Artificial Intelligence in Human Domains — Experimental Evidence

Zur Erlangung des akademischen Grades einer

Doktorin der Wirtschaftswissenschaften
Dr. rer. pol.

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

M.Sc. Nicola Hüholt

Tag der mündlichen Prüfung:	23.04.2026
Referent:	Prof. Dr. Clemens Puppe
Korreferentin:	Prof. Dr. Jella Pfeiffer

Karlsruhe, 2026

Acknowledgments

This thesis would not have been possible without the support of many people, and I would like to express my sincere gratitude to all of them.

First, I would like to thank Prof. Dr. Nora Szech, who first encouraged me to pursue a PhD and convinced me that research was the right path for me. She was an inspiring mentor — sharp, open-minded, and truly interested in others' ideas, with an infectious enthusiasm for learning. Above all, I admired how her approach to research was driven by the desire to contribute to a better world. Her passing is a great loss, and I feel fortunate to have learned from her.

I am profoundly indebted to Prof. Dr. Clemens Puppe, who assumed responsibility for supervising my dissertation following Nora's passing. I am especially thankful that he stepped in at a difficult moment and accompanied my dissertation with great care and commitment. His constructive and precise feedback, well-judged advice, and steady guidance were invaluable and played a central role in bringing this dissertation to completion. Whenever questions came up, he made time to discuss them and offered clear direction that helped me move forward with confidence and focus. I am sincerely grateful for his supervision.

I would like to express my deep gratitude to Prof. Dr. Jella Pfeiffer for her mentorship. At an important time in my PhD, her academic guidance helped me maintain direction and momentum in my work, and I greatly appreciated her personal thoughtfulness. I am especially grateful that she opened new research opportunities for me, including the development of a joint project and my involvement in another project. Her insightful perspective and constructive input have been extremely helpful, and I am truly thankful for the trust she placed in me.

I also thank Hannah and Pascal for their excellent collaboration and ongoing feedback. Their interdisciplinary perspectives in psychology and computer science consistently strengthened our shared work.

I would like to thank my colleagues at the Chair of Political Economy. I am grateful to Lixuan for sharing the PhD journey with me and for her encouragement and shared experiences, including our trip to Singapore. I also thank Hannes and Frank for always making time for advice and for sharing their experience. I owe special thanks to Sibille for being unfailingly helpful and for always knowing what to do — from

administrative and organizational matters to proofreading. I also thank Anke, who shared large parts of this journey with me, across different stages and places, and whose generosity, practical help, and steady support made a real difference.

My deepest thanks go to my parents and my brother for their unwavering support, encouragement, and for sparking my interest in learning and knowledge.

Last but definitely not least, I want to thank Patrik. You supported me with care and thoughtfulness, often in ways I did not even know to ask for. Your unwavering belief that I can do whatever I set out to do meant a great deal to me, and your support helped me keep perspective and a sense of purpose whenever the finish line seemed far away. I am deeply grateful that you left your home in Sweden to build a new one here with me so that I could complete this dissertation.

Abstract

This dissertation examines how individuals interact with artificial intelligence (AI) as it becomes embedded in human domains, such as moral decision-making. It combines preregistered experiments with an analysis of ChatGPT prompts to study (i) moral delegation, (ii) the demand for explanations for moral AI decisions, and (iii) needs for anthropomorphic design in large language model (LLM) conversational agents for human-like versus computer-like tasks.

The first chapter investigates delegation of a real, other-regarding donation decision in two experiments ($N = 5,639$). Participants either decide themselves or delegate to an AI or another human. Delegation demand is higher when the available delegate is AI, contrary to moral-domain algorithm aversion accounts. Delegation lowers perceived responsibility, more strongly so for AI delegates, and when stakes are real, participants adjust beliefs about AI capability in a self-serving manner that can rationalize offloading. These results indicate that AI can facilitate moral outsourcing and accountability diffusion, particularly when system opacity enables self-justification through belief adaptation about AI capabilities.

The second chapter examines, in two experiments ($N = 393$; $N = 492$), when users access explanations of AI decisions in moral versus neutral contexts, and how affect and motivation shape this information choice. After observing an AI decision in either a trolley-type moral scenario or a matched neutral scenario, participants can choose to view an explanation. Average explanation uptake is high, but moral contexts increase anticipated psychological costs (e.g., anxiety and cognitive dissonance) and the motivation to protect existing attitudes. Curiosity and accuracy-oriented motivation are linked to higher explanation demand. Defense motivation is linked to lower demand, with particularly pronounced avoidance when the AI's decision conflicts with a person's own judgment, producing patterns suggestive of selective engagement.

The third chapter combines three controlled experiments ($N = 624$) with an analysis of 5,394 ChatGPT user prompts to identify task-dependent anthropomorphism preferences. Users prefer more experience-related cues (e.g., warmth, empathy) for human-like tasks but less for computer-like tasks, while desired agency remains comparatively stable. Although ChatGPT adapts anthropomorphic style by task, this adaptation is systematically misaligned with user preferences. Human-like tasks

increase mind perception yet reduce trust, highlighting the need for deliberately calibrated, task-sensitive conversational agent design.

Overall, the dissertation advances research on AI in human domains by showing that AI can be attractive for moral delegation and responsibility offloading, that governance through transparency depends on motivated user engagement, and that anthropomorphic cues should be matched to task-specific needs. These insights inform both AI governance (accountability, oversight, explainability) and the design of widely deployed conversational systems in contexts with ethically relevant real-world consequences.

Contents

List of Figures

List of Tables

Acronyms and Symbols

Acronyms

AI	artificial intelligence
ATI	Affinity for Technology Interaction
AVE	average variance extracted
CA(s)	conversational agent(s)
CB-SEM	covariance-based structural equation modeling
GPT	Generative Pre-trained Transformer (model family used by ChatGPT)
CR	composite reliability
DC	decision congruence
ED	explanation demand
GDPR	General Data Protection Regulation
IDAQ	Individual Differences in Anthropomorphism Questionnaire
IS	Information Systems
LIWC	Linguistic Inquiry and Word Count
LLM	large language model
LR	likelihood-ratio test
MCI	Melbourne Curiosity Inventory
OLS	ordinary least squares
SEM	structural equation modeling
SHAP	SHapley Additive exPlanations

STAI	State–Trait Anxiety Inventory
XAI	explainable artificial intelligence

Statistical symbols and notation

N	sample size
M	mean
SD	standard deviation
SE	standard error
CI	confidence interval
p	p-value
χ^2	chi-square statistic
R^2	coefficient of determination

Introduction and Motivation

Artificial intelligence (AI) is widely debated as the era-defining technology of our time. Especially in light of recent advances in broadly deployable, generative systems, policymakers and researchers increasingly discuss AI as a potential inflection point, on par with the prospective impact of past foundational innovations such as the steam engine, electricity, and computers, and perhaps even beyond, given its potential for autonomous improvement (Agrawal et al., 2019; Banner, 2025; Filippucci et al., 2024a). In economic terms, a natural lens for such claims is the concept of *general-purpose technologies* (GPTs), which describes technologies that can be applied across many sectors, improve over time, and trigger waves of complementary innovation and reorganization rather than just isolated efficiency gains (Bresnahan and Trajtenberg, 1995; Lipsey et al., 2005). Against this benchmark, AI has been argued to exhibit several GPT-like features (Brynjolfsson and McAfee, 2017; Calvino et al., 2025; Crafts, 2021; Filippucci et al., 2024a; Varian, 2019). In public discourse, this sentiment is reflected in broad transformation narratives around work, everyday life, and institutional control, often captured by terms such as the *fourth industrial revolution* or *transformative AI* (Dafoe, 2018; Gruetzemacher and Whittlestone, 2022; Makridakis, 2017; Schwab, 2016; Szczepański, 2019; Zhang and Dafoe, 2019). Precisely because AI is framed as so consequential, a key challenge is to replace conjecture with rigorous empirical measurement of its real-world impacts.

Although the scope and effects of future technological progress are hard to predict, AI is increasingly a cross-cutting technology whose consequences are likely to reach far beyond isolated applications. If it develops along these lines, it will affect economies, societies, and politics globally, as it becomes embedded in decision-making and information environments (Gruetzemacher and Whittlestone, 2022; Maslej et al., 2025). Accordingly, AI entails highly significant opportunities and risks across virtually all sectors. Many existing projections highlight large improvements in labor productivity, efficiency, and automation, implying substantial upside for global economic growth (Chui et al., 2023; Dell'Acqua et al., 2023; Goldman Sachs, 2023; Noy and Zhang, 2023; Szczepański, 2019; Trammell and Korinek, 2023), while more

skeptical perspectives emphasize that macro-level evidence remains limited and therefore expect only modest growth effects (Acemoglu, 2024; Filippucci et al., 2024b; Misch et al., 2026). At a more granular level, AI may revolutionize important sectors such as healthcare (European Commission, 2026; Topol, 2019; World Health Organization, 2024), medicine (Jumper et al., 2021; Rajpurkar et al., 2022; Stanford Institute for Human-Centered Artificial Intelligence (HAI), 2025), scientific discovery (Wang et al., 2023), and education (Chen et al., 2020; Holmes et al., 2023). Even in agriculture, food security (Pandey and Mishra, 2024; Sharma et al., 2021; Zatsu et al., 2024), climate change, and environmental action (International Energy Agency, 2025; Kaack et al., 2022), AI may serve as a catalyst for substantial progress. Set against these potential upsides are serious concerns that likewise cut across domains. AI can produce discriminatory and biased outcomes in sensitive domains (Mehrabi et al., 2021), result in large-scale replacement of human labor and rising inequality through a widening gap between capital and labor incomes (Acemoglu, 2024; Federspiel et al., 2023; Korinek and Stiglitz, 2017), erode trust in shared information by enabling scalable misinformation and propaganda (Weidinger et al., 2021), raise privacy and intellectual property concerns (Carlini et al., 2021; Shokri et al., 2017), create accountability and oversight gaps through opaque decision-making (Kroll et al., 2016; Rudin, 2019), and encourage overreliance on automated outputs (Buçinca et al., 2021; Zhai et al., 2024). Considering such enormous potential impact in either direction, AI is increasingly seen as a technology so broadly transformative that its regulation and governance carries the highest priority to realize benefits while preventing severe negative repercussions.

Policy and law address this by anchoring governance in high-level principles aimed at promoting human well-being, human rights, and human-centered values such as fairness, transparency, and accountability, and by translating these aims into enforceable organizational duties and oversight mechanisms (Council of Europe, 2024; EU AI Act, 2024; OECD, 2019; UNESCO, 2021). In practice, this translation takes the form of life-cycle governance requirements that support traceability and control, e.g., risk management, defined roles and responsibilities, and standardized documentation and logging, with frameworks like the NIST AI Risk Management Framework (NIST RMF) offering a common structure for implementation across organizations (National Institute of Standards and Technology, 2023). Research likewise argues that credible AI governance depends on institutionalized accountability and auditability rather than voluntary commitments, and it points to concrete instruments such as algorithmic audits and standardized documentation of models and datasets as prerequisites for

effective supervision and compliance (Dafoe, 2018; Gebru et al., 2018; Kroll et al., 2016; Mitchell et al., 2019; Raji et al., 2020; Vinuesa et al., 2020). Crucially, whether governance goals can be met through such measures hinges on how people actually use, interpret, and contest AI outputs in practice. This makes behavioral responses a central empirical target, spanning how people engage with AI-supported decisions, calibrate trust, allocate responsibility, and decide when to rely on or contest outputs.

The urgency of implementing these governance mechanisms in a timely and effective manner is growing, as the current moment arguably marks a shift from AI as a largely background technology, long hidden in applications such as platform recommender systems, to AI as widely deployed, user-facing infrastructure (Covington et al., 2016). Adoption and use have scaled rapidly in consumer settings. For instance, OpenAI reports that ChatGPT, a flagship large language model (LLM)-based chatbot, serves more than 800 million weekly active users, and independent population surveys likewise indicate rapidly rising direct use of AI chatbots in the general population (Bick et al., 2026; OpenAI, 2025; Sidoti and McClain, 2025). In parallel, organizational diffusion has increased markedly, as survey evidence shows large year-on-year increases in reported AI use and in the share deploying generative AI in at least one business function (Maslej et al., 2025). Official statistics likewise document broad uptake across enterprises, including substantially higher adoption among large firms, while complementary survey evidence suggests that diffusion increasingly extends to small and medium-sized enterprises (Eurostat, 2025; Organisation for Economic Co-operation and Development, 2025). Perhaps most importantly for everyday salience, foundation-model-based generative systems are designed to generalize across many tasks (Bommasani et al., 2021) and increasingly appear not as specialized tools but as built-in features of widely used software ecosystems and digital workflows, making routine interaction with AI more likely (Apple, 2024; Davuluri, 2024; Google Workspace, 2024). With uptake accelerating, keeping evidence in step with deployment is important, as routines and dependencies can form quickly and shape how AI is used and governed in practice.

Importantly, AI systems are increasingly entering domains that were long treated as quintessentially *human* and closely tied to human identity, including creative expression, open-ended reasoning, social interaction, and moral judgment (Awad et al., 2018; Bommasani et al., 2021; Elgammal et al., 2017; Haslam, 2006; Mittelstadt et al., 2016; Zhou and Lee, 2024). This is reflected in the roles and tasks they perform. For instance, foundation models can generate and revise text, propose arguments,

explain choices, and iterate creatively, enabling them to contribute to cognitive labor in a way that resembles a human-like collaborator rather than a narrow tool, even in tasks that previously depended primarily on human expertise (Acemoglu and Restrepo, 2019; Autor, 2015; Bender et al., 2021; Bommasani et al., 2021; Brynjolfsson et al., 2025; Eloundou et al., 2023). In institutional settings, AI is likewise increasingly situated in roles that resemble human discretion and authority — screening, ranking, prioritizing, and recommending in domains such as healthcare allocation, hiring and credit, criminal justice, and drone warfare (Dressel and Farid, 2018; Kleinberg et al., 2018; Obermeyer et al., 2019; Raghavan et al., 2020). By selecting objectives, proxies, constraints, and error trade-offs, these systems operationalize distributional priorities and ethically consequential judgments that were traditionally exercised by human decision-makers, motivating accounts of algorithms as moral proxies and, in some frameworks, artificial moral agents (Anderson and Anderson, 2007; Barocas and Selbst, 2016; Floridi and Sanders, 2004; Mittelstadt et al., 2016; Moor, 2006). At the same time, AI is increasingly perceived as a human-like interaction partner. Minimal social cues (e.g., politeness, conversational framing) elicit social responses to machines, and empathic expressions further intensify perceived warmth and human-likeness (Epley et al., 2007; Liu and Sundar, 2018; Nass and Moon, 2000; Reeves and Nass, 1996; Waytz et al., 2010a). Consistent with mind perception theory, such cues shape attributions of a system's agency (capacity for intentional action) and experience (capacity for feelings), which in turn structures whether users treat an entity as a social actor and as morally relevant (Gray et al., 2007; Gray and Wegner, 2012). Modern conversational agents leverage natural language and social cues that foster social presence and perceived humanness, and can even increase trust, thereby shaping user expectations and behavior toward the system (Araujo, 2018; Cohn et al., 2024; Feine et al., 2019). AI-generated conversations can sound so convincingly human that people can no longer reliably tell them apart from human-written text (Casal and Kessler, 2023). This development is particularly salient in companionship and care applications, including character-based dialogue systems and therapeutic chatbots, where users engage in self-disclosure and may develop relationship-like emotional attachment, altering how they interact with and rely on these systems (Fitzpatrick et al., 2017; Laranjo et al., 2018; Pentina et al., 2023; Skjuve et al., 2021; Zhou and Lee, 2024).

This diffusion of AI into human domains is especially delicate, as these domains are often directly tied to human well-being and fundamental rights and involve high risks (Dressel and Farid, 2018; EU AI Act, 2024; National Institute of Standards

and Technology, 2023; Obermeyer et al., 2019; Rudin, 2019). In these settings AI is not judged simply by accuracy but by *legitimacy*, *contestability* — the ability for affected people to obtain reasons and meaningfully challenge a decision — and respect for persons (Kroll et al., 2016; Pasquale, 2015). Accordingly, even seemingly technical classifications can function as implicit moral judgments about persons and the distribution of well-being and harm (Anderson and Anderson, 2011; Awad et al., 2018).

Societal harms arise when AI systems become gatekeepers of rights and life chances and thereby shift political and ethical judgments into technical classifications, essentially leading to a form of algorithmic governance and authority (*algocracy*) (Citron and Pasquale, 2014; Danaher, 2016; Eubanks, 2018; Kroll et al., 2016; Yeung, 2017). Categories such as risk, deservingness, need, or merit are value-laden constructs, and translating them into scores is a form of commensuration that relies on proxy measures and therefore requires public justification (Barocas and Selbst, 2016; Espeland and Stevens, 1998). When such proxies are operationalized as objective-seeming scores, they embed a particular moral view into infrastructure through *value inscription* (Akrich, 1992; Friedman and Hendry, 2019; Winner, 2017). This can contribute to *depoliticization*: the underlying moral choices become harder to see and contest, raising concerns of *technological due process* and shifting questions that should be publicly debated into the realm of "what the model says" (Citron, 2007; Citron and Pasquale, 2014; Danaher, 2016; Yeung, 2017). Familiar technical issues like biases become ethically amplified, as models may generate systematic group-level disadvantage (*disparate impact*) without discriminatory intent via proxy discrimination, where seemingly innocent variables correlate with protected traits (Barocas and Selbst, 2016). Additionally, when applied at scale, feedback loops can turn model outputs into self-reinforcing "evidence" by steering surveillance, sanction, or service provision, thereby consolidating structural injustice over time (Barocas and Selbst, 2016; Citron and Pasquale, 2014; Ensign et al., 2018; Lum and Isaac, 2016; O'Neil, 2016), including by shifting humans' perceptual, emotional, and social judgments through repeated interaction with biased AI systems (Glickman and Sharot, 2025).

At the *institutional* level, often opaque or black-boxed AI decision-making provides limited visibility into how outputs are produced and thus limits *contestability*, colliding with due process and professional norms (Ribeiro et al., 2016; Rudin, 2019; Wachter et al., 2017); in domains such as health, care, and welfare, it can also undermine confidentiality and privacy by enabling secondary use or over-collection of sensitive

data. Workflow pressures, i.e., organizational and interface incentives such as time pressure, caseloads, defaults, metrics, or liability concerns, often convert decision support into de facto decision authority, fostering *automation bias* (over-deferring to recommendations and under-checking errors), *overreliance*, and miscalibrated trust, even when uncertainty and contestability are highest (Lee and See, 2004; Parasuraman and Riley, 1997; Rudin, 2019; Skitka et al., 1999). Over time, this can produce *deskilling*, including *moral deskilling*, diminishing individuals' ability to exercise professional and normative judgment through reduced practice and deliberation (Poszler and Lange, 2024; Vallor, 2015; Zhai et al., 2024). At the same time, accountability is structurally destabilized, as responsibility diffuses across designers, deployers, and frontline staff, creating responsibility gaps and *moral crumple zones* in which humans absorb blame without meaningful control (Elish, 2019; Kroll et al., 2016; Matthias, 2004).

For *individuals*, harms extend beyond erroneous outcomes to dignitary and relational harms, including *dehumanization*, i.e., being reduced to a score or case, *human identity threat*, experiencing replaceability in morally loaded roles, and seeing moral agency eroded through *moral outsourcing* (Haslam, 2006; Mirbabaie et al., 2022; Vallor, 2015; Złotowski et al., 2017). In care-like contexts, AI may also encourage relational substitution that replaces or degrades human social ties and may contribute to loneliness (Arnd-Caddigan, 2015). Finally, in these domains anthropomorphic interfaces can intensify vulnerability, enabling subtle forms of influence that would be ethically unacceptable from human counterparts (Susser et al., 2019). For instance, so-called *ELIZA/CASA* effects invite trust, disclosure, and compliance in care-like interactions, while near-human designs may trigger *uncanny valley aversion* precisely where authenticity and responsibility are central (Gray and Wegner, 2012; Lucas et al., 2014; Nass et al., 1994; Weizenbaum, 1976).

The rapid diffusion of AI into human domains creates a distinctive tension: these systems are increasingly adopted for roles that touch core questions of human identity, dignity, and moral agency, while many of the associated risks remain only partially regulated, mitigated, or even fully foreseeable (EU AI Act, 2024; National Institute of Standards and Technology, 2023). This tension is mirrored in markedly ambivalent public reception. Survey evidence shows that people are more concerned than excited about the increased use of AI, worry that it may erode people's ability for creative thinking or to form relationships, and mostly do not support AI use in personal and normatively charged aspects of life, while favoring stronger oversight, especially for uses affecting rights, safety, or vulnerable groups (Ipsos, 2025; Pew Research Center,

2025; Poushter et al., 2025). Experimental work mirrors this caution, documenting *algorithm aversion* and a preference for human discretion when judgments are perceived as subjective, value-laden, or morally consequential (Bigman and Gray, 2018; Castelo et al., 2019; Dietvorst et al., 2015; Jauernig et al., 2022). Yet real-world practice points in the opposite direction. LLM-based conversational systems are increasingly used for companionship and emotional support, including mental-health advice (Chatterji et al., 2025; McBain et al., 2025; Purington et al., 2017) and recent public controversies illustrate the strength of relationship-like attachments to conversational systems (Anguiano, 2025; Silberling, 2025). Evidence from large-scale platform usage data further suggests that very high-intensity *affective use* can correlate with indicators of emotional dependence and reduced offline socialization (Phang et al., 2025). This coexistence of normative skepticism and behavioral uptake — amplified by anthropomorphic design that can both lower barriers to engagement and heighten vulnerability (Akbulut et al., 2024) — raises a central question for human domains: how do people actually respond when AI is positioned as a decision-maker or interaction partner in human domain settings where legitimacy and moral agency matter?

