/10.1126/science.185.4157.1124. _eprint: https://www.science.org/doi/pdf/10.1126/science.185.4157.1124.

[199] A. Tversky and D. Kahneman. The Framing of Decisions and the Psychology of Choice. 211, 1981.

[200] A. Tversky and D. Kahneman. Rational Choice and the Framing of Decisions. In B. Karpak and S. Zionts, editors, *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*, pages 81–126, Berlin, Heidelberg, 1989. Springer Berlin Heidelberg. ISBN 978-3-642-74919-3.

[201] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, Oct. 1992. ISSN 1573-0476. doi: 10.1007/BF00122574. URL https://doi.org/10.1007/BF00122574.

[202] R. Uuk, C. I. Gutierrez, D. Guppy, L. Lauwaert, A. Kasirzadeh, L. Velasco, P. Slattery, and C. Prunkl. A Taxonomy of Systemic Risks from General-Purpose AI, Nov. 2024. URL http://arxiv.org/abs/2412.07780. arXiv:2412.07780 [cs].

[203] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1):129:1–129:38, Apr. 2023. doi: 10.1145/3579605. URL https://doi.org/10.1145/3579605.

[204] B. Victor Syiem and E. Velloso. Better Assumptions, Stronger Conclusions: The Case for Ordinal Regression in HCI. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–21, Barcelona Spain, Apr. 2026. ACM. ISBN 979-8-4007-2278-3. doi: 10.1145/3772318.3790821. URL https://dl.acm.org/doi/10.1145/3772318.3790821.

[205] V. Visschers, P. Wiedemann, H. Gutscher, S. Kurzenhäuser, R. Seidl, C. Jardine, and D. Timmermans. Affect-inducing risk communication: current knowledge and future directions. *Journal of Risk Research*, 15(3):257–271, Mar. 2012. ISSN 1366-9877, 1466-4461. doi: 10.1080/13669877.2011.634521. URL http://www.tandfonline.com/doi/abs/10.1080/13669877.2011.634521.

[206] V. H. Visschers and R. M. Meertens. Associative and Cognitive Processes in Risk Perception and Communication. In J. G. Lavino and R. B. Neumann, editors, *Psychology of Risk Perception*, Psychology of Emotions, Motivations, and Actions. Nova Science Publishers, New York, illustrated edition, 2010. ISBN 978-1-60876-960-5. Section: 3.

[207] V. H. M. Visschers and R. M. Meertens. Gut Feeling vs. Common Sense: Associative and Cognitive Processes in Risk Perception and Communication. In *Probabilistic Safety Assessment and Management (PSAM 9)*, Hong Kong, 2008.

[208] V. H. M. Visschers, R. M. Meertens, W. F. Passchier, and N. K. De Vries. An Associative Approach to Risk Perception: Measuring the Effects of Risk Communications Directly and Indirectly. *Journal of Risk Research*, 10(3):371–383, Apr. 2007. ISSN 1366-9877, 1466-4461. doi: 10.1080/13669870701252463. URL http://www.tandfonline.com/doi/abs/10.1080/13669870701252463.

[209] V. H. M. Visschers, R. M. Meertens, W. W. F. Passchier, and N. N. K. De Vries. Probability Information in Risk Communication: A Review of the Research Literature. *Risk Analysis*, 29(2):267–287, Feb. 2009. ISSN 0272-4332, 1539-6924. doi: 10.1111/j.1539-6924.2008.01137.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.2008.01137.x.

[210] H. D. Walpole and R. S. Wilson. Extending a broadly applicable measure of risk perception: the case for susceptibility. *Journal of Risk Research*, 24(2):135–147, Feb. 2021. ISSN 1366-9877, 1466-4461. doi: 10.1080/13669877.2020.1749874. URL https://www.tandfonline.com/doi/full/10.1080/13669877.2020.1749874.

[211] H. D. Walpole and R. S. Wilson. A Yardstick for Danger: Developing a Flexible and Sensitive Measure of Risk Perception. *Risk Analysis*, 41(11):2031–2045, Nov. 2021. ISSN 0272-4332, 1539-6924. doi: 10.1111/risa.13704. URL https://onlinelibrary.wiley.com/doi/10.1111/risa.13704.

[212] X. Wang and M. Yin. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, College Station TX USA, Apr. 2021. ACM. ISBN 978-1-4503-8017-1. doi: 10.1145/3397481.3450650. URL https://dl.acm.org/doi/10.1145/3397481.3450650.