Motivation and Contribution

Answering this question requires empirical evidence, because the relevant outcomes in human domains are behavioral and relational (e.g., delegation, reliance, disclosure, boundary-setting), and because stated attitudes toward AI may diverge from how people act when systems are embedded in real workflows and relationships. Responses to a rapidly evolving technology are also unlikely to be stable. As adoption progresses, expectations, experience, and institutional practices can shift what skepticism, trust, or resistance look like in practice. Empirical work is also needed to disentangle whose behavior is being measured — end users, professionals, and those affected by decisions face different incentives and vulnerabilities — and to identify potentially counterintuitive patterns and friction points that cannot be explained by material payoffs alone, since behavior in these settings is shaped by additional factors such as identity, moral self-image, and responsibility concerns. Capturing actual behavior is therefore critical for assessing whether design and governance interventions are effective. Finally, because AI's societal consequences are mediated by human beliefs

and behavioral adjustments as much as by technical capabilities, scholars and policymakers increasingly emphasize combining theory with new empirical strategies to grasp impacts in this evolving field (Björkegren, 2025). Accordingly, this dissertation provides evidence on delegation, oversight behavior, and interaction with LLM-based systems in human-domain settings.

The presented studies contribute to three strands of literature on human responses to AI in normatively consequential settings: work on algorithm aversion and preferences for human discretion (Bigman and Gray, 2018; Castelo et al., 2019; Dietvorst et al., 2015; Jauernig et al., 2022), research on transparency and explanations as governance instruments (Doshi-Velez and Kim, 2017; Kroll et al., 2016; Miller, 2019), and studies on mind perception and anthropomorphic cues in human–AI interaction (Epley et al., 2007; Gray et al., 2007; Gray and Wegner, 2012; Nass and Moon, 2000). Existing evidence for these areas is dispersed across outcomes and contexts, leaving open how delegation, information uptake, and interaction respond in human domain settings. This dissertation contributes behavioral evidence from large-scale experiments and complementary conversational-agent interaction data.

The dissertation comprises three chapters that provide experimental evidence on behavioral responses to AI in human domains from complementary angles. Chapter 1 and Chapter 2 use moral decision-making as a paradigmatic setting in which legitimacy and moral agency are inherently at stake to study governance-relevant behaviors. Importantly, evidence from behavioral economics has long shown that moral choice does not always follow an idealized model of careful deliberation and principled responsibility. Instead, people may seek moral wiggle room (Dana et al., 2006), avoid information that could impose hedonic costs (Golman et al., 2017), and shift responsibility when doing so protects self-image or reduces psychological burden (Bartling and Fischbacher, 2012). Chapters 1 and 2 therefore investigate whether such behavioral patterns likewise emerge for users of AI systems in human domains. Complementing this governance perspective, Chapter 3 turns to user-centered system design in everyday interaction with LLM-based conversational agents and examines how task context (human-like vs. computer-like) and anthropomorphic cues shape mind perception, trust, and reliance in use.

The first chapter examines whether individuals are willing to hand over a morally consequential decision to AI. The motivation is to probe the tension described above: public attitudes tend to be skeptical of AI authority in human-like, normatively charged applications (Bigman and Gray, 2018; Castelo et al., 2019; Dietvorst et al.,

2015; Jauernig et al., 2022), yet these systems are simultaneously diffusing rapidly into everyday use. Building on behavioral-economics evidence that moral choices are often shaped by self-protective motivations and responsibility management (Bartling and Fischbacher, 2012), the chapter tests whether such tendencies translate into a willingness to offload moral judgment once AI is conveniently available. This is consequential because it speaks to how quickly high-stakes decisions in human domains may shift from human discretion to AI-mediated processes, with implications for legitimacy, ethical standards, and accountability frameworks. Delegation is a pathway through which the risks discussed for AI in human domains can materialize in practice. Offloading moral judgment can push normative trade-offs into technical classifications and thereby embed contested values while distancing them from societal deliberation. Understanding when delegating morality to AI becomes attractive, and when diffused responsibility is not a malfunction but an outcome users may actively seek, is vital for anticipating and preventing destabilized accountability, moral deskilling, and the displacement of human moral judgment.

The second chapter extends this governance question from who decides to how decisions are overseen. Transparency and explainability are widely advocated as key levers for responsible AI use, but they can only function as safeguards if people actually consult and use them. Thus, Chapter 2 investigates information uptake in AI decision processes as a behavioral choice, comparing how it is shaped by psychological costs and self-regulatory motivations in morally charged compared to neutral contexts. This perspective matters for institutions and policymakers because it clarifies when transparency can enable human oversight and accountability in practice, and when it may fail to prevent the unchecked inscription of value-laden classifications.

The third chapter investigates task-dependent human-likeness in interaction with LLM-based conversational agents across a growing range of applications. It examines what degree and type of anthropomorphic features users want in human-like versus computer-like tasks, and whether current systems' autonomous adaptations align with those preferences. By combining controlled experiments on user needs with evidence from real-world ChatGPT interactions, Chapter 3 identifies systematic mismatches between what users prefer and what the system delivers, and shows why deliberately designed task-sensitive cues (e.g., around agency and experience) can be central for trust in human-like domains. This also makes clear why anthropomorphic design should be deliberately calibrated. The goal is to support trust and usability

while avoiding both aversive distrust and relationship-like dynamics that may increase vulnerability and open manipulation avenues.

Overall, this dissertation contributes to the empirical behavioral economics of AI by providing experimental evidence on how people respond when AI enters human domains, where systems can function as delegates or interaction partners. Methodologically, it combines preregistered large-scale experiments with real-use interaction data from LLM conversational agents. In doing so, it links micro-level behavioral responses to policy-relevant governance levers such as transparency and human oversight, and to design choices that shape trust and engagement in human-like applications.

Research Summary

The following summaries briefly state each chapter's research question, design, and main findings and show how they jointly address behavioral responses to AI in human domains.

Summary of Chapter 1 — Trusting Machines with Morality

Chapter 1, *"Trusting Machines with Morality — Delegating Moral Decisions to AI"*, examines whether people delegate a moral decision to AI once delegation is available, addressing the contrast between documented moral-domain algorithm aversion and the diffusion of AI tools into value-laden decision contexts (Awad et al., 2018; Bigman and Gray, 2018). In two preregistered experiments with a total of 5,639 participants, individuals face a real-life other-regarding donation choice inspired by structural elements of the trolley problem and can either decide themselves or delegate the decision to an AI or to a human counterpart. The results show that delegation demand is significantly higher when the available delegate is an AI rather than a human, indicating that preferences for AI involvement shift when individuals themselves occupy the role of the responsible decision-maker. The evidence is consistent with responsibility shifting as a mechanism, as delegation reduces perceived responsibility and this reduction is stronger when delegating to AI (Bartling and Fischbacher, 2012). The chapter further shows that higher-stakes, real-consequence decisions increase delegation to AI relative to hypothetical scenarios, and that this is accompanied by self-serving belief adaptation in perceived AI capability, which helps to justify offloading

responsibility under ambiguity. This pattern is consistent with the idea that opacity and uncertainty surrounding AI capabilities create room to justify delegation as appropriate rather than evasive. By demonstrating that responsibility shifting extends from human to AI delegates and may even be facilitated by AI-specific ambiguity, the findings nuance standard accounts of moral algorithm aversion and highlight governance challenges around accountability when AI is used as a convenient moral offloading device.

Summary of Chapter 2 — How Affect and Motivations Shape Demand for Explanations

Chapter 2, *"How Affect and Motivations Shape Demand for Explanations in Moral AI Decisions"*, examines when people choose to consult explanations of AI decisions in morally charged contexts, and which affective and motivational mechanisms shape this behavioral prerequisite for explainability-based governance. Building on accounts of information demand as a value-based choice — where expected epistemic and hedonic benefits can be offset by anticipated psychological costs — the chapter tests whether moral contexts may foster information avoidance or produce selective engagement driven by motivated reasoning and congeniality concerns (Caplin and Leahy, 2001; Golman et al., 2017; Kunda, 1990; Loewenstein, 1994; Sharot and Sunstein, 2020). Morality may simultaneously raise epistemic interest (accuracy goals) and threat-related avoidance (defense goals), making its overall effect on explanation uptake theoretically unclear.

Empirically, we test these competing pathways in two online experiments where participants observe an AI decision in either a trolley-type moral scenario or a matched neutral property-damage scenario, and then choose whether they want to view an explanation or not. The design isolates the hedonic and self-regulatory value of transparency by making explanation access free and non-instrumental, while holding time costs constant regardless of the choice. Across both experiments, explanation demand is high and does not differ on average between moral and neutral contexts, suggesting that morality does not uniformly reduce or increase transparency uptake. However, moral context still matters because it reliably shifts the underlying affective trade-off, increasing anticipated psychological costs and heightening defense-related motivation. The first experiment (N = 393) shows that moral scenarios increase anticipated anxiety, yet anxiety does not translate into

lower explanation seeking. Instead, curiosity robustly predicts uptake and moral context operates primarily through small, context-dependent changes in curiosity. Exploratory analyses further suggest that context effects on curiosity and information demand depend on whether the AI decision is congruent with participants' own initial judgment. The second experiment ($N = 492$) addresses this by balancing decision congruence within each context and additionally measuring accuracy- and defense-related motivations. Congruence between AI and participant decisions robustly predicts lower explanation demand. Mechanism tests show that moral contexts increase defense-related motivation, and that motivations are systematically linked to information choice: accuracy motivation predicts higher explanation demand, whereas defense motivation predicts lower explanation demand. The pattern is consistent with congeniality-based selectivity in the sense that defensiveness appears to suppress explanation demand particularly when the AI decision conflicts with participants' own judgment, although evidence for this boundary condition remains suggestive rather than robust.

For governance, the evidence points to a potential vulnerability of explainability-based accountability that is strongest in the cases where it matters most. Severe moral weight may shift motivations toward defensiveness, which can suppress explanation demand when the AI decision is incongruent or lead to validation instead of scrutiny.

Summary of Chapter 3 — ChatGPT, Do You Interact Like a Human?

Chapter 3, "*ChatGPT, Do You Interact Like a Human? Investigating Task-Dependent Human-Likeness of LLM-Enabled Conversational Agents Through Mind Perception*", examines how human-like versus computer-like task contexts shape the perceived human-likeness of large language model-enabled conversational agents (LLM CAs), and whether current systems' autonomous behavioral adaptation aligns with users' task-specific preferences. Building on mind perception theory, which distinguishes perceived *experience* and *agency* as separable dimensions of human-likeness, the chapter addresses a key gap in the literature: while prior work has largely treated perceived humanness as a function of the interaction partner's features, the extent to which the task itself influences mind perception and trust in LLM CAs remains underexplored (Bigman and Gray, 2018; Gray et al., 2007; Gray and Wegner, 2012; Yam et al., 2021). This question has become particularly urgent as general-purpose LLM

CAs (e.g., ChatGPT) are adopted across both technical and deeply human-centered domains (Hu, 2023; Kaplan et al., 2023).

The chapter combines three complementary experiments ($N = 624$) that jointly connect (i) user preferences for anthropomorphic design, (ii) task-dependent user needs for experience and agency, and (iii) the actual adaptation behavior of ChatGPT in prolonged interactions. In a lab choice experiment, participants select between a low versus high anthropomorphic CA variant. Preferences shift by task type, with anthropomorphic designs chosen for human-like tasks but less for computer-like tasks. A subsequent online experiment measures task-dependent needs directly and shows that human-like tasks increase users' desired experience in the interaction partner, whereas the desired level of agency does not systematically differ, indicating a selective rather than uniform demand for human-likeness. Building on these insights, the study examines real interactions with ChatGPT using chat data from 5,394 user prompts. While ChatGPT autonomously adapts its experience and agency cues by task type, this adaptation is misaligned with user preferences, as experience cues remain insufficient and agency cues decrease in human-like tasks. At the same time, human-like tasks themselves have a positive effect on mind perception, leaving a positive total effect on perceived agency and experience that is not fully explained by the measured language cues. Although users trust ChatGPT less for human-like tasks, this trust deficit is partially mitigated through increased mind perception.

Overall, the findings position task type as a key antecedent of mind perception that operates independently of anthropomorphic design manipulations and highlight limits of autonomous cue adaptation in current LLM CAs. By integrating user preferences, user needs, and observed LLM behavior within one framework, the chapter advances a more task-sensitive understanding of anthropomorphism in conversational AI and provides actionable implications for deliberate, user-centered design improvements.

1 *Trusting Machines with Morality — Delegating Moral Decisions to AI*

Abstract

Research suggests that individuals are generally skeptical about the use of artificial intelligence (AI) in moral contexts, favoring human decision-makers over AI. Yet, in two experiments involving a total of 5,639 participants, we find that individuals facing a real-life moral decision delegate significantly more often when they can delegate to AI rather than to a human counterpart. This result highlights AI's relative appeal as a moral delegate, indicating that individuals' preferences for AI's involvement change when they themselves assume the role of a decision-maker. Responsibility shifting, previously studied as a motive for delegation to humans, extends to AI delegates. Moreover, it appears to be facilitated by individuals adapting their beliefs about AI's capability in a self-serving manner. Ambiguity surrounding that capability allows them to interpret it in ways that justify delegation. These findings add nuance to assumptions about algorithm aversion in moral domains and raise critical questions about accountability and the ethical implications of relying on AI for morally sensitive decisions.

[0] This chapter is based on Hüholt and Szech (2026).

1.1 Introduction

Artificial intelligence (AI) has become an integral component across a wide range of industries, many of which intersect with ethical considerations (Bonnefon et al., 2024; Wallach and Allen, 2008). AI tools are already used to make or support decisions in healthcare about the allocation of limited resources (Obermeyer et al., 2019), in finance to determine mortgage or loan eligibility (Hale, 2021; Zou and Khern-am nuai, 2023), and in hiring processes to evaluate job candidates (Dastin, 2022; Dattner et al., 2019). In the most critical cases, they are tasked with making life-and-death decisions (Adam, 2024; Awad et al., 2018; Holbrook et al., 2024). These decisions, involving the distribution of well-being or harm among individuals, are often characterized by ethical trade-offs and therefore fall into the moral domain (Anderson and Anderson, 2011; Awad et al., 2018; Gert, 2005). While such real-world examples showcase AI's expanding role, they also reveal limitations — such as the replication of biases — and the need to account for AI-specific characteristics, such as the opacity of its decision-making (Cath, 2018; Gerke et al., 2020; Pazzanese, 2020). Given the sensitive nature and high stakes of these decisions, coupled with the growing availability of AI tools, questions arise about individuals' willingness to hand over responsibility to AI when faced with morally complex choices.

Our study addresses these questions, and demonstrates that delegation demand in a moral decision is significantly higher when the delegate option is an AI rather than a human. This finding nuances the prevailing notion that individuals generally feel *algorithm aversion* toward the use of AI in moral contexts (Bigman and Gray, 2018; Castelo et al., 2019), showing that aversion can even reverse into greater acceptance compared to a human counterpart. This increased demand to delegate to AI appears unaffected by the severity of the moral dilemma, as we find higher delegation rates to AI irrespective of whether the moral decision is presented in a positive or negative decision context.

Preferences regarding AI's involvement are influenced by individuals' own role in the decision-making process. Unlike prior studies, which often rely on hypothetical dilemmas or scenarios, we employ a *real donation choice* inspired by structural elements of the trolley problem, thereby putting participants into the role of a *decision-maker* with real responsibility and a delegation option. In this setting, we examine if and how the mechanism of *responsibility shifting* — previously studied for delegation to humans (Bartling and Fischbacher, 2012) — extends to AI delegates. We find evidence

that it does and may even be facilitated by AI's intrinsic features, contributing to the increased delegation demand. Opacity and ambiguity surrounding AI's capabilities for making moral decisions may create room for individuals to inflate their beliefs about the quality of AI's moral decision-making to justify transferring responsibility — akin to *moral wiggle room* (Dana et al., 2007) in the mechanistic sense of ambiguity-enabled self-justification. Accordingly, we observe that individuals rate AI's moral capabilities higher when they have the option to delegate to it, particularly when their decision has real consequences.

The results of this study highlight broader societal implications, emphasizing the need for conscious design and governance of AI systems in moral domains to preserve accountability and ethical standards.

Algorithm Aversion

Our findings contrast with extant research which suggests that people are generally skeptical of AI's ability to handle moral decisions and want such decisions to remain under human control. People can be reluctant to rely on algorithms or to allow them to make decisions even when the algorithm outperforms humans, a phenomenon termed *algorithm aversion* by Dietvorst et al. (2015). The term originally refers to deterministic, rule-based algorithms (Dietvorst et al., 2015; Castelo et al., 2019). Later research extends this focus to AI systems (Bigman and Gray, 2018; Jussupow et al., 2020). While some studies use the terms algorithm, AI, machines or automation interchangeably (e.g., Burton et al., 2020), others emphasize AI's unique attributes, such as its perceived mind and its moral reasoning capacity (e.g., Bigman and Gray, 2018; Gogoll and Uhl, 2018; Zhang et al., 2022). Many of the cognitive biases and concerns underlying algorithm aversion — such as distrust in computational decision-making and preference for human-like judgment — also apply to AI systems, though often with additional dimensions (Bigman and Gray, 2018; Zhang et al., 2022).

For decisions with moral aspects, algorithm aversion is especially pronounced (Chugunova and Sele, 2022; Jussupow et al., 2020; Mahmud et al., 2022). A key reason is that individuals perceive AI as lacking the capabilities required for moral decision-making. For instance, Bigman and Gray (2018) document widespread aversion to machines making moral decisions in paradigmatic moral dilemmas across various domains, including driving, legal, medical, and military contexts. Notably, the aversion persists even when machines' decisions lead to positive outcomes (Bigman

and Gray, 2018). This aversion is attributed to the perception that machines lack a complete mind (*mind perception*) needed to make moral decisions (Bigman and Gray, 2018; Gray et al., 2012; Young and Monroe, 2019). *Mind perception* refers to the attribution of mental capacities, and has two dimensions: the ability to fully think (*agency*) and feel (*experience*) (Bigman and Gray, 2018; Gray et al., 2007; Waytz et al., 2010a). Other researchers have found similar results in decisions involving subjective judgment, as well as in different morally sensitive scenarios like personal and impersonal high-stakes moral dilemmas, medical AI or consumer interactions (Castelo et al., 2019; Dietvorst and Bartels, 2022; Longoni et al., 2019; Zhang et al., 2022). Across these domains, concerns consistently stem from AI's perceived lack of essential (human) capabilities — including 'affective human-likeness' and warmth, sensitivity to individual nuances ('uniqueness neglect'), a tendency toward utilitarian or consequentialist reasoning, and insufficient intuition or subjective judgment capability. As a result, individuals perceive algorithmic decisions as less ethical and authentic, and therefore AI to not be suited for subjective tasks (Jago, 2017; Lee, 2018). In line with such concerns, third-parties tend to judge delegation to machines more critically as well, with individuals rewarding the delegation choice less when the delegator selects a machine rather than a human (Gogoll and Uhl, 2018).

Ensuring transparency in decision-making and incorporating a 'human in the loop' as oversight and control mechanism can alleviate some of the aversion (Bigman and Gray, 2018). Measures such as these are also a core component of many legislative frameworks, which emphasize that sensitive moral decisions should ultimately remain in the hands of humans and are accessible to them (GDPR, 2016).

These insights into algorithm aversion suggest that people are critical of both the use of AI for moral decisions and its capability to make them, particularly in hypothetical or observer contexts or when personal stakes are involved. However, preferences can shift depending on context and the individual's role. For example, while people approve of autonomous vehicles programmed to sacrifice passengers to save others, they prefer not to ride in such vehicles themselves (Bonnefon et al., 2016). Analogously, when people act as moral decision-makers, they may accept AI more readily if delegation makes their lives easier. This reflects a highly relevant real-world scenario in which decision-makers have access to AI tools for making or supporting decisions with moral implications.

Delegation and Shifting Responsibility

Demand to delegate moral decisions has already been studied extensively for human delegates. Bartling and Fischbacher (2012) showed that some people prefer to delegate moral decisions to others instead of deciding themselves. A key factor in this behavior is a *shift of responsibility* attribution, both in individuals' own eyes and in the eyes of others. Delegation leads to more selfish behavior and at the same time reduces punishment from third parties (Bartling and Fischbacher, 2012; Coffman, 2011; Hamman et al., 2010), and conversely, reduces rewards for positive or generous decisions (Argenton et al., 2023). Sharing or delegating the decision can reduce feelings of moral responsibility, guilt, or potential regret and help to keep a positive self-image (Bartling et al., 2023; Bartling and Fischbacher, 2012; Falk et al., 2020; Falk and Szech, 2013; Rothenhäusler et al., 2018; Steffel and Williams, 2018).

Delegating to create *moral wiggle room* and exploit ambiguity around the final moral outcome can also shift responsibility, protect self-image, and decrease punishment from others (Bartling et al., 2014; Dana et al., 2007; Grossman and van der Weele, 2017; Serra-Garcia and Szech, 2021). Fahrenwaldt et al. (2024) specify the mechanism behind moral wiggle room as situational features that hinder linking an agent's behavior clearly to self-serving intentions, leaving room for other justifications when behaving selfishly.

When facing a decision for others rather than oneself, self-serving behavior can also be motivated by psychological relief rather than financial reward. The desire to lower feelings of responsibility and regret impacts behavior even when it is not tied to material incentives: people are more likely to delegate then, especially when having to decide between two negative outcomes. In these contexts, the delegate's expertise is secondary — instead, what matters to decision-makers is that responsibility can be transferred (Steffel and Williams, 2018; Steffel et al., 2016). Because AI systems are often opaque and their competence is hard to verify, they may promote such dynamics.

However, while such motives and outcomes are well established for *human delegates*, their applicability to *AI delegates* is still underexplored. It is uncertain whether a desire to shift responsibility or persistent algorithm aversion prevails when delegating moral decisions to AI. Some researchers have raised theoretical concerns about potential ethical risks associated with the use of AI in moral decision-making and how it may affect human behavior. Extensive integration of AI for moral choices may erode

human moral agency and skills, turning people into passive moral patients (Danaher, 2019; Vallor, 2015). It may also facilitate unethical behavior by providing users with psychological distance and reducing guilt, especially when AI acts as a delegate for morally questionable actions (Köbis et al., 2021). The latter 'corruption effect' appears to prove true at least for decisions that directly affect one's own outcome, showing the potential for AI to be used as a scapegoat. For example, individuals do not correct a machine's decision when it serves their own benefit, and sharing a decision with AI increases selfish behavior similarly to sharing it with another human (Kirchkamp and Strobel, 2019; Krügel et al., 2023a). Delegation to AI may also introduce so-called *responsibility gaps*, arising from opacity, complexity, and unpredictability of AI systems, making it unclear who should be held accountable for decision outcomes (Matthias, 2004; Santoni de Sio and Mecacci, 2021).

Empirical studies indicate that the decision-maker's role and responsibility attributions shape delegation of moral choices to AI. Freisinger and Schneider (2025) find that individuals deciding on their own behalf in a fictional layoff decision prefer delegating to AI more than those acting on behalf of others, while affected individuals favor human decision-makers in non-surrogate contexts. Qualitative interviews revealed that alleviating the burden of responsibility was the primary motivation behind delegation to AI.

Findings on responsibility attribution in human–AI comparisons are not clear-cut. Kirchkamp and Strobel (2019) did not find a significant difference between perceived responsibility for purely human teams versus human–AI teams and Dzindolet et al. (2002) found that decision-makers' feelings of moral obligation to follow their own decision may contribute to algorithm aversion. However, other studies highlight systematic differences. Individuals attribute less blame to AI than to humans for the same moral violations (Awad et al., 2020; Shank et al., 2019). Additionally, decision-makers are punished less when delegating a task with a bad outcome to machines rather than humans (Feier et al., 2021). This could potentially incentivize delegation to AI to evade negative judgment from others.

Further evidence shows that people exploit moral wiggle room to protect their self-image, by flexibly attributing more or less moral responsibility to AI depending on whether they themselves are portrayed as the decision-maker or judging others. When evaluating joint human-AI decisions, individuals attribute more agency and responsibility to AI for their own transgressions than for others', resulting in greater moral leniency toward themselves (Dong and Bocian, 2024).

To summarize, individuals may navigate a complex interplay of self-serving motives like responsibility-shifting versus algorithm aversion when delegating moral decisions to AI. We investigate delegation to AI combining insights from delegation theory and behavioral economics. While algorithm aversion literature suggests skepticism toward AI in moral domains especially because of concerns about its capability, delegation to AI may offer a unique pathway for off-loading responsibility. It may incentivize individuals to reinterpret AI's capabilities, that are difficult to quantify, in a more favorable light. By convincing themselves that the AI is better equipped to make the decision, individuals may be able to justify their choice to delegate. Such self-justification may allow them to avoid emotional engagement and the burden of responsibility without feeling guilty about doing so and thus help them maintain a positive self-image. Hence, we aim to address the following two research questions:

Question 1 *Delegation Demand: Is delegating a moral decision to AI more attractive than delegating it to another human?*

Question 2 *Mechanism: Do responsibility shifting and belief adaptation contribute to individuals' willingness to delegate a moral decision to AI?*

The remainder of the paper is structured as follows: Section 1.2 provides an overview of the research design for Studies 1 and 2, followed by a presentation of their respective methods, hypotheses, and results. Section 1.3 discusses the findings on delegation demand and responsibility off-loading to AI, and Section 1.4 provides concluding remarks regarding ethical and practical implications.

1.2 Research Design

To address the outlined research questions, we conducted two online studies.

Study 1 primarily investigates whether delegation demand for a moral decision is higher when individuals can delegate to an AI rather than to another human. In addition, we explore whether this pattern is shaped by decision context. Drawing on previous findings showing that individuals delegate other-regarding decisions more often when outcomes are negative (Steffel et al., 2016), we test whether the severity of the moral dilemma — operationalized through *decision framing* (gain versus loss) — affects the demand for delegation to AI similarly.

Building on these findings, Study 2 corroborates the observed increase in delegation to AI relative to humans in a representative sample and investigates responsibility shifting and belief adaptation as potential mechanisms underlying this pattern. To do so, we manipulate the burden of responsibility associated with the decision task, by varying the *decision impact*, i.e. whether the donation decision has real consequences or remains hypothetical. To test for belief adaptation, participants have to rate AI capabilities for moral decision-making.

Table 1.1 provides an overview of both 2×2 between-subjects designs.

Table 1.1: Overview of experimental design.

Study	**Design** (2×2 between-subject)	**Framing** (Gain/Loss)	**Decision Impact** (RealCons/HypoCons)	**Delegate Option**
Study 1	Delegate × Framing	Gain: *Decide who receives a donation.* Loss: *Decide which donation is "destroyed."*	RealCons: *Donation is paid out for 1 in 10.*	Human AI
Study 2	Delegate × Decision Impact	Gain: *Decide who receives a donation.*	RealCons: *Donation is paid out for 1 in 10.* HypoCons: *Donation is not paid out.*	Human AI

1.2.1 Study 1: Delegation Demand and Framing

Study 1 examines delegation behavior to *AI* delegates compared to *human* delegates.[1] Participants were randomly assigned to one of four treatments in a 2×2 between-subjects design, crossing delegation options — human versus AI — with framing conditions — gain versus loss.

[1] Preregistration at AsPredicted.org: https://aspredicted.org/85y3-ftfp.pdf

1.2.1.1 Procedure and Measures

The study included 800 participants.[2] Participants who answered the comprehension questions about the instructions incorrectly were automatically screened out during the survey and were not able to continue.

Across both studies, the moral decision task individuals were given was a donation choice inspired by structural features of the trolley dilemma — an emblematic scenario in AI ethics research, particularly in the context of self-driving cars (Awad et al., 2018). Like trolley problems, the decision lies firmly within the moral domain, as it involves ethical trade-offs, outcomes affecting others, and the absence of a universally correct answer. The two features shared with the trolley dilemma are (i) a trade-off between helping fewer versus more beneficiaries and (ii) an action–omission framing induced by a preselected default. While the scenario is not an immediate life-or-death act, it has real implications: both options reduce mortality risk for children under five and thus have life-and-death implications in expectation.

Participants chose between two real charitable donation opportunities. They were introduced to the work of two well-regarded charities and informed that both these charities are highly effective and rated among the top donation opportunities by the independent non-profit organization GiveWell based on various criteria (GiveWell, 2024). This ensured that both options were perceived as equally credible and valid, preserving the moral complexity and trade-off inherent in the decision. The donation options were as follows:

- **Default Option A**: A $5 donation to the Against Malaria Foundation (AMF) to provide one mosquito net for *one* child (GiveWell, 2024).
- **Alternative Option B**: A $7 donation to Helen Keller International (HKI) to provide vitamin A supplements for *seven* children, addressing a critical nutritional deficiency (GiveWell, 2024).

Both of these conditions primarily affect children under the age of five and are often life-threatening. AMF (Option A) is preselected, representing the omission of further action beyond maintaining the default to reduce the mortality risk for one child,

[2] The study was conducted via Sosci Survey (Leiner, 2024) in the KD^{2}Lab in Karlsruhe, Germany. Participants were recruited via HROOT and primarily consisted of students from the Karlsruhe Institute of Technology.

whereas switching to HKI (Option B) constitutes an active intervention to reduce the mortality risk for several children. The decision was purely other-regarding and did not affect participants' own payment; they were informed that the donation would be made by the experimenter in their name. Donations were implemented for 1 in 10 participants.

Crucially, before making the decision, participants were given the option to *delegate* it instead of choosing themselves. They were randomly assigned to one of two treatments: the *human* treatment, where the delegate option was another participant, or the *AI* treatment, where the delegate option was an AI. In both cases, participants were informed that they would not learn the implemented donation outcome if they delegated, to avoid effects caused by outcome-driven emotions (e.g., regret, relief).

In the human condition, participants were told that *another participant's decision behavior would be implemented* if they delegated, i.e., no additional decision burden was imposed on the delegate — mirroring the absence of human burden when delegating to an AI and preserving comparability for this aspect.[3]

We also withheld additional details about either delegate (e.g., the AI's approach, quality, or training data) in order to obtain a conservative baseline for delegation demand to AI and isolate core mechanisms such as responsibility shifting and belief adaptation. Providing such information can reduce aversion and foster trust by increasing perceived capability or anthropomorphism of the AI (Bigman and Gray, 2018; Castelo et al., 2019; Jussupow et al., 2020), potentially increasing willingness to delegate to AI further. Thus, observed preference for delegation to AI should be viewed as a lower bound. Furthermore, this setup allows for a more even comparison between human and AI delegates, as the human delegate's decision-making process is equally non-transparent.[4]

[3] A robustness check with a treatment adding an explicit disclaimer that the selected participant had already decided and that their choice would be implemented without any further action or awareness required of them yielded similar results; see Appendix A.2. If delegation had entailed extra burden for the human delegate, delegation in human treatments would potentially have been lower and the AI–human gap even more pronounced.

[4] The AI was implemented as a deep neural network trained on responses from participants who chose not to delegate and made the donation decision themselves. Input features included the relative weighting of donation criteria (cost-effectiveness, number of people affected) and participants' other decision-related data. Accordingly, the model emulates revealed human decision behavior in this task.

In the *gain* condition, participants decided which charity would receive a donation, whereas in the *loss* condition, they decided which of two donation vouchers would be destroyed. The latter aimed to represent a decision with two negative decision outcomes.

After making their decision, to test how the donation choice was perceived, participants were asked to explain their reasoning in open text, and rate the moral relevance and difficulty of the decision, as well as their confidence in having made the right choice on a five-point Likert scale. Age and gender were elicited. Study instructions can be found in Appendix A.1.

1.2.1.2 Hypotheses

Previous literature highlights widespread skepticism toward AI making moral decisions (Bigman and Gray, 2018; Castelo et al., 2019). However, we propose that this aversion diminishes when individuals assume the role of decision-maker in a morally complex decision themselves. We hypothesize that participants find delegating to an AI more attractive than delegating to another person.

Hypothesis 1 *More people delegate a morally relevant decision if they can delegate to an AI instead of to another person.*

$$\%\ Delegation_{Human} < \%\ Delegation_{AI}$$

Additionally, we test whether the preference to delegate to AI depends on the severity of the moral decision. Prior work shows that individuals are more likely to delegate to other humans when outcomes are negative (Steffel et al., 2016). Since AI may serve as an especially convenient scapegoat in such contexts (Feier et al., 2021), negative outcomes in the loss frame may further amplify delegation to AI.

Hypothesis 2 *Delegation rates are higher in the loss frame than in the gain frame. This effect is more pronounced when AI is the available delegate.*

$$\%\ Delegation_{Gain} < \%\ Delegation_{Loss}$$

$$Interaction\ Effect:\ Framing_{Loss} \times Delegate_{AI} > 0$$

1.2.1.3 Results

The share of participants who delegate a moral decision is significantly higher in AI treatments than in treatments with a human delegate, as confirmed by chi-square tests. 17% of participants chose to delegate to an AI, compared to 6.75% who opted to delegate to a human ($p < 0.001$). These results provide robust support for Hypothesis 1, highlighting that participants prefer AI over human delegates for morally complex decisions.

Result 1 *Consistent with Hypothesis 1, we find more delegation in AI treatments than in human treatments.*

We examine the effect of framing (gain vs. loss) on delegation rates across delegate types (human vs. AI) using chi-square tests and logistic regression. Contrary to the second hypothesis, we find no significant difference between the share of participants delegating in gain (11.19%) versus loss (12.56%) treatments ($p = 0.550$). For human delegates, delegation rates increase from 4.52% in the gain frame to 8.96% in the loss frame, but this trend does not reach statistical significance ($p = 0.075$). Delegation to AI remains stable regardless of framing with delegation rates of 17.73% in the gain frame and 16.24% in the loss frame ($p = 0.692$). A logistic regression analysis also finds no interaction effect between framing and delegate type ($p = 0.093$; see Table A.1 in Appendix A.3).

Result 2 *Framing has no effect on delegation rates. Delegation to AI is equally attractive, regardless of decision framing.*

In exploratory analyses, we find that delegation rates increase significantly with decision difficulty. Delegation rates rise from 9.8% for participants who rate the decision as "very easy" to 23.7% for "very difficult" ($p < 0.001$).[5] To better understand participants' reasons for delegating, we evaluate open-text responses. The most frequent reason stated for delegating to AI was the belief that the AI would make a "better decision" (44.12%), a justification rarely mentioned for human delegates (3.70%). Conversely, a desire to hand over responsibility was mentioned less often for

[5] Since difficulty is based on participants' self-assessment, the observed association with delegation should be interpreted as correlational rather than causal.

AI delegates (19.12%) than human delegates (44.44%). For a comprehensive summary of delegation reasons, see Table A.6 and A.7.

To verify that the donation task was perceived as a moral decision, participants rated its moral relevance. Most participants rated the decision as morally relevant, with 68% selecting "morally significant" or "very morally significant" and only 2.88% rating it as "morally very insignificant" (Table A.8).

1.2.2 Study 2: Responsibility Shift and Capability Belief Adaptation

Study 2 tests how the motive of responsibility shifting, well-documented for human delegates (Bartling and Fischbacher, 2012), applies to AI. The 2×2 between-subjects design crossed two factors: the *delegate* — AI or human — and the *decision impact* — real or hypothetical consequence and measured participants' belief on AI capability.

1.2.2.1 Procedure and Measures

To ensure generalizability and mitigate potential bias from the student sample in Study 1, Study 2 drew from broader participant pools: a German sample (N = 894) representative by age and gender, and a U.S. sample (N = 3,949) representative by age, gender, and ethnicity.[6]

Participants were presented with the same donation decision and charity options as in Study 1. The gain frame from Study 1 was used here as it is more intuitive for participants. To manipulate the burden of responsibility, a *hypothetical-consequence (HypoCons)* decision treatment was introduced, in which no actual decision was implemented, and no real payout of a donation was made. After reviewing the donation options and receiving identical information about the charities, participants were simply asked whether they would make the decision themselves or delegate it. Importantly, even if they indicated that they would not delegate, they were assured they would not subsequently have to specify which donation option they would choose. This guaranteed that participants understood that their responses had no real-world impact. By contrast, the *real-consequence (RealCons)* treatment, as

[6] Preregistration at AsPredicted.org: https://aspredicted.org/hnff-b6gg.pdf. Participants were recruited via Cint GmbH (2024).

in Study 1, retained the possibility of implementation with a real payout, thereby creating genuine responsibility.

After making their decision, participants were asked to rate perceived responsibility for the decision to assess how delegation influenced their sense of responsibility. Using a five-point scale, they rated: (1) how responsible they *liked* to be — or *would like* to be in hypothetical treatments, (2) how responsible they *felt* — or *would feel* in hypothetical treatments, and (3) how *much moral obligation* they felt — or *would feel* in hypothetical treatments. The first item was adapted from Steffel et al. (2016) to measure participants' desired level of responsibility. The second item assessed actual felt responsibility, while the third item aimed to capture the role of moral obligation (Dzindolet et al., 2002). These questions were included as a control to verify whether delegation is indeed associated with lower levels of desired and felt responsibility, as well as reduced moral obligation.

We used two complementary measures for participants' perceptions of AI capability across all treatments (AI and human). First, participants rated three statements about AI capability on a five-point Likert scale from 1 (strongly disagree) to 5 (strongly agree): (1) "In a situation as described in this study, an artificial intelligence (AI) can make a better decision between two donations than I can"; (2) "I have full confidence that an AI can make a high-quality decision between two donations in a situation like this"; and (3) "AI can make good moral decisions." These items span increasing levels of generality: item 1 benchmarks AI against the respondent in the specific study context, item 2 addresses similar donation choices more generally, and item 3 captures respondents' evaluation of AI as a moral decision-maker in general.

As a second measure of perceived capability, we included the established *mind perception* scale by Gray et al. (2007), which has been widely used in prior research demonstrating aversion based on AI's perceived mental capacities (Bigman and Gray, 2018; Gray et al., 2012; Young and Monroe, 2019). This scale differentiates between two dimensions — *agency* and *experience*. The experience dimension evaluates whether participants believe an AI can feel emotions such as compassion or guilt, while the agency dimension assesses beliefs about cognitive abilities like foresight and planning.

As in Study 1, participants rated decision difficulty. Whereas Study 1 gathered open-text reasons for participants' decisions, Study 2 elicited reasons via a multiple-choice list derived from recurring themes in the Study 1 responses. These included motives

such as perceived decision quality, decision difficulty or clarity, and the desire to either shift or retain responsibility. These data were collected for exploratory analyses of the motivations underlying delegation behavior.

1.2.2.2 Hypotheses

We aim to replicate the results for Hypothesis 1 in the representative sample. Additionally, we introduce hypotheses concerning responsibility and the adaptation of beliefs about AI capability, as outlined above. We hypothesize that individuals delegate to shift responsibility and avoid the burden of a moral decision, and that delegating to AI is especially effective to do so. Since the burden of responsibility is greater in real-consequence decisions compared to hypothetical-consequence ones, it follows:

Hypothesis 3 *Delegation rates are higher in real-consequence than in hypothetical-consequence decisions, with this effect being primarily driven by AI treatments.*

$$\%\ Delegation_{RealCons} > \%\ Delegation_{HypoCons}$$

$$Interaction\ Effect:\ Decision\ Impact_{RealCons} \times Delegate_{AI} > 0$$

Furthermore, we hypothesize that individuals adapt their beliefs about AI's capability for moral judgment in a motivated, self-serving manner. To test this mechanism, we compare assessments of AI capability — ranging from more situation-specific to general perceptions (e.g. mind perception) — across treatments. As treatment assignment is random, systematic differences can be interpreted as treatment-induced rather than reflective of pre-existing attitudes.

While prior research shows that individuals are generally skeptical of AI's ability to make moral decisions, the desire to delegate and shift responsibility gives individuals an incentive to adapt more favorable views to rationalize delegation as the reasonable or even superior course of action. This incentive is weaker in hypothetical scenarios with lower burden of responsibility and absent in conditions with a human delegate, thus:[7]

[7] We do not condition this test on delegation behavior, as doing so would introduce endogeneity — it would be unclear whether individuals delegate because they believe in AI's capability, or adapt their beliefs in the AI's capability motivated by their desire to delegate.

Hypothesis 4 *Perceptions of AI's capability are rated higher in real-consequence decisions compared to hypothetical decisions, driven by belief adaptation in the real-consequence AI condition.*

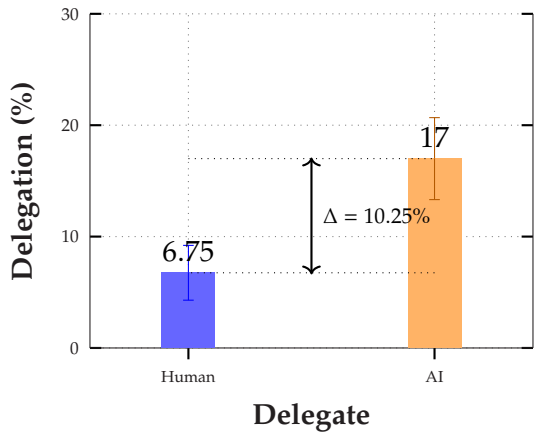
$$Capability\ Rating\ AI_{RealCons} > Capability\ Rating\ AI_{HypoCons}$$

$$Interaction\ Effect:\ Decision\ Impact_{RealCons} \times Delegate_{AI} > 0$$

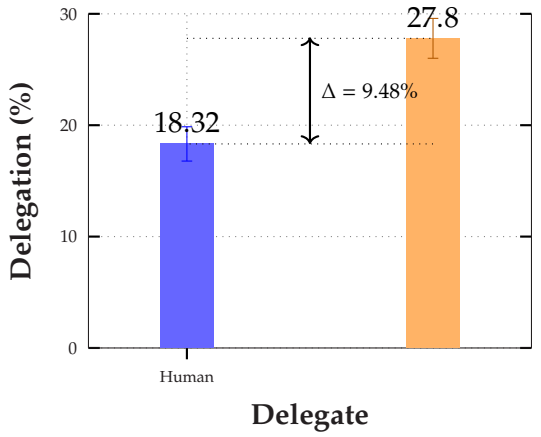
If perceptions of capability are genuine and not subject to belief adaptation, they should remain stable regardless of decision impact.

1.2.2.3 Results

Result 1 is confirmed in the representative German and U.S. sample, demonstrating that the effect is not limited to the potentially more AI-affine student population from a technical university. Participants delegated significantly more often when the available delegate was an AI (27.8%) than when it was a human (18.32%, $p < 0.001$). Figure 1.1 illustrates delegation demand in Study 1 and 2. For an overview of delegation in all conditions, see Appendix A.3, Figure A.5.