[213] X. Wang and M. Yin. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–36, Dec. 2022. ISSN 2160-6455, 2160-6463. doi: 10.1145/3519266. URL https://dl.acm.org/doi/10.1145/3519266.

[214] Z. J. Wang, C. Kulkarni, L. Wilcox, M. Terry, and M. Madaio. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–40, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 979-8-4007-0330-0. doi: 10.1145/3613904.3642335. URL https://dl.acm.org/doi/10.1145/3613904.3642335.

[215] P. Wason and J. Evans. Dual processes in reasoning? *Cognition*, 3(2):141–154, Jan. 1974. ISSN 00100277. doi: 10.1016/0010-0277(74)90017-1. URL https://linkinghub.elsevier.com/retrieve/pii/0010027774900171.

[216] A. Waytz, J. Cacioppo, and N. Epley. Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232, May 2010. ISSN 1745-6916. doi: 10.1177/1745691610369336. URL https://doi.org/10.1177/1745691610369336.

[217] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 214–229, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533088. URL https://dl.acm.org/doi/10.1145/3531146.3533088.

[218] L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. Isaac. Sociotechnical Safety Evaluation of Generative AI Systems, Oct. 2023. URL http://arxiv.org/abs/2310.11986. arXiv:2310.11986 [cs].

[219] C. D. Wickens, S. Rice, D. Keller, S. Hutchins, J. Hughes, and K. Clayton. False Alerts in Air Traffic Control Conflict Alerting System: Is There a "Cry Wolf" Effect? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 51(4):446–462, Aug. 2009. ISSN 0018-7208, 1547-8181. doi: 10.1177/0018720809344720. URL https://journals.sagepub.com/doi/10.1177/0018720809344720.

[220] R. Wiczorek and D. Manzey. Supporting Attention Allocation in Multitask Environments: Effects of Likelihood Alarm Systems on Trust, Behavior, and Performance. *Human Factors*, 56(7):1209–1221, Nov. 2014. ISSN 0018-7208. doi: 10.1177/0018720814528534. URL https://doi.org/10.1177/0018720814528534.

[221] R. S. Wilson, A. Zwickle, and H. Walpole. Developing a Broadly Applicable Measure of Risk Perception. *Risk Analysis*, 39(4):777–791, Apr. 2019. ISSN 0272-4332, 1539-6924. doi: 10.1111/risa.13207. URL https://onlinelibrary.wiley.com/doi/10.1111/risa.13207.

[222] K. Xu, Y. Shen, L. Yan, and Y. Ren. Cognitive Agency Surrender: Defending Epistemic Sovereignty via Scaffolded AI Friction, Apr. 2026. URL http://arxiv.org/abs/2603.21735. arXiv:2603.21735 [cs.HC].

[223] R. V. Yampolskiy. Taxonomy of Pathways to Dangerous Artificial Intelligence. In *Workshops of the Thirtieth AAAI Conference on Artificial Intelligence: AI, Ethics, and Society*, Technical Report WS-16-02, pages 143–148, Phoenix, Arizona, USA, 2016. AAAI Press.

[224] S. Yang, N. M. Krause, L. Bao, M. N. Calice, T. P. Newman, D. A. Scheufele, M. A. Xenos, and D. Brossard. In AI We Trust: The Interplay of Media Use, Political Ideology, and Trust in Shaping Emerging AI Attitudes. *Journalism & Mass Communication Quarterly*, 102(2):382–406, June 2025. ISSN 1077-6990, 2161-430X. doi: 10.1177/10776990231190868. URL https://journals.sagepub.com/doi/10.1177/10776990231190868.

[225] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Glasgow Scotland Uk, May 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300509. URL https://dl.acm.org/doi/10.1145/3290605.3300509.

[226] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 295–305, New York, NY, USA, Jan. 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372852. URL https://doi.org/10.1145/3351095.3372852.

[227] B. J. Zikmund-Fisher, H. O. Witteman, M. Dickson, A. Fuhrel-Forbis, V. C. Kahn, N. L. Exe, M. Valerio, L. G. Holtzman, L. D. Scherer, and A. Fagerlin. Blocks, Ovals, or People? Icon Type Affects Risk Perceptions and Recall of Pictographs. *Medical Decision Making*, 34(4):443–453, May 2014. ISSN 0272-989X. doi: 10.1177/0272989X13511706. URL https://doi.org/10.1177/0272989X13511706.