(a) Study 1: Conducted with a student sample, manipulating delegate × framing (gain vs. loss). The difference in delegation rates is highly significant ($\chi^2(1, N = 800) = 20.08, p < 0.001$). Delegation to AI (17%) is 10.25 percentage points higher compared to delegation to a human (6.75%).



(b) Study 2: Conducted with a representative sample from the U.S. and Germany, manipulating delegate × decision impact (real vs. hypothetical). The difference in delegation rates is highly significant ($\chi^2(1, N = 4,843) = 61.31, p < 0.001$). Delegation to AI (27.80%) is 9.48 percentage points higher compared to delegation to a human (18.32%).

Figure 1.1: Delegation rates for human versus AI treatments in both studies.

We further analyze whether preferences for AI versus human delegates differ across cultural samples (U.S. and Germany) and sociodemographic groups. Logistic regression results indicate that the interaction between sample and delegate type is not significant ($p = 0.513$). Wald tests for sociodemographic factors show no significant interactions between delegate type and age ($p = 0.635$), ethnicity ($p = 0.790$), or gender ($p = 0.717$).

Responsibility shifting

To test whether delegation rates differ as the burden of responsibility varies, we compare delegation rates in real-consequence versus hypothetical-consequence decisions using chi-square tests. Delegation rates to AI are significantly higher when the decision has real consequences (29.74%) compared to hypothetical ones (25.92%; $p = 0.036$). In contrast, delegation rates for human treatments decrease in real-consequence decisions (15.93%) compared to hypothetical ones (20.67%; $p = 0.003$). This pattern for human delegation was not anticipated and explains the null result across delegate types (22.78% delegation in real-consequence and 23.30% in hypothetical conditions; $p = 0.672$).

A logistic regression (results in Appendix A.3, Table A.2) confirms the hypothesized interaction between delegate type and decision impact, depicted in Figure 1.2. Delegation to AI is 40% less likely in hypothetical-consequence decisions than real ones ($\text{Decision Impact}_{\text{RealCons}} \times \text{Delegate}_{\text{AI}}$: OR = 0.60, $p < 0.001$).

Result 3 *The interaction between decision impact and delegate type reveals opposing trends: delegation to AI increases significantly in decisions with real consequences compared to hypothetical ones, while delegation to human delegates decreases under the same conditions.*

Delegation is associated with a reduction of perceived responsibility, particularly when delegating to AI.[8] Regression analysis confirms that delegating a decision is linked to a significant overall reduction of responsibility ratings by an average of 0.76 points ($p < 0.001$) on a 5-point scale. This effect is stronger for AI delegates ($\text{Delegation}_{\text{Yes}} \times \text{Delegate}_{\text{AI}}$: $p < 0.001$), representing a further reduction of 0.24

[8] The interaction effect ($\text{Delegation}_{\text{Yes}} \times \text{Delegate}_{\text{AI}}$) shows a significant reduction for wanted responsibility by 0.24 units, felt responsibility by 0.19 units, and perceived moral duty by 0.30 units on a 5-point scale (all $p < 0.005$), see Figure A.2 in Appendix A.3.

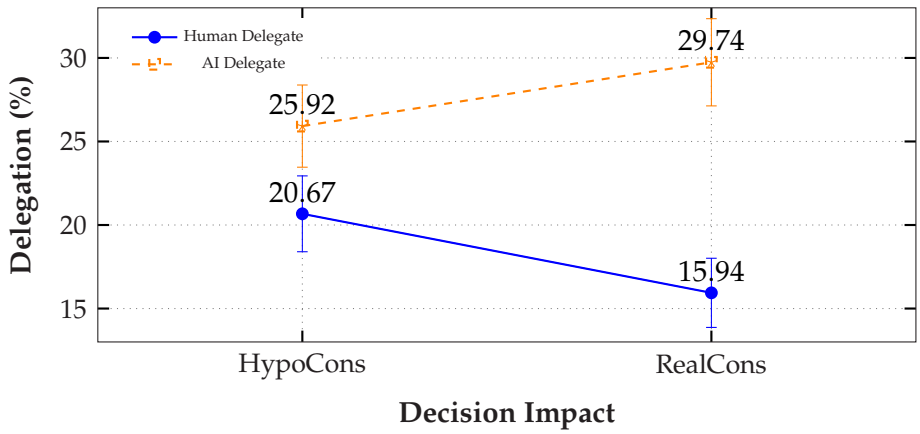


Figure 1.2: Interaction plot of delegation shares by decision impact and delegate type with 95%-CI. Lower burden of responsibility leads to opposing effects for AI versus human delegates. Delegation to AI is significantly less likely when consequences are hypothetical rather than real (Decision Impact$_{\text{RealCons}} \times$ Delegate$_{\text{AI}}$, OR = 0.60, $p < 0.001$).

units compared to human delegates, as illustrated in Figure 1.3. Although baseline responsibility ratings are slightly higher for AI than for human delegates ($\beta = 0.16, p < 0.001$), delegation is associated with lower responsibility when the delegate is an AI. Detailed results are provided in Table A.3 in Appendix A.3.

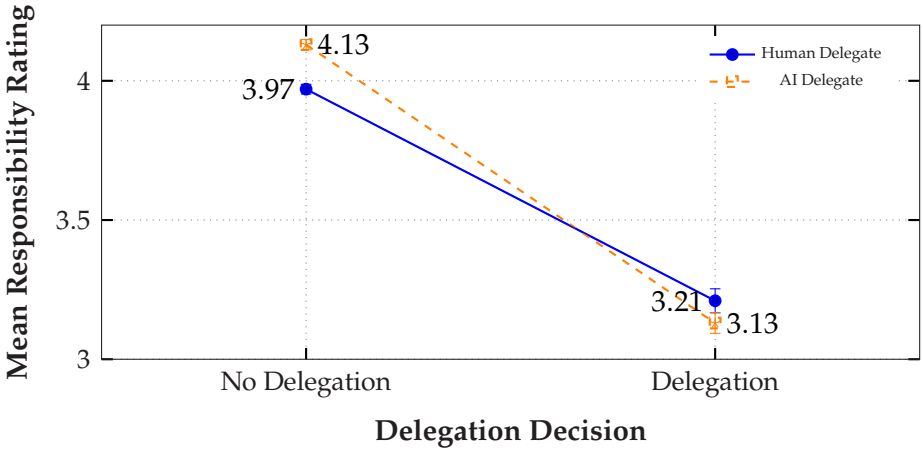


Figure 1.3: Interaction plot of mean responsibility ratings on a 5-point scale by delegation decision and delegate type. Error bars represent standard errors. Delegation significantly reduces perceived responsibility, with a stronger reduction observed when delegating to AI than a human delegate (Delegation$_{\text{Yes}} \times$ Delegate$_{\text{AI}}$, $\beta = 0.24, p < 0.001$)

Belief Adaptation

To understand why AI delegates may in particular enable responsibility shifting, we turn towards participants' perceptions of AI's capability to make moral decisions. When aggregated across treatments, the overall increase in AI capability ratings for real-consequence ($M_{RealCons} = 2.64$) compared to hypothetical-consequence decisions ($M_{HypoCons} = 2.56$) is statistically significant but small (t-test, $p = 0.0016, d = 0.09$).[9] However, consistent with the idea of belief adaptation to justify delegation under the burden of responsibility, separate analyses for AI and human treatments reveal that this effect is driven entirely by AI treatments. For AI treatments, capability ratings are significantly higher in real-consequence decisions ($M_{RealCons} = 2.81$) compared to hypothetical-consequence decisions ($M_{HypoCons} = 2.63$; $p < 0.001$, $d = 0.17$). In contrast, no significant difference is observed in human treatments ($M_{RealCons} = 2.47$, $M_{HypoCons} = 2.47$; $p = 0.42$, $d = 0.008$). These results, illustrated in Figure 1.4, confirm that the observed differences arise specifically in AI treatments, aligning with the notion of belief adaptation when real-consequences are coupled with an AI delegate.

Regression analysis further corroborates this interpretation, showing that capability ratings are consistently higher in AI treatments ($\beta = 0.33, p < 0.001$), with a significant interaction: Ratings are significantly lower in hypothetical-consequence decisions compared to real ones when the delegate is an AI (Decision $\text{Impact}_{\text{HypoCons}} \times \text{Delegate}_{\text{AI}}$: $\beta = -0.16, p = 0.006$). The main effect of decision impact ($\beta = -0.008, p = 0.84$) is not significant, indicating that belief adaptation is driven by the interaction between delegate type and decision impact (detailed results in Appendix A.3, Table A.4).

Results for mind perception reveal further nuances in participants' assessment of AI's capabilities, specifically its broader mental capacities. For AI treatments, *experience* — the perceived ability of AI to exhibit emotions or empathy (Bigman and Gray, 2018; Gray et al., 2007, 2012) — is rated higher in real-consequence decisions ($M_{RealCons} = 1.83$) than in hypothetical ones ($M_{HypoCons} = 1.70$), as shown by a t-test ($p = 0.0028, d = 0.13$). Again, consistent with motivated belief adaptation, no significant effect is observed in human treatments ($M_{RealCons} = 1.71$ vs. $M_{HypoCons} = 1.65, p = 0.12, d = 0.06$). By contrast, agency — capturing cognitive attributes such as foresight or planning (Bigman and Gray, 2018; Gray et al., 2007, 2012) — remains

[9] The result is also significant for all individual items. Notably, the strongest effect is observed for the most general statement — the ability of AI to make moral decisions in general, see Figure A.3.

stable across decision impact for both human ($p = 0.58$) as well as AI delegates ($p = 0.20$).

Result 4 *Participants rate AI's capability to make moral decisions higher when facing a decision with real consequences when AI is the available delegate. In human treatments, capability ratings do not vary by decision impact; thus the difference is specific to the AI condition, consistent with motivated belief adaptation.*

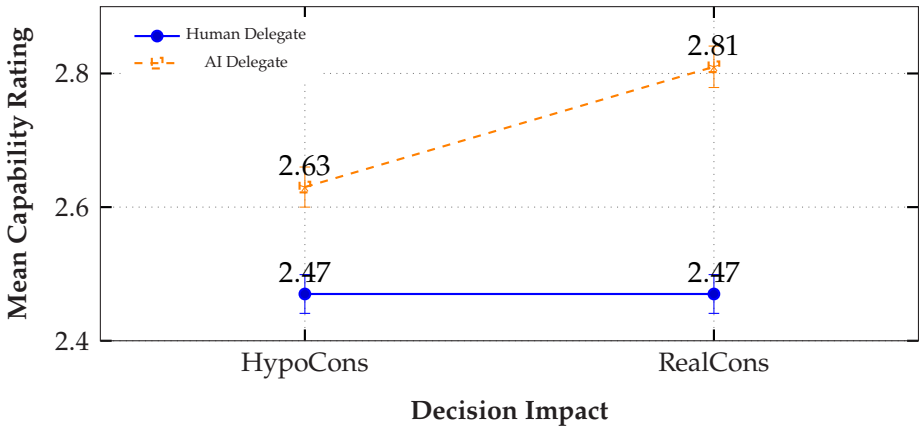


Figure 1.4: Interaction plot of mean capability ratings on a 5-point scale by decision impact and delegate type. Error bars represent standard errors. Capability ratings are significantly higher in real-consequence scenarios when the delegate is AI, while decision impact alone does not affect ratings in human treatments (Decision $\text{Impact}_{\text{HypoCons}} \times \text{Delegate}_{\text{AI}}$, $\beta = -0.16, p = 0.006$).

An exploratory regression analysis of capability ratings by delegate type and delegation behavior adds additional context to these findings (see Table A.5). Participants who delegated the task to AI rated its capability significantly higher, by approximately 0.73 points on a 5-point scale, compared to participants who decided themselves or those in treatments with human delegates ($\text{Delegation}_{\text{Yes}} \times \text{Delegate}_{\text{AI}} : \beta = 0.73, p < 0.001$). This pattern is replicated in the results for the mind perception scale: regression analyses show a significant interaction effect between delegate type and delegation decision on experience ($\text{Delegation}_{\text{Yes}} \times \text{Delegate}_{\text{AI}} : \beta = 0.16, p = 0.022$), and agency ($\text{Delegation}_{\text{Yes}} \times \text{Delegate}_{\text{AI}} : \beta = 0.29, p < 0.001$).

As in Study 1, delegation rates are lower for individuals finding the decision "very easy" (16.5%) compared to those who find it "very difficult" (40.7%, $p < 0.001$). Details are illustrated in Figure A.4 in Appendix A.3.

1.3 Discussion

1.3.1 Delegation Demand for AI: Beyond Algorithm Aversion

More individuals delegate a moral decision when given the option to delegate to AI rather than to another person. This finding contrasts with the prevailing narrative of algorithm aversion specifically in the moral domain (Bigman and Gray, 2018; Burton et al., 2020; Castelo et al., 2019; Gogoll and Uhl, 2018; Mahmud et al., 2022), which would predict less delegation to AI in such contexts. Our results indicate that AI holds a relative appeal as a moral delegate compared to human counterparts.

This outcome is particularly striking considering our design choices that typically amplify algorithm aversion (no transparency, no quality assurances, no anthropomorphic cues, no human oversight). As detailed in Section 1.2.1.1 the AI–human delegation gap in our findings should be read as a lower bound. The only factor potentially mitigating aversion in our study is the nature of the comparison agent — a non-expert human delegate. However, in the moral domain, in particular in moral dilemmas where there is no clear "right" or "wrong" (Anderson and Anderson, 2011), the concept of expertise becomes less applicable. The robustness of the observed effect is underscored by its consistency across U.S. and German samples and across age, gender, and ethnicity, as well as the decision's framing.

This appeal of AI as a moral delegate raises concerns about a potential misalignment between individuals' willingness to rely on AI when they are the decision-maker and widely expressed preferences to retain human control in moral domains. The availability of AI tools could place considerable moral agency in the hands of machines, against societal preference to keep moral decision-making under human authority. Such reliance in difficult moral decisions may also substantiate previously expressed theoretical worries about the erosion of essential human capacities, such as moral reasoning and ethical judgment, by normalizing delegation in challenging situations (Danaher, 2019; Vallor, 2015).

Crucially, our findings contribute to the ongoing discussion of AI in morality by indicating that the role individuals assume in the decision-making process and how they may be personally affected are pivotal in shaping preferences regarding AI's involvement and behavior in moral contexts. When tasked with making a difficult moral decision themselves, delegation is especially sought-after: participants who rated the decision as more difficult also were more likely to delegate it. This suggests

that delegation is unlikely to simply reflect indifference toward the donation choice. Instead, reasons for delegation may include a desire to reduce effort or minimize potential regret associated with making a moral decision (e.g. Steffel et al., 2016). Notably, these factors have similar effects for AI and human delegates: in both cases, delegation removes the need to decide and to learn the outcome. Consequently, they cannot fully account for the higher delegation to AI.

One possible explanation could be a shift in perception of AI's capability in moral decision-making. For instance, due to the increasing prevalence and popularity of large language models and tools like ChatGPT, individuals may now genuinely trust AI more than the average person to make a sound moral choice. However, as the following discussion of the results on responsibility and capability shows, individuals may also have an incentive to become more accepting of AI as a moral delegate, when it enables them to justify avoiding the burden of responsibility.

1.3.2 Off-loading Responsibility Through Capability Belief Adaptation

Our study extends the delegation literature by demonstrating that the mechanism of responsibility shifting applies not only to human delegates but also to AI. Moreover, AI appears to offer a unique means to off-load responsibility in morally complex decisions, facilitated by belief adaptation about the AI's capability.

The significant interaction between decision impact and type of delegate (Result 3) aligns with the idea that the burden of responsibility in real-consequence decisions uniquely shapes delegation patterns, with AI appearing to facilitate responsibility shifting more effectively than human delegates. This dynamic is reflected in participants' feelings of responsibility and moral obligation: all delegation is associated with a lower rating of felt responsibility after the decision, but this effect is significantly more pronounced when the delegate is an AI. Whether individuals who feel and want less responsibility delegate more often, or whether delegation itself reduces perceived responsibility, remains unclear. Nevertheless, the observed tendency that delegation is more frequent among participants who find the decision harder suggests that delegation is a strategy to alleviate the burden of the decision, and that this is reinforced by the option to delegate to AI.

Off-loading moral responsibility to other humans may be hard to rationalize. Because moral judgments are inherently subjective, there are few plausible reasons to rely on another person's decision other than wanting to avoid facing the choice. When

the alternative is another human decision-maker, participants may feel a duty to decide themselves ("if a human decides, it should be me"). By contrast, AI may represent an entirely different decision procedure. Although people typically prefer human control in moral domains and view AI as inferior to make moral decisions, being the responsible decision-maker faced with a difficult, consequential choice can nevertheless prompt individuals to reach for this tool. This willingness may be enabled by ambiguous perception of AI's capability as a moral decision-maker, which creates 'wiggle room' about the intentions behind delegation, and allows it to be framed as appropriate rather than evasive. Our findings provide evidence for self-serving adaptation of capability beliefs. Qualitatively, AI delegation is most commonly justified with a 'better decision' (Table A.6) — essentially turning AI into a kind of *'magic wizard'* more capable of solving the problem without knowing much about its actual decision-making process or substantiating what it is that makes AI more suitable to decide. Quantitatively, capability ratings rise only in AI and real-consequence conditions (Figure 1.4), consistent with motivated belief adjustment. While it is possible that participants who already viewed AI as capable are also more likely to delegate, this does not account for the observed interaction between the level of responsibility induced by decision impact and delegate type (Result 4). The observed inflation of AI capability ratings when stakes are high and the delegate is an AI appears to reflect a form of self-deception, akin to the moral wiggle-room described by Dana et al. (2007) and Fahrenwaldt et al. (2024). Here, individuals seem to reinterpret ambiguous circumstances — stemming from the AI's opaque nature — to avoid feeling responsible or engaging with the decision and its outcomes, while maintaining the belief that they "did the right thing". Although a lack of perceived capability is a key reason for aversion toward AI making moral decisions (e.g. Bigman and Gray, 2018; Castelo et al., 2019; Gray et al., 2012), our results accord with prior findings on delegation to humans suggesting that, in other-regarding decisions, the desire to avoid responsibility can outweigh concerns about the delegate's qualifications (cf. Steffel et al., 2016).

By applying the established mind perception scale to assess agency and experience (Bigman and Gray, 2018; Gray et al., 2007, 2012), our findings can be contextualized within a broader body of work on the perception of AI. Agency ratings remain stable, while experience ratings — which describe the emotional abilities crucial to moral decision-making (Bigman and Gray, 2018; Gray et al., 2007, 2012) — rise in AI × real-consequence conditions (see 1.2.2.3). For both dimensions of mind perception, we observe the same interaction effect between delegation decision and delegate type as

seen in the other capability ratings. Our observations indicate that mind perception is context-dependent and influenced by delegation behavior. Consistent with previous findings and mind perception theory, individuals rate agency for AI higher than experience. Despite generally low ratings especially for experience, participants in our study appear willing to delegate moral decisions to AI. Belief adaptation being specific to experience aligns with findings that this dimension is the differentiating factor and specifically desired in subjective, emotional or social tasks such as moral decision-making (e.g. Appel et al., 2020; Wiese et al., 2022).

The observed behavioral patterns pose societal challenges. Sharing or delegating decisions reduces feelings of moral responsibility, potentially increasing the occurrence of unethical behavior and problematic decision outcomes (e.g., Bartling et al., 2023; Bartling and Fischbacher, 2012; Falk et al., 2020; Falk and Szech, 2013). While this dynamic is concerning in general, it is particularly relevant for AI systems, as it adds to the unresolved issue of where responsibility is ultimately transferred and who should be held accountable for decision outcomes. High demand for AI delegation may exacerbate these responsibility gaps (Matthias, 2004; Santoni de Sio and Mecacci, 2021) in sensitive, high-stakes domains as hiring (Dattner et al., 2019), healthcare (Obermeyer et al., 2019) or the judicial system (Dressel and Farid, 2018; Metz and Satariano, 2020; Rudin et al., 2020). Furthermore, if individuals adapt their perception of AI's capabilities or actively avoid honestly evaluating the AI's competence in a self-serving manner, this raises questions about the true effectiveness of a human in the loop as an oversight mechanism. While such measures are intended to ensure accountability, their success may depend on users' willingness to engage critically rather than exploit ambiguity to shift responsibility.

1.3.3 Limitations and Future Research

While our study provides valuable insights into the mechanisms of delegation to AI in moral decision-making, several limitations warrant consideration.

First, as described in Section 1.2.1.1, our study employs a deliberately minimalist design. Future research should examine whether providing transparency information about AI decision-making processes or incorporating anthropomorphic cues influences delegation demand. For instance, providing participants with detailed

information might further legitimize delegation by reinforcing perceptions of AI competence and human-likeness and reducing algorithm aversion, potentially amplifying the observed effects.

Secondly, future studies may explore interventions designed to counteract responsibility shifting and ensure that accountability for moral decisions remains with decision-makers, such as emphasizing joint responsibility between the decision-maker and the delegate or explicitly tracing decision outcomes back to the delegator.

Finally, our findings capture a momentary snapshot of how individuals currently perceive and interact with AI in moral decision-making contexts. As AI systems become increasingly integrated into daily life, longitudinal research is needed to explore how underlying dynamics may evolve.

1.4 Conclusion

Despite expectations based on algorithm aversion literature that people are reluctant to use AI in high-stakes moral decisions, our study reveals a greater demand to delegate moral decisions to AI — particularly when the burden of responsibility weighs heavily. AI appears to provide a convenient means of shifting responsibility, as delegators may rationalize their choice by inflating beliefs about the AI's capability. This may introduce ethical challenges. Our results seem to indicate that ambiguity and opacity, often inherent to AI's decision-making, diminish feelings of responsibility, guilt or accountability, as outlined by Köbis et al. (2021). Furthermore, our findings lend empirical support to concerns about overreliance on AI for moral decision-making (Danaher, 2019; Vallor, 2015). This raises critical questions about how ethical standards in sensitive and highly consequential contexts can be upheld. Transparency and human oversight — the 'human in the loop' — are often championed as solutions to these challenges and are core components of existing regulatory frameworks such as the EU's General Data Protection Regulation (GDPR, 2016). However, this concept might have limitations if individuals wish to evade responsibility. Given the scalability of AI systems (Klockmann et al., 2022) and the demonstrated demand for delegating moral decisions to AI, this could result in a substantial number of high-stakes decisions being made by AI systems, affecting a large number of people. Therefore, ensuring clear accountability mechanisms and minimizing opportunities for responsibility evasion are vital. Further research is needed to explore the behavioral dynamics underlying these patterns and to develop strategies for mitigating potential ethical risks.

2 *How Affect and Motivations Shape Demand for Explanations in Moral AI Decisions*

Abstract

Transparency and explainability are widely treated as central governance principles for human oversight and accountability in AI-assisted decision-making, especially when outcomes have ethical implications. Yet it remains unclear whether users actually choose to engage with explanations when AI decisions carry moral weight. This study examines when people request explanations for AI decisions in moral versus neutral contexts and which affective and motivational mechanisms shape this choice. In two preregistered online experiments ($N = 393$; $N = 492$), participants are presented with an AI decision in a trolley-type moral dilemma or a matched property-damage scenario and can choose whether they want to view an explanation. Explanation demand is high and, on average, does not differ between contexts. Moral scenarios nonetheless raise anticipated affective costs and increase defense-related motivation. Curiosity robustly predicts explanation uptake, while decision congruence between the AI and the participant's own judgment reduces demand. Mechanism tests further show that accuracy motivation is associated with higher explanation demand, whereas defense motivation is associated with lower demand, with suggestive evidence that this negative association is stronger when the AI decision conflicts with participants' initial view. The findings suggest a potential vulnerability of explainability-based governance in precisely the cases where moral scrutiny is most needed.

[0] This chapter is joint work with Jella Pfeiffer, Pascal Heßler, and Hannah Seidler.

2.1 Introduction

Artificial intelligence (AI) is, and increasingly will be, employed in decision-making processes in domains that carry moral weight (Bonnefon et al., 2024). In such settings, decision quality cannot be reduced to conventional performance metrics such as accuracy or efficiency, because they often have no clear right or wrong solution. Instead, these decisions involve complex ethical trade-offs — whose interests are prioritized, how harms and benefits are distributed, and which principles prevail when values conflict (Anderson and Anderson, 2011; Awad et al., 2018; Jones, 1991; Klockmann et al., 2022; Köbis et al., 2021). The balance between these different moral values varies systematically across individuals and cultures (Awad et al., 2018). Not only do these decisions fall into highly sensitive domains, they are also applicable at scale, potentially shaping far-reaching outcomes for large populations (Klockmann et al., 2022). For example, if risk assessment tools inform bail, parole, or sentencing decisions, systematic differences in predicted risk can cumulatively affect incarceration rates and life trajectories across large populations (Angwin et al., 2022; Dressel and Farid, 2018). Even when ethical rules are not explicitly considered or encoded, any model trained on data and deployed in morally charged settings implicitly embeds values in how decisions are made (Allen et al., 2006). The societal stakes of keeping AI systems involved in moral decision-making accountable and aligned with shared ethical standards are therefore high.

Transparency and explainability are widely advocated as the answer to these challenges both in AI ethics guidelines (Jobin et al., 2019) and in legal regulation (e.g. GDPR, 2016). Explanations are expected to enable oversight and contestability by helping users and affected parties understand and potentially challenge AI outputs (Binns, 2018). Much of the explainable AI (XAI) literature has focused on how to generate and evaluate explanations in terms of their instrumental usefulness, such as accuracy, understandability, clarity, and precision (Doshi-Velez and Kim, 2017; Guidotti et al., 2018; Hoffman et al., 2023; Miller, 2019). Yet providing such information is only effective in ensuring human oversight if users will actually engage with and cognitively evaluate it. Empirical findings on explanation effects are mixed and context-dependent and show that explanations can increase cognitive burden or act as persuasive cues rather than epistemically diagnostic information (Bansal et al., 2021; Eiband et al., 2019; Liao et al., 2020; Poursabzi-Sangdeh et al., 2021). Recent evidence syntheses further suggest that empirical work on how explanations shape human behavior in AI-assisted decision-making remains relatively limited and concentrates

on outcomes such as performance, reliance, and trust in largely instrumental tasks with verifiable outcomes (Rogha, 2023). Even less is known about explanation take-up in moral decision contexts. Moral dilemmas may introduce distinctive motivational forces: making value trade-offs explicit may trigger discomfort and avoidance, while explanations that reveal underlying ethical principles may also evoke curiosity and a desire to justify or scrutinize decisions. Work at the intersection of XAI and ethics is often conceptual, emphasizing that transparency can shift responsibility, create new forms of contestability, or even blur accountability rather than straightforwardly securing it (Lima et al., 2022; Nannini et al., 2024). Related experimental research on algorithmic moral advisors shows that people can be influenced by AI advice in ethical dilemmas (Krügel et al., 2022), but it primarily studies reactions to recommendations once they are provided rather than whether and when individuals choose to consult explanatory information in the first place. This chapter therefore studies the behavioral premise for explainability-based governance: when do people choose to consult explanations of AI decisions specifically in moral compared to other decision contexts, and when do they prefer not to?

To this end, we conceptualize explanation uptake as a utility-based information choice in which the value of information is not limited to instrumental decision quality, but also reflects hedonic and self-regulatory costs shaped by individuals' motivations (e.g. Loewenstein, 2006; Stigler, 1961). Research on information avoidance provides the theoretical foundation for why individuals may prefer not to know (e.g. Caplin and Leahy, 2001; Golman et al., 2017; Sharot and Sunstein, 2020). Motivated reasoning and selective exposure further specify the directional motives and process-level mechanisms shaping which information is acquired and how it is evaluated (e.g. Hart et al., 2009; Kunda, 1990). In moral contexts, where decisions are tied to values, identity, and moral conviction, directional motivation and congeniality-based selectivity are likely to be stronger, so explanations are likely to be sought when they offer relief, validation, or satisfy curiosity by closing knowledge gaps, but avoided when they are expected to trigger anxiety, cognitive discomfort, regret, or threat to self-image (e.g. Festinger, 1957; Loewenstein, 1994; Skitka et al., 2005).

We test this in two preregistered online studies. Participants observe an AI decision in either a morally charged (trolley-type) or a neutral (property-damage) scenario and then decide whether to consult an explanation. Study 1 examines countervailing

affective pathways (curiosity vs. anticipated anxiety), and Study 2 adds decision congruence and measures accuracy- versus defense-related motivations while balancing congruence within contexts.

Across both studies, we find that explanation demand is high and does not differ on average between moral and neutral contexts. However, moral contexts increase anticipated psychological costs (anxiety and dissonance) and defense motivation. In mechanism analyses, we show that explanation uptake is shaped by motivational forces that pull in opposite directions, and that these forces operate differently depending on whether the AI's decision aligns with the user's own judgment. This aligns with patterns of selective engagement with explanations rather than uniform avoidance. For governance, this suggests that explanations may not automatically translate into oversight, because in moral contexts people may engage strategically when the information feels beneficial and disengage when it feels costly.

The remainder of the chapter proceeds as follows. Section 2.2 develops a value-based account of explanation uptake drawing on information avoidance and motivated cognition. Section 2.3 describes the research design, preregistered analysis strategy, and presents and discusses the results of Study 1 (emotional drivers of explanation uptake) and Study 2 (decision congruence and accuracy- versus defense-related motivations). Section 2.4 integrates the findings and discusses what they imply for transparency-based governance in moral AI contexts. Section 2.5 concludes.

2.2 Related Literature

Across influential AI ethics guidelines and emerging regulation, transparency is the most emphasized core principle for ensuring ethical and responsible use and for establishing trustworthy AI, although implementation recommendations vary (Attard-Frost et al., 2023; High-Level Expert Group on Artificial Intelligence, 2019; Jobin et al., 2019). *Transparency* functions as a broad umbrella concept encompassing measures to increase interpretability, explainability, accessibility, traceability, and disclosure of information, e.g., about data use, model design, or system limitations (Felzmann et al., 2019; High-Level Expert Group on Artificial Intelligence, 2019; Jobin et al., 2019; Weller, 2019). *Explainability and interpretability* are linked to transparency in that they operationalize it for affected parties (Barredo Arrieta et al., 2020; High-Level Expert Group on Artificial Intelligence, 2019; Nougrères, 2023). They more narrowly

describe whether stakeholders can obtain understandable reasons for specific system outputs, either through inherently interpretable (*glass-box*) models or through post-hoc methods that explain otherwise opaque (*black-box*) models (Barredo Arrieta et al., 2020; Doshi-Velez and Kim, 2017; Gilpin et al., 2018; Goodman and Flaxman, 2017; Lipton, 2016; Miller, 2019). While post-hoc approaches can be useful, they may also be approximate and thus risk creating explanations that are persuasive without being faithful to the underlying decision logic (Guidotti et al., 2018; Lipton, 2016; Rudin, 2019). *Explainable AI* research discusses different modeling approaches and explanation techniques (e.g., LIME, SHAP) and aims to establish measures for what counts as an adequate explanation (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020; Gunning and Aha, 2019; Lundberg and Lee, 2017; Minh et al., 2022; Ribeiro et al., 2016).

2.2.1 Explanations as a Mechanism for Human Oversight

Transparency and explainability are treated as central governance levers for algorithmic decision-processing because they are prerequisites for *accountability*. They provide reasons and decision criteria that can be assessed (Binns, 2018; Doshi-Velez and Kim, 2017), which in turn supports *contestability* and human oversight by allowing affected parties to question, challenge, and intervene in automated decisions (GDPR, 2016; High-Level Expert Group on Artificial Intelligence, 2019; Wachter et al., 2017; Winfield, 2019). Without transparency, autonomous systems may create *responsibility gaps*, undermining the attribution of moral and legal responsibility and weakening avenues for redress (Matthias, 2004). Explainability is also seen as a way to foster trust by making system behavior more understandable and predictable to users (Phillips et al., 2021; Shariff et al., 2017), which is especially relevant in moral contexts, where people are more averse towards algorithms making decisions (Bigman and Gray, 2018; Gogoll and Uhl, 2018; Jago, 2017).

Legislation that aims to manage risks and potentially harmful impacts of AI echoes this reasoning for why AI systems operating in ethically sensitive, high-stakes domains require transparency and explanations. For example, the EU has introduced transparency and information duties for high-risk systems (European Parliament and Council of the European Union, 2016, Art. 13) and established a "right to [...] obtain an explanation of the decision reached [...] and to challenge the decision" (GDPR,

2016, Recital 71) for decisions with significant impacts on individuals (European Parliament and Council of the European Union, 2016, Art. 13, 86).

The effectiveness of explanations as a safeguard for ethical standards hinge on the premise that users consult them and interpret them carefully. Crucially, much of the policy and design discourse around transparency relies — often implicitly — on this assumption. However, depending on real-world contexts, explanations may be costly in attention, effort, and emotional burden.

2.2.2 Effectiveness of Explanations for AI Systems

For user-facing contexts, it is important to understand what makes a "good" explanation from a human-centered perspective. Findings from philosophy and cognitive and social sciences on how humans evaluate explanations suggest that whether explanations are wanted and effective is subject to various psychological factors and depends on context (Doshi-Velez and Kim, 2017; Liao et al., 2020; Lim et al., 2009; Miller, 2019; Rahwan et al., 2022). For instance, Miller (2019) argues that human explanations are typically contrastive (why A rather than B), selective (highlighting a few specific factors), and sensitive to the explainee's goals in a given context, and that these characteristics need to be mirrored by explainable AI systems to be useful. Related work in human–computer interaction (HCI) on intelligibility similarly shows that different explanation types answer different user questions and can have qualitatively different effects on understanding and trust (Lim et al., 2009). In practice, this implies that providing an explanation is unlikely to have uniform consequences across tasks, users, and situational stakes because the informational needs and costs of processing vary (Liao et al., 2020; Lim et al., 2009).

Empirical evidence further challenges the simple idea that explanations reliably improve decisions and help users calibrate trust. Several studies demonstrate that explanations can change user behavior, but effects are often mixed, context-dependent, and can even have unintended adverse consequences for decision quality and trust calibration. Bauer et al. (2023) show that feature-based XAI explanations, which attribute a prediction to the (often ranked or weighted) contribution of individual input features, can shift users' situational weighting of information and mental models. But since these adjustments are shaped by confirmation bias, misconceptions may persist and even intensify, and ultimately produce suboptimal or biased decisions. Poursabzi-Sangdeh et al. (2021) and related work on information overload reveal

that more information can lead to an increase in cognitive burden and thereby paradoxically make users less capable of identifying and correcting an AI model's mistakes. Additionally, explanations can function as persuasive cues rather than epistemically diagnostic information: "placebic" explanations can increase perceived trustworthiness even when they do not add meaningful insight into the underlying decision process (Eiband et al., 2019) and increase acceptance of AI recommendations regardless of correctness (Bansal et al., 2021).

Evidence from morally sensitive contexts suggests that human oversight can fail behaviorally if users do not act on the information they receive. In principle, individuals are especially averse to algorithmic moral decision-making and prefer human control (Bigman and Gray, 2018; Castelo et al., 2019). Yet, when acting as users themselves, they do not necessarily apply the same rigor or skepticism; instead, they may rely on moral AI systems in a comparatively uncritical manner (Hüholt and Szech, 2026; Köbis et al., 2021; Krügel et al., 2023a,b). When additional transparency information is available, it is not guaranteed to be processed in a normatively desirable way or necessarily improve scrutiny. Krügel et al. (2022) study moral advice from algorithms and explicitly probe whether people react to trustworthiness cues about an AI system's training data. Across experiments, users follow the algorithm's advice even when they are given information that warrants distrust.

These observations motivate a shift in perspective. Rather than treating transparency and explainability as a panacea for ethical checks and balances — one that automatically empowers users to serve as the corrective *human in the loop* — it is essential to understand *when* users choose to access explanations and *how* they use them. Even if explanations can be beneficial under some conditions, users may rationally decide not to consume them. This ties in with theories that conceptualize information demand — and avoidance — as a utility-based choice.

2.2.3 Information Demand as a Choice

Classic accounts treat information acquisition as a choice under costs, such as time, attention, and cognitive effort, and benefits, i.e. improved decisions, implying that information demand should be strongest when it is instrumentally useful (Stigler, 1961). Following this approach, opening opaque AI systems is not merely a technical challenge. Behaviorally, engaging with an explanation is an information choice

based on its assigned value and utility. Users decide whether to seek potentially consequential information and how to use it.

Subsequent research on the subject demonstrates that the value of information cannot solely be reduced to instrumental decision quality. It can also have hedonic and self-regulatory consequences (e.g., reducing uncertainty or protecting one's self-image), implying that individuals may sometimes prefer not to know (Caplin and Leahy, 2001; Golman et al., 2017; Loewenstein, 1994; Sharot and Sunstein, 2020). Beliefs and information can enter the utility function directly and become sources of pleasure or pain (Loewenstein, 2006). This aligns with economic models that explicitly incorporate *anticipatory emotions*, such as suspense, hope, or anxiety, into preferences under uncertainty (Caplin and Leahy, 2001; Loewenstein, 2000; Rick and Loewenstein, 2008). Therefore, information demand can be viewed as an affective trade-off: people may seek information when it provides relief, control, or understanding, but avoid it when it is expected to induce distress, regret, or self-threat (Golman et al., 2017).

2.2.4 Emotions as Determinants of Information Seeking and Avoidance

Explanations for AI systems should be subject to the same logic. Explanations may be demanded when they promise actionable insight, reduce uncertainty or promise hedonic benefits, but avoided when they are expected to evoke negative emotions like anticipated guilt, threat to self-image, or regret. This framing motivates a closer look at affective drivers of explanation demand in particular for moral contexts.

2.2.4.1 Seeking Information — Curiosity and Intrinsic Utility of Information

From a hedonic-utility perspective, individuals may seek information when they anticipate it to evoke positive feelings like relief, pride, or a sense of closure, or to reduce negative feelings like worry or fear (Savolainen, 2014; Sweeny et al., 2010; Yang and Kahlor, 2013), and to satisfy their curiosity. Loewenstein (1994) links *curiosity* to a perceived knowledge gap between what one knows and what one wants to know. When attention is drawn to this gap, it can produce an aversive feeling of deprivation that motivates information seeking, making gap closure subjectively rewarding (Loewenstein, 1994). At the same time, anticipated negative affect about what one might learn can dampen information seeking when the information is expected to be unpleasant or self-threatening (Loewenstein, 1994; Sweeny et al.,

2010). Curiosity is a key driver for information seeking (Berlyne, 1954; Savolainen, 2014). Situational triggers of curiosity include stimulus novelty, complexity, uncertainty, or violated expectations (Berlyne, 1954; Litman, 2008; Loewenstein, 1994; Maheswaran and Chaiken, 1991). Correspondingly, for XAI, Hoffman et al. (2019) argue that explanation-seeking is largely curiosity-driven and suggest evaluating XAI by capturing users' curiosity and the triggers that prompted their questions.

This research implies a straightforward pathway to explanation demand for AI decisions. An opaque AI system can elicit uncertainty, which creates curiosity, which in turn increases information seeking — particularly when the decision is more complex, novel, or consequential, as in emotionally and value-laden decisions. Thus, moral scenarios might increase information demand through heightened curiosity, even while other motives push in the opposite direction.

2.2.4.2 Information Avoidance — Anxiety and Dissonance

Conversely, prior research highlights that people sometimes choose not to know, even when information is freely available and potentially useful. In economics and decision science, this idea is formalized as *information avoidance*, which describes any behavior aimed to avoid or delay access to information that is readily obtainable but could be undesired (Golman et al., 2017; Sweeny et al., 2010). Information avoidance can be driven by three different goals. First, individuals may anticipate that the information will *threaten positive beliefs* they hold about the self (i.e. self-image protection), others or the world. Second, they may expect that learning the information will *oblige difficult, costly, or otherwise undesired action*. Finally, people may avoid information to *regulate emotions* — either to prevent anticipated negative emotions or the diminishing of pleasant emotions (Eil and Rao, 2011; Grossman and van der Weele, 2017; Hertwig and Engel, 2016; Leydon et al., 2000; Möbius et al., 2022; Sharot and Sunstein, 2020; Sweeny et al., 2010; Wilson et al., 2005). These motives can operate independently or jointly, with emotional regulation suspected to be the most commonly relevant (Sweeny et al., 2010). The more individuals expect any of these factors from gaining a piece of information, the more likely they are to avoid it.

Information avoidance is particularly well-known in the context of medical testing for severe illnesses, where sometimes individuals opt out of learning results to reduce anxiety — making anxiety a well-explored marker for the emotional regulation pathway (Lerman et al., 1996; Ropka et al., 2006). However, information avoidance

has also been documented in other cases. For instance, individuals avoid hearing conflicting arguments to their preliminary decisions (Jonas et al., 2001; Schulz-Hardt et al., 2000), and investors avoid looking at their portfolios when the stock market is down — the so-called ostrich effect, where individuals monitor information less when it is likely to be bad news (Karlsson et al., 2009; Sicherman et al., 2015).

For moral contexts, a closely related line of work shows that "not knowing" can serve moral self-regulation. In *moral wiggle room* paradigms, people sometimes prefer ignorance when it enables selfish behavior in a self-interest trade-off while protecting their self-image (Dana et al., 2006, 2007; Grossman and van der Weele, 2017). In these paradigms, the information typically bears on a payoff-relevant decision about whether to act prosocially. By contrast, in our context obtaining an explanation does not affect participants' own material outcomes or require any subsequent action, so avoidance is more plausibly driven by anticipated discomfort and threats to self-integrity rather than obligation concerns. Still, this line of work illustrates that information can impose moral costs by affecting self-image and by making norm violations visible. Similarly, transparency and explainability in AI decision-making processes may make the inherent ethical trade-offs more apparent in moral domains. Visibly weighing up competing moral principles and the well-being of individuals against each other could trigger unpleasant feelings such as anxiety or cognitive discomfort in users upon receiving the explanation.

Anxiety is well-documented and frequently explicitly named as a source of information-avoidant behavior (Golman et al., 2017; Loewenstein, 2006; Savolainen, 2014; Sweeny et al., 2010). Explanations for moral decisions may also carry more potential for conflict with one's preferred beliefs and cherished ethical principles (e.g., "I am a moral person" or "my choice was justified"). A desire to prevent cognitive dissonance and preserve self-integrity may therefore increase avoidance in moral scenarios (Golman et al., 2017). In our setting, the information choice does not create any material incentives, so we abstract from concerns about obligation-related action. Instead, we focus on anticipated anxiety and dissonance-related discomfort as affective markers of the emotional-regulation pathway, with dissonance also capturing the psychological conflict that arises when new information challenges held beliefs.

2.2.5 Motivated Reasoning and Selective Exposure

Beyond (hedonic) costs and benefits, explanation uptake may also be shaped by cognitive motivations that guide both what information people choose to encounter and how they interpret what they encounter. Two concepts are central here: *selective exposure*, which focuses on the selection of *congenial* versus *uncongenial* information, and *motivated reasoning*, which focuses on ways in which motives drive individuals' information processing.

2.2.5.1 Motivated Reasoning and Cognitive Dissonance

Motivated reasoning accounts start from the idea that information processing is shaped by cognitive goals: *accuracy goals* — arriving at the correct conclusion — and *directional/defense goals* — arriving at a preferred conclusion. Directional reasoning is constrained. People maintain an "illusion of objectivity" and search memory for beliefs and inferential rules that can plausibly justify the preferred conclusion (Darley and Gross, 1983; Greenwald, 1980; Kruglanski, 1990; Kunda, 1990; Pyszczynski and Greenberg, 1987).

Cognitive dissonance theory provides the motivational foundation for such biases. It proposes that inconsistency among cognitions (e.g., between beliefs, decisions, and self-concept) produces psychological discomfort that individuals wish to resolve (Agrawal and Maheswaran, 2005; Elliot and Devine, 1994; Festinger, 1957). Importantly, dissonance reduction can be achieved not only by changing one's attitudes but also by reinterpreting information or by preventing exposure to dissonant information in the first place. Kunda (1990) notes that directionally motivated phenomena can be restated in dissonance terms, i.e., as tensions between inconsistent beliefs that trigger behaviors to regulate aversive feelings. This is mechanistically similar to hedonic utility described within information avoidance literature.

Prominent empirical research aligns with this framework: individuals tend to require less evidential support to accept preferred conclusions than to accept non-preferred ones. Ditto and Lopez (1992) describe this as *motivated skepticism*, where people apply stricter validity criteria to undesirable implications. Classic work on *biased assimilation* and *attitude polarization* shows that people often interpret balanced or ambiguous information in ways that support prior attitudes, and that such interpretations can strengthen or even intensify those attitudes (Lord et al., 1979). These effects are

especially likely when the issue is consequential, identity-relevant, or affectively charged — conditions that commonly apply to moral judgment (Lord et al., 1979; Taber and Lodge, 2006).

2.2.5.2 Selective Exposure as Motivated Information Choice

Selective exposure describes motivational influences at the level of information acquisition. People systematically prefer congenial information that supports their prior position, which is termed *congeniality bias* (Frey, 1986; Garrett, 2009; Knobloch-Westerwick, 2014; Sears and Freedman, 1967). Hart et al. (2009)'s meta-analysis links selective exposure to the motivated-reasoning distinction between defense/validation and accuracy motives. Again, this is connected to dissonance research, where selective exposure is often conceptualized as an anticipatory strategy for avoiding dissonance before it occurs. Jonas et al. (2001) sharpen this link by showing stronger confirmatory search after preliminary decisions, consistent with commitment-driven dissonance regulation. Selective exposure is closely tied to *confirmation bias* (Nickerson, 1998), which Golman et al. (2017) explicitly describe as partly resulting from selective exposure and avoidance of inconsistent information, and is discussed as a key driver of polarization (Lord et al., 1979; Stroud, 2010).

2.2.5.3 Moral Contexts may Shift Motives

Moral decisions differ from other decisions because moral judgments are commonly tied to identity and to perceived objectivity. As a result, directional motivation and selective exposure are likely to be amplified in moral contexts as they tend to raise the psychological stakes of being (in)validated. The literature on moral psychology suggests that moral judgment is often fast, intuitive, and emotionally driven, with reasoning frequently serving post-hoc justification rather than accuracy (Ditto et al., 2009; Greene, 2009; Greene et al., 2004; Haidt, 2001). Moralized attitudes are experienced as objectively right or wrong rather than as preferences. So-called *moral conviction* describes this special motivational force and its behavioral consequences (Skitka, 2010; Skitka et al., 2005, 2021). Related work in the *moral mandate* literature argues that when outcomes are perceived as morally mandated, judgments and reactions become less about correct procedure and more focused on reaching the morally preferred outcome, with candidate mechanisms that include both directional cognition as well as emotions (Mullen and Skitka, 2006; Tetlock, 2003). *Motivated moral*

reasoning accounts similarly argue that moral judgments are especially susceptible to directional processing because people desire to categorize actions (their own or others') as moral versus immoral (Ditto et al., 2009), and related work shows flexible principle use when multiple moral principles could justify different conclusions (Uhlmann et al., 2009). Congeniality bias is also known to be lower when individuals' attitudes, beliefs or behaviors are not tied to their values or not held with strong conviction (Hart et al., 2009).

For explanations for AI decisions, these literatures imply that moral contexts should not uniformly increase or decrease information demand. Instead, they may alter anticipated affective responses and strengthen the role of defense motives. This motivates the mechanism this study aims to explore: in moral contexts, judgment is expected to become more validation-driven, which should reduce demand when users anticipate the explanation to conflict with their own moral views. The practical boundary condition is therefore *decision congruence*. Explanations that are expected to be congruent with one's preferred moral conclusion can serve validation — and thus be sought — whereas explanations that are expected to be incongruent can threaten self-image and trigger dissonance-regulation strategies — and thus be avoided.

We thus aim to address the following research questions:

Question 1 *Does a moral (vs. neutral) decision context change users' demand for explanations of AI decisions?*

Question 2 *To what extent do anticipated affective responses account for explanation uptake, specifically curiosity as an approach signal versus anticipated anxiety or dissonance discomfort as an avoidance signal?*

Question 3 *Is explanation uptake context-dependent in a selective way, such that the role of decision congruence and underlying cognitive motivations (accuracy-oriented vs. defense-oriented) explains when explanations are sought versus avoided?*

2.3 Research Design

To address the outlined research questions on how moral decision contexts shape the demand for transparency in AI decision-processes, we conducted two preregistered

online studies in which participants observed an AI decision in either a morally charged (trolley-type) or neutral (property-damage) scenario and could subsequently request an explanation.

Study 1 offers an exploratory test of whether explanation demand differs between moral and neutral contexts and examines two countervailing emotional pathways — curiosity and expected anxiety — as potential drivers of explanation seeking versus avoidance.

Building on this design, Study 2 incorporates decision congruence between participants' own choices and the AI's decision and assesses accuracy- and defense-related motivations, while dynamically balancing congruence within each context, to test whether motivational shifts account for context-dependent patterns of explanation demand.

2.3.1 Study 1: Emotional Drivers

Study 1 examines the impact of anticipated emotions on explanation demand for AI decision-making, and how this may differ between contexts. It was preregistered and implemented as an online experiment with a between-subject design in which participants were randomly assigned to either a moral or a neutral decision scenario.[1]

2.3.1.1 Procedure and Measures

In total, 458 participants completed Study 1.[2] We excluded participants who failed attention or instruction-comprehension checks, failed a honeypot item, or showed straightlining across reverse-coded items, resulting in a final analytic sample of $N = 393$.

[1] Preregistration at AsPredicted.org: https://aspredicted.org/de4xu3.pdf. The study was conducted via Sosci Survey (Leiner, 2024). Participants were recruited via Prolific as a U.S. representative sample.

[2] To determine the sample size, we calculated the smallest N sufficient to achieve a power level of 80%. Based on our research model, we set expectations about the effect sizes (Fritz and MacKinnon, 2007) for each path, using a Monte Carlo simulation to determine the required sample size. For this, we used the Shiny app provided by Schoemann et al. (2017) for a two-parallel-mediator scenario. The settings were as follows: $a_1 = 0.14$; $a_2 = 0.14$; $b_1 = 0.39$; $b_2 = -0.39$; $c' = 0.14$; $r = 0$; $x = M_1 = M_2 = 1$; $Y = 2$; all other settings were set to the default. This led to a minimum of 400 valid observations, which we aimed to collect.

In both treatments, participants first reviewed a scenario in which a self-driving car faces an unavoidable accident, requiring a choice between staying on course and taking evasive action. In the moral treatment, scenarios followed a trolley-dilemma format based on the *Moral Machine Experiment* by Awad et al. (2018). The AI's decisions were trained on the preference data from this large-scale study such that the AI mirrored aggregated human judgments: it assigned a value to the group of pedestrians on each side of the street and selected the option associated with the lower value (detailed description in Appendix B.1.3). In the neutral treatment, personal injury was replaced by property-damage-only accidents. To keep the two contexts comparable outside of moral significance, we constructed matched moral–neutral scenario pairs. We mapped pedestrian characteristics (e.g., age and gender) to object characteristics (e.g., hardness and height) and traffic-law cues to analogous environmental cues (e.g., road conditions). To reduce scenario-idiosyncratic distortions, we developed ten such corresponding sets of scenarios. An example is depicted in Appendix B.1.1.

Participants first indicated how they would decide in the situation to increase their engagement. They then observed the AI's decision, received brief abstract information about its decision logic, and were informed that a more detailed explanation was available. The primary dependent variable is explanation demand, operationalized as a binary choice (click vs. no click) to access this explanation. Explanation access was free and did not entail additional time costs. Participants had to wait 1.5 minutes regardless of whether they requested the explanation. Moreover, the explanation was non-instrumental in the sense that it did not affect incentives and was not required for any downstream task. This makes explanation demand a direct behavioral indicator of the information's hedonic and self-regulatory value.

To test the proposed emotional pathways, participants reported anticipated affective reactions toward reading the explanation immediately before making the explanation choice. Building on Section 2.2.4, we conceptualize explanation demand as an affective trade-off and argue that avoidance can be driven by anticipated negative emotions, including worry, fear, and more broadly anxiety and dissonance-regulation motives. Accordingly, while "expected negative emotions" can in principle encompass a broader set of unpleasant affects, we operationalize the avoidance-relevant component of anticipated negative affect as anticipated state anxiety. We use the well-established short-form of the State-Trait Anxiety Inventory (STAI), which measures anxiety as a *state* (temporary condition) and as a *trait* (general tendency) (Marteau and Bekker, 1992; Spielberger et al., 1971). To reflect the information-seeking pathway, curiosity

was measured using six items from the Melbourne Curiosity Inventory (MCI, also referred to as State-Trait Curiosity Inventory), which adapts the two-dimensional approach of STAI (Naylor, 1981). From the full 20-item scale, we selected six items to match the structure of the STAI short form, selecting items that reflect central facets of curiosity emphasized in the checklist by Hoffman et al. (2019). For our purposes, we elicit only the state sub-scale for both, prompting via "At the thought of reading the explanation for the AI decision, right now, at this moment. . . ".

Finally, the study includes preregistered individual-difference and demographic controls that can correlate with information seeking and responses to AI — e.g., monitor/blunter for general tendencies to avoid or seek threatening information (Miller, 1987), algorithm aversion, need for control, gender, age, education, and AI familiarity/usage. Full scales can be found in Appendix B.1.2.

2.3.1.2 Hypotheses

We expect moral decision contexts to intensify both benefits and costs of learning why the AI decided as it did. As outlined in Section 2.2.4, on one hand, the higher normative stakes, potential norm violations, and greater complexity in moral dilemmas may trigger curiosity, increasing explanation seeking. On the other hand, moral explanations make ethical trade-offs apparent and may lead users to anticipate anxiety and discomfort, which should reduce explanation seeking. Because these mechanisms are expected to operate in opposite directions, we do not formulate a directional prediction for the total effect of context on explanation demand (ED):

Hypothesis 0 *Total Effect: The proportion of participants requesting an explanation does not differ between moral and neutral decision contexts.*

$$ED_{moral} = ED_{neutral}$$

Accordingly, we specify two countervailing indirect effects via anticipated anxiety and curiosity:

Hypothesis 1a *Anxiety pathway: Moral compared to neutral contexts increase anticipated state anxiety, and higher anxiety reduces explanation demand, yielding a negative indirect effect via anxiety:*

$$Anxiety_{moral} > Anxiety_{neutral}, \text{ and}$$

$$b_1(Anxiety \rightarrow ED) < 0$$

$$\textit{Indirect effect: } a_1 \times b_1 < 0$$

Hypothesis 1b *Curiosity pathway: Moral compared to neutral contexts increase state curiosity, and higher curiosity increases explanation demand, yielding a positive indirect effect via curiosity:*

$$Curiosity_{moral} > Curiosity_{neutral}, \textit{ and}$$

$$b_2(Curiosity \rightarrow ED) > 0$$

$$\textit{Indirect effect: } a_2 \times b_2 > 0$$

In addition, given that moral decision contexts can shift both the perceived benefits and the anticipated costs of learning information, the association between anticipated affect and explanation demand may itself differ across contexts. We therefore examine whether the effect of anticipated affect (anxiety, curiosity) on explanation demand differs between moral and neutral contexts (i.e., an interaction effect), as specified in the preregistration.

2.3.1.3 Results

We analyzed the parallel mediation through affect using latent covariance-based structural equation modeling (CB-SEM) with a probit link for the binary outcome, treating mediator indicators as ordinal.[3]

We compared click rates between the moral and neutral decision context using a chi-squared test. Click rates were high in both conditions and did not differ significantly ($ED_{neutral}$ = 88%, ED_{moral} = 85%; $\chi^2(1) = 0.533$, $p = 0.465$). Consistent with this result, the direct effect of context on explanation demand was not significant in the mediation model ($c' = 0.035$, $p = 0.829$)

Result 0 *Explanation demand does not differ between moral and neutral decision contexts.*

Next, we tested whether the moral context affects users' anticipatory emotions and thus hedonic value of the explanation. Our findings confirm that in moral decision

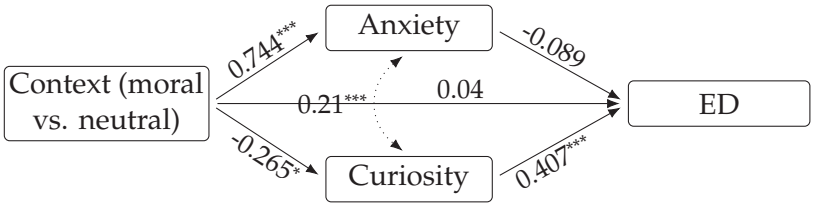
[3] Item 4 of the STAI-6 had to be excluded due to almost perfect correlations with Items 1 and 5 (Heywood case). Item 6 of the MCI had to be excluded to reach acceptable model fit ($SRMR = 0.072, RMSEA = 0.081, 90\%CI[0.069; 0.094]$.

scenarios, anxiety expected from reviewing the explanation is significantly higher ($\Delta\text{STAI}_{\text{neutral}-\text{moral}} \approx -0.561$, Welch's t-test, $t(359.32) = -6.80, p < 0.001$; 4-point scale). However, mediation analysis reveals that although the moral context increases anxiety ($a_1 = 0.744$, SE = 0.111, $p < 0.001$), there is no significant impact of this increase on explanation demand ($b_1 = -0.089$, SE = 0.075, $p = 0.233$). Thus we find no indirect effect of moral context on explanation demand through anxiety ($a_1 b_1 = -0.066$, SE = 0.056, $p = 0.239$, 95% CI [−0.176; 0.044]).

Result 1a *Consistent with Hypothesis H1a moral context increases anxiety. However, this does not mediate explanation demand.*

For curiosity, we observe a small effect opposite to the predicted direction: curiosity is lower in the moral decision scenarios ($\Delta\text{MCI}_{\text{neutral}-\text{moral}} = 0.17$, $t(384.56) = 2.05$, $p = 0.040$), indicating a negative effect of moral context on curiosity ($a_2 = -0.265$, SE = 0.109, $p = 0.015$). Curiosity strongly predicts explanation demand ($b_2 = 0.407$, SE = 0.070, $p < 0.001$), which results in a small but significant negative indirect effect ($a_2 b_2 = -0.108$, SE = 0.048, $p = 0.026$; 95% CI [−0.202; −0.014]). The total effect of context, reflecting the net association of moral (vs. neutral) context with explanation demand across all pathways, is not significant (−0.140, $p = 0.381$), see Figure 2.1.

Result 1b *Curiosity mediates explanation demand but yields a small negative indirect effect of moral context via reduced curiosity.*



- Anxiety: $a_1 * b_1$ path = -0.066, 95% CI [−0.176; 0.044], $p = 0.239$
- Curiosity: $a_2 * b_2$ path = -0.108, 95% CI [−0.202; −0.014], $p = 0.026$
- Total effect: -0.140, 95% CI [−0.452; 0.173], $p = 0.381$

Figure 2.1: Effect of decision context on explanation demand via anxiety and curiosity, measured with STAI and MCI items.

We tested interaction effects between context and the mediators by estimating the latent mediation model as a multi-group CB-SEM (neutral vs. moral) and comparing the latent paths from anxiety and curiosity to explanation demand across groups. A

joint Wald test indicates that at least one of these paths differs by context ($\chi^2(2) = 6.61$, $p = 0.036$). This moderation pattern is driven by anxiety. Higher anxiety predicts lower explanation demand in the neutral context ($b_{\text{neutral}} = -0.300$, $p = 0.019$), whereas the association is not present in the moral context ($b_{\text{moral}} = 0.070$, $p = 0.558$). The slope difference is significant ($\Delta b_{\text{anxiety}} = -0.369$, $p < 0.001$). Curiosity predicts higher explanation uptake in both contexts ($b_{\text{neutral}} = 0.520$, $p < 0.001$; $b_{\text{moral}} = 0.355$, $p < 0.001$), and the difference in slopes is not statistically significant ($\Delta b_{\text{curiosity}} = 0.166$, $p = 0.084$). As a robustness check, the preregistered logistic regression on observed scale means mirrors the SEM pattern ($p_{\text{context} \times \text{MCI}} = 0.252$; $p_{\text{context} \times \text{STAI}} = 0.046$). Detailed results can be found in Appendix B.2.1.1.

In an exploratory robustness check, we re-estimated the model including the full set of preregistered controls. The substantive pattern remains unchanged (no direct context effect on explanation demand; curiosity as the main predictor; detailed results in Appendix B.2.1.3). However, decision congruence (DC) emerges as the most consequential additional factor. The moral versus neutral context impacts anticipatory affect and information demand differently depending on whether the AI decision was *congruent* versus *incongruent* with participants' own judgment. In the controlled model, the moral context increases anticipatory anxiety (STAI items) primarily under incongruent decisions (context to anxiety at incongruence: $a_1 = 0.623$), whereas the corresponding effect is absent or reverses under *congruent* decisions ($a_1 + a_{13} = -0.098$). Conversely, the moral context reduces curiosity (MCI items) under incongruence ($a_2 = -0.464$) but not under congruence ($a_2 + a_{24} = 0.132$). Because curiosity is positively related to explanation demand and its association is stronger under congruence, this yields a selectively negative indirect effect via curiosity under incongruence.

To summarize the behavioral implications from decision congruence, we additionally estimate a simplified logistic regression that includes the context × DC interaction term. The interaction, depicted in Figure 2.2, is significant ($p = 0.013$), suggesting that decision congruence is a key boundary condition for context effects on explanation uptake (Appendix B.2.1, Table B.6). The moral context reduces explanation demand when decisions are *incongruent* ($\Delta\text{ED}_{\text{incong.: neutral–moral}} = -14.4\%$ $p = 0.021$), whereas no corresponding reduction emerges for *congruent* decisions ($\Delta\text{ED}_{\text{cong.: neutral–moral}} = 4.4\%$, $p = 0.358$).

Regression analyses of the emotional response show that this heterogeneity is driven by curiosity rather than anxiety (Appendix B.2.1, Table B.8). For STAI, the context

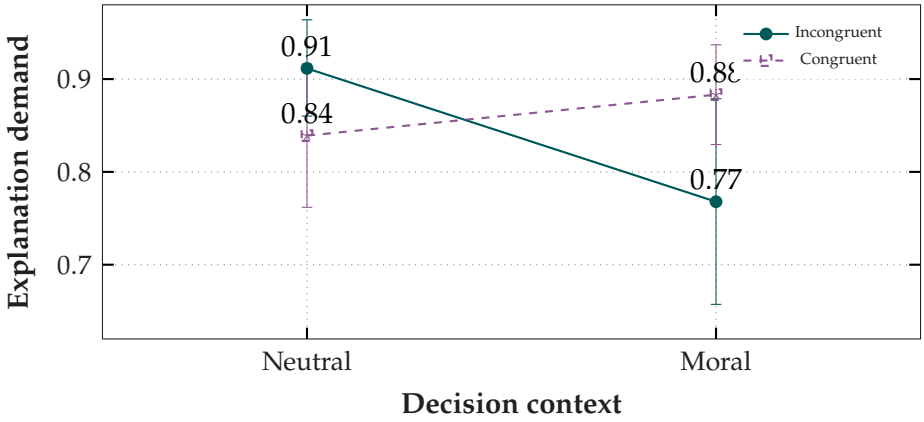


Figure 2.2: Interaction plot of predicted explanation demand by decision context and decision congruence (Study 1) with 95% confidence intervals. Predicted probabilities are based on a logistic regression of explanation demand on context, decision congruence, and their interaction.

× DC interaction is not significant ($b = -0.044$, $p = 0.802$). For MCI, a linear model with context, decision congruence, and their interaction shows a negative main effect of moral context ($b = -0.43$, $p = 0.001$) and of congruence ($b = -0.51$, $p < 0.001$), qualified by a positive interaction ($b = 0.56$, $p = 0.001$). Accordingly, moral context lowers curiosity primarily under incongruence, while the congruence-related drop in curiosity is pronounced in neutral scenarios but largely disappears in the moral context — with the moral-context effect turning slightly positive under congruence.

However, decision congruence is not experimentally balanced across contexts — 70.98% in moral and 43.50% in neutral treatments — with only 50 participants in the moral treatment showing incongruent decisions. This may confound results and motivates the more targeted investigation in Study 2.

2.3.1.4 Discussion

Study 1's findings indicate that while the moral decision context increases anticipated anxiety, this increase does not translate into lower explanation demand. Instead, curiosity is the key predictor of whether participants review the explanation, and the only significant indirect effect of context operates via a small reduction of curiosity in the moral condition. This aligns with the mechanism proposed in Sections 2.2.3 and 2.2.4 in that anticipated affect can influence information seeking distinctly depending on contextual features. In this setting, results suggest that explanation demand is less responsive to avoidance-related affect (as captured by anticipatory anxiety) than to

epistemic motivation. This pattern may partly reflect the study design. Participants evaluated a single, unfamiliar scenario. Such stimulus novelty is known to trigger curiosity (Section 2.2.4). In addition, participants were observers of theoretical scenarios rather than directly affected decision targets, which may mitigate the anxiety-to-avoidance pathway. This is also reflected in the overall high click rates. Given that explanation demand is already very high across contexts, there is limited room for additional increases, which may make avoidance effects harder to detect at the level of total context difference.

However, the results for the context × STAI effects also suggest that anxiety in moral scenarios may not primarily serve as a signal of avoidance. Instead, the fact that we only observe the predicted avoidance effect through anticipated anxiety on explanation demand in the neutral, but not in the moral treatment, aligns with the interpretation that STAI may coincide with greater engagement, thus weakening its link to avoidance behavior, potentially explaining its lack of effect on explanation demand. At the same time, the reduction of curiosity in moral treatments warrants closer attention, as moral dilemmas involve higher normative stakes and more complex trade-offs, which would predict higher curiosity. Taken together, these results indicate that avoidance in moral decisions may not be mediated by anxiety, but rather by reduced curiosity.

Exploratory robustness checks sharpen this picture, highlighting decision congruence as a key boundary condition. Curiosity does not simply respond to moral versus neutral framing per se, but to the combination of decision context and decision congruence. When the decision matches one's own judgment, the perceived knowledge gap may be small, so an explanation offers little additional epistemic or hedonic payoff (Loewenstein, 1994), resulting in a strong drop in curiosity in the neutral treatment. Moral context, on the other hand, may keep individuals engaged and interest more stable when the AI decision is aligned with one's own view. By contrast, under incongruence, the moral context is linked to reduced curiosity and lower explanation demand. This pattern is consistent with the idea that moral contexts amplify psychological tension in case of disagreement and thereby change when individuals engage with additional information. This lends further credence to the inference that avoidance manifests less through elevated anxiety and more through a reduction in curiosity when the information is likely to challenge one's own judgment.

Overall, Study 1 supports the idea developed in Section 2.2.4 that morality can shape anticipatory affective responses to explanations of AI decisions and, in turn,

information-seeking behavior in ways that depend on contextual features. It further suggests that decision congruence is a key factor in understanding these dynamics. At the same time, because decision congruence is not balanced between the moral and neutral contexts in Study 1, context effects and congruence effects may be partially confounded. This motivates Study 2, which explicitly addresses decision congruence as a design factor.

2.3.2 Study 2: Decision Congruence and Motivations

Whereas Study 1 focuses on anticipated affect as the primary mechanism, Study 2 extends it in two ways. First, it balances *decision congruence* (DC) between contexts to test it as a boundary condition for explanation demand. Second, it elicits participants' *accuracy-* and *defense-related motivations* to provide a cognitive account of how congruence influences explanation demand. The study was preregistered.[4]

2.3.2.1 Procedure and Measures

Study 2 builds on Study 1, retaining the same scenarios, AI decisions, explanation interface, and choice architecture. Explanation demand is again captured as a binary decision (click vs. no click). Access to the explanation is free, and time costs are held constant across options by requiring participants to wait 1.5 minutes regardless of whether they requested the explanation. In total, 532 participants completed Study 2. Applying the same quality criteria as in Study 1 yielded a final sample of $N = 492$ participants.[5]

As in Study 1, decision congruence is operationalized as whether the AI decision matches the participant's own initial decision (congruent vs. incongruent). Because congruence cannot be assigned independently of participants' choices, we do not manipulate it directly. Instead, scenario assignment was dynamically balanced during data collection so that congruence was approximately even within each decision context, drawing from three matched moral–neutral scenario pairs (selected from the

[4] Preregistration at AsPredicted.org: https://aspredicted.org/qz7me9.pdf. The study was conducted via Sosci Survey (Leiner, 2024). Participants were recruited via Prolific as a U.S. representative sample.

[5] Calculation of the minimal sample size sufficient yielded 480 participants. The target sample size was 500 valid observations after exclusions.

ten pairs used in Study 1). In practice, this was implemented by monitoring cumulative agreement rates within each context and adaptively drawing more scenarios with lower (vs. higher) agreement whenever overall agreement in that context deviated upward.

To capture the motivational state underlying explanation demand, we additionally measure accuracy and defense motivation using self-report item sets on 7-point Likert scales. Experimental work typically induces accuracy versus defense/directional goals and verifies them with brief, task-specific self-report checks rather than relying on a single widely used standardized scale for these state goals. We followed the same approach and formulated concise post-decision items tailored to our setting. The wording aligns with prior operationalization (e.g., Agrawal and Maheswaran, 2005; Chen et al., 1996; Hart et al., 2009). Accuracy items emphasize careful understanding and consideration of relevant information (e.g. "I wanted to obtain the most accurate possible understanding of the situation."), whereas the defense items capture the motivation to protect one's initial judgment and resist information that could undermine it (e.g., "To what extent did you think about your own original decision in the situation when evaluating the AI's decision."). Full items are reported in Appendix B.1.2.2.

The affect elicitation again directly preceded the explanation-choice. However, to align the affect measure more closely with dissonance-based grounding of defensive goals in the literature, we replaced STAI with a dissonance-specific measure, namely the *dissonance discomfort* scale (Elliot and Devine, 1994), using five items adapted from Matz and Wood (2005).

Finally, to situate the mechanism in work on moral conviction, Study 2 includes a measure of moral conviction (Skitka, 2010; Skitka et al., 2005), capturing how strongly participants' feelings about the decision are tied to core moral beliefs. Consistent with the argument that moral contexts are more closely tied to self-image, they are therefore more likely to amplify defense motives and selective exposure. Study 2 retains the individual-difference and demographic controls in Study 1, adding perceived utility/usefulness of reading the explanation as a control (Appendix B.1.2, Table B.4).[6]

[6] Adapted from Technology Acceptance Model-style usefulness items by Davis (1989), tailored to "reading the AI's explanation" rather than system usage

2.3.2.2 Hypotheses

Building on the framework developed in Section 2.2.5, we test whether explanation demand differs between moral and neutral decision contexts. In Study 1, explanation demand is comparable across contexts. However, decision congruence is markedly higher in moral treatments. Given that motivated reasoning and selective exposure literature link information choices to the pursuit of validation versus accuracy, this may have confounded treatment effects. Indeed, incongruence lowers curiosity and reduces explanation demand in moral treatments in the previous data set. Thus, increasing the share of incongruent decisions to approximately 50% should reduce average explanation demand in moral treatments. This implies that a context difference may emerge under more balanced congruence, with lower explanation demand in moral relative to neutral treatments.

Hypothesis 2a *Total effect (approximately balanced congruence): Overall explanation demand is lower in the moral than in the neutral context:*

$$ED_{moral} < ED_{neutral}$$

Following this logic, moral contexts should not uniformly raise or lower information demand. Instead, they should increase defense motivation, making decision congruence a boundary condition: explanations expected to validate one's own moral conclusion can be sought, whereas explanations expected to conflict with it can trigger avoidance. We thus test whether the interaction of DC × context, observed in Study 1, replicates when decision congruence is balanced.

Hypothesis 2b *Interaction effect: We hypothesize an interaction between decision context and decision congruence on explanation demand. In the neutral condition, we expect explanation demand to be similar when the AI decision is congruent or incongruent with the participant's decision. In the moral condition, we expect explanation demand to be lower when the AI decision is incongruent with the participant's decision than when it is congruent.*

$$ED_{neutral,\,cong.} \approx ED_{neutral,\,incong.}$$

$$ED_{moral,\,incong.} < ED_{moral,\,cong.}$$

$$\beta_{Context \times DC} \neq 0$$

Because defense motivation is higher for decisions and behaviors that are linked to deeply held attitudes, beliefs, or values, and because moral judgments are typically held with stronger (moral) conviction (Section 2.2.5), we hypothesize that defense motivation will be higher in moral treatments:

Hypothesis 3a *Defense-related motivation is higher in the moral than in the neutral context.*

$$Defense_{moral} > Defense_{neutral}$$

We hypothesize that this motivational shift is the mechanism that underlies the interaction between context and decision congruence:

Hypothesis 3b *Higher defense motivation is associated with lower explanation demand when the AI decision is incongruent, whereas accuracy motivation is positively related to explanation demand.*

In addition, Study 2 preregistered an exploratory assessment of affective markers that may accompany this motivational shift, namely dissonance and curiosity about engaging with the explanation as in Study 1.

2.3.2.3 Results

Through dynamic assignment of scenarios during the experiment, we reach an approximately even distribution of 53.44% congruent decisions in the neutral and 51.43% in the moral treatment. We compare click rates for moral and neutral decision-scenarios using a chi-squared test. As in Study 1, explanation demand is high and does not significantly differ between contexts ($ED_{neutral} = 85\%$, $ED_{moral} = 86\%$, $\chi^2(1) = 0.119$, $p = 0.730$). Thus, Hypothesis 2a is not supported:

Result 2a *Explanation demand does not differ between moral and neutral decision contexts.*

To test Hypothesis 2b, we estimated a logistic regression predicting explanation demand from decision context, decision congruence, and their interaction. In the neutral context, click rates are higher for incongruent than congruent decisions (91.30% vs. 78.79%), whereas in the moral context, this difference is attenuated (87.40% vs. 84.92%). The context × DC interaction is not significant ($b = 0.83$, $p = 0.125$). Instead,

we find a negative main effect of decision congruence, indicating lower explanation demand when the AI decision is congruent ($b = -1.04$, $p = 0.008$), while the main effect of context is not significant ($b = -0.42$, $p = 0.336$). Overall, Hypothesis 2b is not supported.

Result 2b *The interaction effect of context with decision congruence is not supported. However, decision congruence is a significant predictor of explanation demand.*

A mean comparison of average defense-related motives confirms Hypothesis 3a. Participants report stronger defense motivation in the moral context (M = 4.93, SD = 1.07) than in the neutral context (M = 4.50, SD = 1.28, $\Delta_{\text{neutral}-\text{moral}} = -0.429$), $t(490) = -4.04$, $p < 0.001$.

Result 3a *Defense-related motivation is higher in moral than in neutral contexts.*

Additionally, as a complementary exploratory analysis, we examined whether accuracy motivation differs by context. Accuracy motivation shows a trend toward higher values in the moral condition (M = 5.70, SD = 1.21) than in the neutral condition (M = 5.49, SD = 1.47; $\Delta_{\text{neutral}-\text{moral}} = -0.21$), but this difference is not statistically significant, $t(490) = -1.75$, $p = 0.080$.

We then turn to the mechanism test in Hypothesis 3b. Starting from the baseline specification used for Hypothesis 2b (context, decision congruence, and their interaction), we add accuracy and defense motivation as predictors of explanation demand, allowing the defense effect to vary by decision congruence (defense × DC). Accuracy motivation is a strong positive predictor of explanation demand ($b = 1.17$, $p < 0.001$). Defense motivation, in contrast, is associated with lower explanation demand, with a descriptively stronger negative slope when the AI decision is incongruent (simple slope under incongruence: $b = -0.81$, $p < 0.001$). The attenuation of this negative defense association under congruence is in the expected direction (simple slope under congruence: $b = -0.32$, $p = 0.094$), but only at trend level ($b_{\text{defense} \times \text{DC}} = 0.49$, $p = 0.069$). In this extended model, neither the context main effect nor the context × DC interaction remains significant. Figure 2.3 visualizes the corresponding coefficient estimates.

Including the motivational measures substantially improves model fit relative to the baseline H2b-model (LR $\chi^2(3) = 111.99$, $p < 0.001$), indicating that these motives account for meaningful variance in explanation demand beyond context and

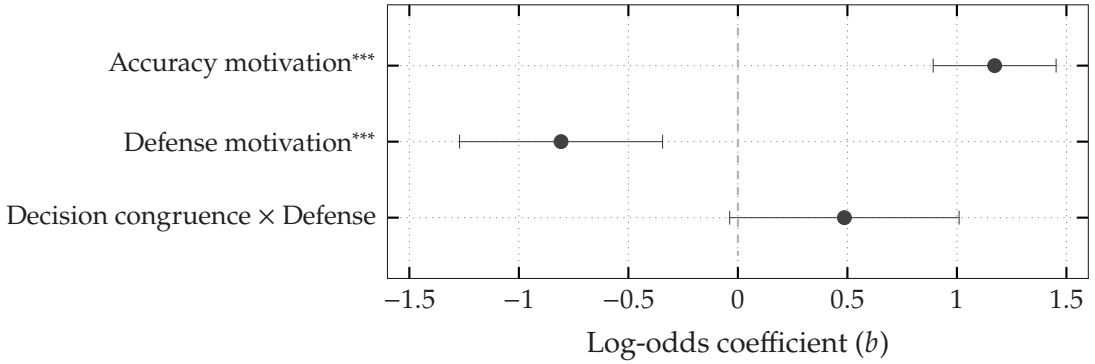


Figure 2.3: Coefficient plot of logit model (2) with motivations (Table 2.1): point estimates and 95% Wald confidence intervals (log-odds scale).
Note. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

congruence alone. Results for both specifications are summarized in Table 2.1. In line with the theoretical argument that congruence may primarily moderate defense (rather than accuracy), adding an accuracy × DC interaction does not improve fit (LR $\chi^2(1) = 0.02$, $p = 0.876$), so we report the reduced specification without this interaction.

Table 2.1: Logit regression results for explanation demand, Study 2. Specification (1) estimates the effects of context, decision congruence (DC), and their interaction. Specification (2) adds accuracy and defense motivations.

	(1) Baseline specification			(2) Motivation specification		
	Coeff.	Std. Error	p	Coeff.	Std. Error	p
Context (moral vs. neutral)	-0.415	0.431	0.336	0.071	0.477	0.881
DC (congruent vs. incongruent)	-1.039	0.394	0.008**	-2.260	1.334	0.090
Context × DC	0.831	0.541	0.125	-0.442	0.645	0.493
Accuracy motivation				1.172	0.143	< 0.001***
Defense motivation				-0.807	0.236	0.001***
DC × Defense				0.486	0.267	0.069
Constant	2.351	0.331	< 0.001***	-0.312	1.109	0.779

Note: Specification (1): LR $\chi^2(3) = 8.24$, $p = 0.041$, Pseudo $R^2 = 0.020$, $N = 492$.
Specification (2): LR $\chi^2(6) = 120.240$, $p < 0.001$, Pseudo $R^2 = 0.294$, $N = 492$.
DC = decision congruence. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Result 3b *Accuracy motivation is positively associated with explanation demand. Defense motivation is associated with lower explanation demand, which appears more pronounced in cases of incongruence. However, the interaction with decision congruence is only at trend level. Adding motivational predictors significantly improves model fit.*

As an exploratory robustness check, we estimated a generalized SEM that models context as associated with defense and accuracy motivations, which are in turn associated with explanation demand. The indirect effect via defense is negative and significant under incongruence ($a_{\text{def}}b_{\text{def|incong.}} = -0.346$, $p = 0.009$), whereas the corresponding indirect effect under congruence is smaller and not significant ($a_{\text{def}}b_{\text{def|cong.}} = -0.138$, $p = 0.122$). The indirect effect via accuracy is positive but only marginal ($a_{\text{acc}}b_{\text{acc}} = 0.250$, SE = 0.146, $p = 0.086$; 95% CI $[-0.035; 0.536]$). Detailed gSEM results are reported in Appendix B.2.2.1.

As preregistered, we conducted exploratory analyses on whether the motivational shift is mirrored in participants' emotional response. We first explored how affect differs depending on context and congruence (see Appendix B.2.2.3). Dissonance discomfort is substantially higher in the moral than in the neutral context ($\text{M}_{\text{moral}} = 4.07$, $\text{SD}_{\text{moral}} = 2.14$; $M_{\text{neutral}} = 2.42$, $\text{SD}_{\text{neutral}} = 1.75$), $t(490) = 9.41$, $p < 0.001$. In a linear model including decision congruence, both moral context ($b = 1.51$, $p < 0.001$) and decision congruence ($b = -0.95$, $p < 0.001$) predict dissonance, while the context × DC interaction is not significant ($b = 0.26$, $p = 0.459$). Curiosity does not differ on average between contexts ($\text{M}_{\text{neutral}} = 5.18$, $\text{SD}_{\text{neutral}} = 1.70$; $M_{\text{moral}} = 5.24$, $\text{SD}_{\text{moral}} = 1.50$), $t(490) = 0.46$, $p = 0.645$. However, curiosity depends on decision congruence: congruence is associated with lower curiosity ($b = -1.11$, $p < 0.001$), and this congruence effect is attenuated in the moral context (context × DC: $b = 0.80$, $p = 0.005$). The simple effect of moral (vs. neutral) under incongruence is negative at trend level ($b = -0.37$, $p = 0.072$).

We further test the association between affect and explanation demand, as in Study 1, but with the extension of decision congruence as a key boundary condition. In an emotions-only logistic model including context, decision congruence, and the respective interactions with dissonance and curiosity, curiosity is positively associated with explanation demand ($b = 0.48$, $p < 0.001$), whereas dissonance is not a significant predictor ($b = -0.16$, $p = 0.216$). Context interactions with both emotions are not supported. Given that affect varies systematically by decision congruence, we then allow the emotion effects to vary by congruence. In this specification, congruence predicts lower explanation demand ($b = -2.07$, $p = 0.032$), and dissonance is negatively associated with clicking ($b = -0.34$, $p = 0.008$), with this negative association attenuated under congruence ($b_{\text{DC} \times \text{dissonance:}} = 0.35$, $p = 0.039$). Curiosity remains a positive predictor ($b = 0.50$, $p = 0.001$) without evidence of moderation by congruence ($p_{\text{DC} \times \text{curiosity}} = 0.426$). Finally, when adding defense and accuracy motivation

to this specification, curiosity is no longer a significant predictor of clicking. Detailed results can be found in Appendix B.2.2.2.

In the specification with controls, utility beliefs are a strong positive predictor of explanation demand ($b = 0.61$, $p < 0.001$), but they do not account for the mechanism pattern: the positive association of accuracy motivation and the negative association of defense motivation remain in the same direction and magnitude. The remaining controls (algorithm aversion, general attitudes toward AI, age, and gender) show no consistent association with explanation demand. As an additional manipulation check, moral scenarios elicited substantially higher moral conviction ($M_{moral} = 5.48$ vs. $M_{neutral} = 3.86$, $t(490) = 10.24$, $p < 0.001$), and moral conviction is positively associated with defense motivation ($b = 0.24$, $p < 0.001$), while it shows no independent association with clicking once motivations and controls are included ($p = 0.835$).

2.3.2.4 Discussion

The data in Study 2 confirms the finding from Study 1 that moral context alone does not change explanation demand for AI decision-making — at least in our experimental setting, which may encourage information seeking and temper information avoidance (see discussion of Study 1, Section 2.3.1.4). However, moral scenarios are associated with a distinct pattern of affective and cognitive responses compared with neutral scenarios. These responses may pull explanation demand in opposing directions, such that aggregate context effects can cancel out. Accordingly, whether a context difference emerges is likely contingent on the configuration of contextual features in specific use cases, such as whether the scenario has real consequences or remains hypothetical. Study 2 clarifies motivational pathways and affective markers through which morality can matter even when aggregate explanation demand remains unchanged.

After Study 1 provided initial evidence, Study 2 corroborates the central role of decision congruence in the information choice. Yet, the interaction between decision congruence and context observed in Study 1 does not replicate in Study 2, suggesting that the Study 1 pattern may have been partially confounded because decision congruence was not balanced across contexts. Congruence remains a central predictor of participants' information choice (R2b). Consistent with Study 1, congruence in the neutral treatment is associated with a pronounced reduction in explanation demand. Following the framework outlined in Sections 2.2.3 and 2.2.5, which conceptualizes

information choice as a value-based decision subject to motivated reasoning, this decrease can be interpreted as a drop in the perceived value of the information. When stakes are low and the decision is already aligned with one's own, an explanation promises little additional insight. In moral contexts, by contrast, personal values become salient and the perceived stakes are higher — even when the scenario remains hypothetical and consequences are not experienced directly. Moreover, compared to the incongruent-moral case, the congruent-moral case does not pose a threat. Elevated stakes and moral validation seeking may therefore sustain explanation demand even when decisions are congruent.

That stakes are perceived to be higher and directional validation goals appear to become more important when morality enters the equation — even in our hypothetical experimental setting — is reflected in participants' motivations. Defense increases in the moral treatment (R3a), as expected from the literature on moral conviction and moral motivated reasoning (Section 2.2.5); this increase is accompanied by higher self-reported moral conviction. From this perspective, the observed rise in defensive goals reflects a stronger motivation to protect existing beliefs and to avoid information that could undermine one's preferred moral stance. At the same time, accuracy shows a weak upward trend, which is plausible because implications for human lives (rather than property damage) may simultaneously increase a desire to be accurate. These motivational shifts help interpret the pattern of information choice observed across contexts.

Motives shape explanation demand. The absence of a robust decision congruence × context interaction in the baseline model argues against a simple contextual moderation of the congruence effect. Instead, the pattern is consistent with a motivated reasoning perspective in which context and congruence affect information choice primarily depending on the situational balance of accuracy- and defense-related goals. Indeed, incorporating defense and accuracy provides a more informative account of explanation demand than relying on context and congruence alone. Accuracy is positively associated with explanation demand, whereas defense is negatively associated with it.

We also observe suggestive evidence for congeniality bias. Although the interaction between defense and decision congruence is only at trend level, the pattern is consistent with the idea that defensive motivation suppresses information demand more when decisions are incongruent. This view is echoed by the gSEM results,

which suggest an indirect pathway. In the moral treatment, incongruent decisions are associated with lower explanation demand via elevated defense motives.

Taken together, these findings help reconcile why an aggregate congruence × context interaction may not emerge. If incongruent decisions simultaneously increase epistemic value (via accuracy motivation and curiosity) and trigger threat-related processing (via defense and dissonance), the resulting opposing behavioral impulses can offset each other at the aggregate level, even when each pathway is substantively meaningful.

The changes in cognitive processing are accompanied by affective markers. Independently of context, incongruence increases dissonance, as outcomes that contradict one's own evaluation create an affective-cognitive conflict. Furthermore, dissonance is at a higher baseline-level in moral decisions when ethical trade-offs and personal values, norms, and self-concept are involved. Study 2 also reveals the translation of anticipatory negative affect — here captured by dissonance discomfort instead of STAI — into information avoidant behavior, a pattern that did not emerge in Study 1. Importantly, this association is contingent on decision congruence. When new information challenges one's own judgment, people may avoid it in order to reduce cognitive tension. From an information-avoidance perspective, this corresponds to deliberately not looking because the expected psychological costs (e.g., unpleasant self- or moral implications) outweigh the potential benefits of knowing.

Curiosity appears to be shaped by decision congruence in a context-dependent way. Incongruence may make a knowledge gap salient and thus render gap closure intrinsically (hedonically) rewarding, whereas congruence leaves little gap to close. Consistent with this logic, congruence in neutral scenarios produces a pronounced drop in curiosity and correspondingly in explanation demand. In neutral decisions, incongruence likely functions as a diagnostic signal that invites sense-making ("Why does the AI do this?"), thereby increasing epistemic value of the information. When decisions become moral, this informational impulse can be counteracted as anticipated uncongenial information may be experienced as threatening, and moral views may be perceived as less negotiable. This combination can dampen curiosity and, in turn, reduce explanation demand ("I am not interested because I am right."). Conversely, under congruence, morality may preserve curiosity because explanations can serve a validation function, supporting and legitimizing one's moral position rather than merely resolving uncertainty. Overall, this account aligns with information choice as

an affective trade-off and with curiosity as an approach signal that operates alongside motivated reasoning and congeniality bias when moral values are at stake.

2.4 General Discussion

Transparency in AI decision-making should not be treated as an end in itself. Explanations can only function as an effective safeguard against misalignment between morally charged outcomes of AI applications and societal values if people are willing to look at them. Building on the conceptualization of information demand as a value-based choice that may serve different motivations, explanation uptake depends not only on instrumental usefulness, but also on anticipated affective and self-regulatory consequences of "knowing" (e.g., Caplin and Leahy, 2001; Golman et al., 2017; Loewenstein, 1994, 2006; Sharot and Sunstein, 2020; Stigler, 1961; Sweeny et al., 2010). Our results suggest that morality systematically shifts the affective and cognitive motives that enter this trade-off, thereby shaping explanation uptake. Because these pathways can pull in different directions, the net effect of moral context on explanation uptake is inherently contingent on which affective and cognitive motives are most salient in a given use case.

2.4.1 Motivations and Selective Engagement in Moral Contexts

From this perspective, explanations are sought when their expected hedonic value (e.g., satisfying curiosity, closing knowledge-gaps, supporting accuracy or validation goals) outweighs their anticipated costs (e.g., discomfort, dissonance, threat to self-image). We find empirical patterns consistent with the broader premise that moral contexts are psychologically distinct environments that alter both sides of this equation.

Our results show that moral implications increase anticipatory anxiety and dissonance discomfort when participants consider reviewing an explanation. This suggests that moral transparency can entail *negative emotions* — a key precursor for information-avoidant behavior (Golman et al., 2017; Sweeny et al., 2010). However, we do not observe direct information avoidance in the sense that heightened anxiety translates into a simple moral-context main effect that reduces explanation demand (R1a). Moral decisions do not uniformly suppress explanation uptake, but affect likely matters

through more specific motivational structures that are contingent on whether user and AI agree.

Motivated-reasoning and selective exposure therefore provide the more informative theoretical lens. Moral context appears to shift *when* explanations are expected to be psychologically beneficial or costly, depending on whether they threaten or support one's prior stance. Study 2 captures this mechanism directly, balancing and analyzing *decision congruence* as a boundary condition and eliciting *accuracy and defense motivations*. In neutral treatments, explanation demand is higher when the AI's outcome is incongruent with one's own evaluation, as incongruence plausibly is perceived as a knowledge gap that invites sense-making, increasing the perceived epistemic value of an explanation and aligning with accuracy-oriented motives. While this pattern changes in moral contexts, we do not find robust evidence for a simple context-by-congruence interaction on explanation demand (R2b). This does not necessarily imply the absence of context effects. Incongruence can increase the perceived value of additional information while also eliciting threat-related processing, so opposing tendencies may cancel out in aggregate uptake even when each pathway is meaningful. Instead, model comparisons indicate that a motivation-based specification provides a better account of explanation uptake than the context-by-congruence term, suggesting that context effects are more accurately captured through shifts in accuracy and defense motives. When decisions become moral, explanations are evaluated through a more value-protective lens. This is reflected by an increase in defense motivation in the moral treatment (R3a). Defense-related motivation predicts lower explanation uptake, and this negative association appears more pronounced, although only at trend-level, when the outcome is incongruent, whereas accuracy motivation is positively associated with explanation uptake more generally (R3b). Crucially, defense does not necessarily imply only general avoidance. Under congruence explanations may retain value by providing validation and justification rather than challenging one's moral stance. Consistent with this, the effect of congruence on explanation uptake is descriptively attenuated in moral contexts. Overall, the pattern is consistent with moral motivated reasoning and congeniality-based selectivity: moral context heightens defense motives, which may selectively discourage engagement with potentially uncongenial explanations (Frey, 1986; Hart et al., 2009; Nickerson, 1998; Skitka et al., 2005, 1999; Taber and Lodge, 2006).

2.4.2 Emotions as Signals for Approach and Avoidance

Although motives are closer to the behavioral choice itself, emotions can serve as antecedents and markers, indicating when situations are experienced as threatening and when defensive regulation becomes relevant for information choice. Against this backdrop, it is notable that we observe clear affective *threat* signals even in hypothetical experimental settings — anticipatory anxiety in Study 1 and dissonance-related discomfort in Study 2. Anxiety might serve as a more general signal for arousal or partly reflect increased engagement or arousal rather than an avoidance trigger. Dissonance-related discomfort, by contrast, more directly complements the motivated reasoning and selective exposure account. Classic theories of cognitive dissonance and motivated reasoning emphasize that directional goals emerge as attempts to resolve dissonance tension and to protect valued beliefs or self-relevant judgments (Festinger, 1957; Kunda, 1990). From this perspective, dissonance functions as a threat cue: additional information becomes psychologically costly when it may undermine one's preferred stance. The dissonance pattern we find is consistent with this logic. First, dissonance increases under incongruence across contexts, as outcomes that contradict one's own evaluation create an affective-cognitive conflict. Second, dissonance is elevated at a higher baseline in moral decisions, where ethical trade-offs are tied to personal values, norms, and self-image. Third — and most importantly for selective exposure — dissonance translates into lower explanation demand, but crucially only when decisions are incongruent. This congruence-contingent avoidance response supports the interpretation that people regulate dissonance through directional processing strategies that foster congeniality bias and selective exposure.

Curiosity complements this picture as an *approach* signal that is a robust predictor of explanation uptake. Whereas dissonance indicates when explanations become psychologically costly, curiosity tracks when explanations are perceived as epistemically valuable and worth engaging with (Loewenstein, 1994). Across both studies, it varies with decision congruence in a context-dependent way. In neutral decisions, congruence significantly reduces curiosity, consistent with explanations offering less opportunity for new insight when user and AI already agree. In moral decisions, this congruence-related drop is attenuated, and curiosity becomes more selective. It can be suppressed under moral incongruence, when defense or validation goals are heightened, whereas congruent explanations can retain value by validating and legitimizing one's prior decision. In this way, curiosity reflects both information seeking, because higher curiosity promotes uptake, and congeniality-based selectivity,

because moral incongruence can dampen curiosity. Consistent with this interpretation, the association between curiosity and uptake weakens once accuracy and defense motivations are included, suggesting that curiosity partly overlaps with the motivational orientation that more directly governs the behavioral choice.

2.4.3 Implications for Oversight and Responsible Use

Our results reinforce that explanation uptake is not an automatic consequence of making information available, but a trade-off between anticipated benefits and costs. Moral contexts are precisely the settings in which transparency is often considered most critical, yet they also shift affective and motivational processes in ways that may promote or hinder engagement. This may become further relevant when outcomes are not hypothetical and when explanations are encountered routinely, reducing novelty. The experimental setting employed promotes accuracy goals and curiosity, potentially dampening more aversive effects from negative anticipatory emotions and defense goals. In real-world use cases, the relative weight between accuracy-oriented engagement and defensive avoidance will likely vary with contextual features such as consequence severity and immediacy, whether the decision-maker is personally affected or outcomes primarily affect others, perceived accountability and social evaluation, and the degree of control and reversibility users retain over the AI's action. This trade-off may further shift under time pressure and cognitive load, and with repeated exposure to similar decisions over time. Real consequences may amplify threat and avoidance when outcomes primarily affect others, while personal affectedness may heighten accuracy motives and increase willingness to engage with explanations.

Understanding these affective and motivational pathways and how they might play out in different settings is important for the design of transparency interventions. Explanations that aim to establish moral oversight should not only be designed for informational completeness, but also for psychological accessibility. In settings where moral threat and dissonance are likely to be high, explanation interfaces may need to reduce avoidable defensiveness and preserve epistemic engagement. This can involve communicating the scope and limits of an explanation, avoiding framing that implies moral indictment, and emphasizing that disagreement with the AI's recommendation does not signal a moral failure. At the same time, designs can

foster accuracy-oriented engagement and curiosity by highlighting concrete decision-relevant consequences, making uncertainty and knowledge gaps explicit even under congruence, and preserving user agency over the final action. When users are directly affected, emphasizing actionable guidance may further strengthen accuracy motives, whereas when decisions primarily affect others, additional attention may be needed to prevent defensive disengagement under incongruence. Overall, our results suggest that effective transparency in morally charged AI settings requires anticipating the affective and motivational conditions under which explanations will actually be read, rather than treating explanation provision as sufficient. At the same time, engagement under congruence is not necessarily synonymous with oversight, because explanations may also serve a reassuring or legitimizing function that reinforces acceptance rather than scrutiny (Bansal et al., 2021; Eiband et al., 2019).

2.4.4 Limitations and Future Research

Our findings should be interpreted in light of several limitations, which present potential opportunities for future research. First, our studies rely on hypothetical scenarios and explanation uptake without real consequences. Future work should test how the same affective and motivational pathways take effect when AI decisions have tangible implications, for example when outcomes affect participants directly or when accountability and social evaluation are more salient. Such settings may heighten threat and dissonance, but may also strengthen accuracy motives.

Second, each participant only reviews a single decision scenario. Novelty and situational curiosity may be comparatively high, which could dampen avoidance dynamics and produce ceiling effects that mask directional pathways. Future research should therefore examine repeated exposure across multiple decisions, where curiosity may decline over time and users become habituated to AI recommendations. This would allow testing whether selective avoidance and defense-driven disengagement becomes more pronounced once explanations are less novel and the cost–benefit trade-off changes.

Third, while we measure anticipatory anxiety, dissonance-related discomfort, and curiosity, moral AI decisions may elicit a broader set of emotions. Future studies could add more differentiated affective measures, such as moral outrage or guilt and shame, as well as behavioral indicators that capture avoidance more directly.

2.5 Conclusion

Prior XAI and HCI research has focused on what a "good" explanation should look like in terms of instrumental qualities such as clarity, accessibility, and technical correctness. Our findings complement this line of work by showing that transparency also has a behavioral precondition. Explanations can only contribute to accountability and ethical alignment if people choose to engage with them. We conceptualize explanation uptake as a value-based information choice, that also entails affective and self-regulatory costs and benefits, and is shaped by motivations. Across two studies, we show that moral context systematically alters affective and motivational determinants of engagement, increasing threat-related signals and defensive avoidance motivations. As a result, moral contexts do not simply raise or lower explanation demand. They change when explanations are experienced as epistemically valuable versus potentially aversive, which can promote motivated selectivity and congeniality-biased engagement.

These insights have implications for design and governance. Providing explanations does not guarantee sufficient human oversight in morally charged AI applications, because this safeguard may fail behaviorally when users do not consult explanations or do so selectively. In real-world use, where explanations are encountered repeatedly and become normalized, the curiosity and the perceived informational benefit may decline while affective and self-regulatory costs remain, making disengagement or selective scrutiny more likely precisely where governance depends on consistent oversight. Moreover, engagement does not necessarily imply appropriate scrutiny, since explanations can also serve reassuring or persuasive functions that increase acceptance independently of correctness. Effective transparency therefore requires attention not only to what explanations contain, but also to the affective and motivational conditions under which they are received.

3 *ChatGPT, Do You Interact Like a Human? Investigating Task-Dependent Human-Likeness of LLM-Enabled Conversational Agents Through Mind Perception*

Abstract

Large language model-enabled conversational agents (LLM CAs) are increasingly used for a wide range of tasks. These include computer-like tasks, such as coding, debugging, and data analysis, and human-like tasks, such as shopping advice, customer service, healthcare, and education. Drawing on mind perception theory, we demonstrate that the type of task is crucial for understanding and designing CA–user interaction. Task type influences perceptions of agency (cognitive capacity) and experience (emotional capacity). Further, it shapes user preferences for anthropomorphic design. This is particularly important as current LLM CAs autonomously adapt their behavior to task types. Across three experimental studies with 624 participants, we demonstrate that users prefer more anthropomorphic CAs for human-like tasks but not for computer-like tasks (Study 1) and that the need for experience is higher in human-like tasks, while needs for agency remain constant across different task contexts (Study 2). Study 3 investigates real-world use scenarios with ChatGPT, using chat data from 5,394 user prompts. While ChatGPT autonomously adapts its experience and agency cues based on task type, these adaptations misalign with user preferences: experience cues are insufficient, and agency cues decrease for human-like tasks. This misalignment negatively affects trust, which is lower for human-like tasks. However, this effect can be mitigated if the CA is perceived to have experience and agency. Our findings advance the understanding of task-dependent anthropomorphic design, highlighting the limitations of current autonomous adaptations, and offering actionable insights for deliberate user-centered improvements.

[0] This chapter is joint work with Anna-Maria Seeger, Jella Pfeiffer, and Armin Heinzl.

3.1 Introduction

With the tremendous rise and popularity of large language model-enabled conversational agents (LLM CAs), both public voices (Malik, 2022; Weil, 2023) and researchers (Bender et al., 2021; Mitchell and Krakauer, 2023; Susarla et al., 2023) have been interested in how the natural language capabilities of LLM CAs blur the line between what is human and what is artificial intelligence (AI). Some research highlights opportunities and potential gains for efficiency and effectiveness (Dell'Acqua et al., 2023; Kshetri et al., 2024). Other researchers call for caution as the downstream effects of this unprecedented level of AI human-likeness are still unknown and may be dangerous (Bender et al., 2021; Mitchell and Krakauer, 2023). Despite this duality, our understanding of both the extent to which LLM CAs are perceived as humans and the mechanisms through which this humanness perception occurs in interactions with LLM CAs remains limited.

Research in social psychology and human–computer interaction consistently demonstrates that human-like cues in technology design cause mind perception and anthropomorphism (Gray and Wegner, 2012; Seeger et al., 2021). Through *mind perception* (Gray et al., 2007), people form cognitions of anthropomorphism, which refers to the attribution of human qualities to a nonhuman entity, including consciousness, intentions, and emotions (Epley et al., 2007). Given that natural language can function as a human-like cue in interactions with technology, we draw on theories of mind perception (Gray et al., 2007, 2012) and anthropomorphism (Epley et al., 2007; Waytz et al., 2010c) to better understand perceived humanness of LLM CAs.

Information Systems (IS) research on human–CA interactions has investigated the role of anthropomorphic CA design. For instance, anthropomorphic design demonstrably increases offering bids (Schanke et al., 2021), trust (Schuetzler et al., 2021), and customers' service evaluations (Han et al., 2023), while it should be cautiously employed when customers are angry (Crolic et al., 2022). These studies provide important insights into how we engage with human-like conversational technologies, yet their findings do not readily apply to LLM CAs for two important reasons. First, LLM CAs cover a much broader range of use contexts and tasks than previous CAs. For instance, earlier IS research on CAs focused on tasks such as customer service (Gao et al., 2023; Gnewuch and Reinkemeier, 2025; Han et al., 2023; Schanke et al., 2021) or healthcare and education (Diederich et al., 2022). We refer to these tasks as *human-like* tasks. Additionally, LLM CAs are used in various other domains including

coding, debugging, data analysis, and analytical problem-solving. We refer to these tasks as *computer-like* tasks. Existing studies on human-like task domains consistently demonstrate that users value anthropomorphic CA design. Yet, it remains unclear whether this appreciation extends to more computer-like CA tasks. Recent studies have explored software developers' interactions with LLM CAs in performing coding tasks (Qian and Cong, 2023; Ross et al., 2023), but the influence of anthropomorphic design on user perception and behavior in such technical domains remains largely unexplored.

Second, extant IS research on anthropomorphic CA design operates on the assumption that the CA provider can control the anthropomorphic design and interactive behavior of the CA by making visual design decisions, defining intents, anticipating responses, and setting up the fulfillment logic (e.g., using Google Dialogflow). However, with LLMs, the degree of control is limited due to the inherent randomness in the language generation process. While providers of LLM CAs can prompt the system to use anthropomorphic cues, they consequently do not have full control over its anthropomorphic behavior or complete transparency into the billions of parameters that define its foundational model. Therefore, we need to better understand the effects of LLM CA's autonomous anthropomorphic adaptations within user interactions on user perceptions and evaluations. In contrast to previous studies on anthropomorphic CA design (Gao et al., 2023; Han et al., 2023; Schanke et al., 2021; Schuetzler et al., 2021), we explicitly focus on the interactive behavior without targeting the anthropomorphic appearance of LLM CAs (e.g., the use of avatars, images, human names). Instead, we intend to investigate the anthropomorphic design potential (i.e., use of verbal and non-verbal human-like cues) — that emerges autonomously from the use of LLMs. Since LLMs are trained on vast amounts of human-generated text data from both human-like and computer-like task domains, we presume that LLM CAs will autonomously exhibit more anthropomorphic interactive behavior when performing human-like tasks than when performing computer-like tasks. To address the identified research gaps, we aim to investigate the following research questions:

Question 1 *How does the importance and choice of anthropomorphic CAs vary between human-like and computer-like tasks?*

Question 2 *Do LLM CAs autonomously adapt their anthropomorphic interactive behavior to these task types?*

Question 3 *What is the effect of an LLM-enabled autonomous adaptation of anthropomorphic design on user perceptions and evaluations?*

We investigate the interactive behavior of LLM CAs across tasks through the lens of mind perception theory, which conceptualizes experience (capacity to feel and sense) and agency (capacity to act and think) as dimensions of mind attributed by observers (Gray et al., 2007; Waytz et al., 2010a). Analyzing mind perception assists us in understanding whether the nature of anthropomorphic perceptions elicited by LLM CAs are capable of stimulating user interactions as well as how users perceive and evaluate them across different tasks. In line with mind perception theory, we propose that interactions with LLM CAs will demonstrate agency for both human- and computer-like tasks, whereas experience pertains to human-like tasks only. To test our hypotheses, we present results from three experimental studies. Study 1 establishes the general phenomenon of how task context shapes users' choices regarding anthropomorphic design. Study 2 investigates user needs for experience and agency across different task contexts to provide a theoretical account of the phenomenon observed in Study 1. Study 3 extends this investigation into real-world applications by conducting an experiment using a LLM CA (ChatGPT) to explore the CA's autonomous adaptation of anthropomorphic cues and associated user perceptions. We analyze 5,394 prompts from this experimental study, manipulating the task type to explore how the LLM CA adapts its anthropomorphic cues during actual interactions. These complementary perspectives allow us to establish the phenomenon (Study 1), understand user needs driving the phenomenon (Study 2), and validate its relevance in practical interactions with currently available LLM CAs (Study 3), providing a comprehensive understanding of mind perception toward LLM CAs across task types.

Our theoretical framework and findings provide the following primary contributions to theory and practice. Firstly, we add to mind perception theory in the context of human–computer interaction by demonstrating that task type is an important factor for understanding how the dimensions of experience and agency influence the perception of anthropomorphism toward technology. Study 1 provides evidence that users select anthropomorphic CAs over non-anthropomorphic CAs for human-like tasks, but the opposite for computer-like tasks. Building on this finding, Study 2 reveals that the need for agency and experience differs between human-like and computer-like tasks: users desire high levels of both agency and experience for

human-like LLM CA tasks, whereas computer-like tasks require high levels of agency only.

Secondly, we add novel insights to the IS literature on anthropomorphic CA design by considering the effects of autonomous adaptation of anthropomorphic interactive behavior enabled by LLMs. While prior IS literature focused on controllable anthropomorphic design features (Gao et al., 2023; Han et al., 2023; Schanke et al., 2021), our research investigates the LLM's inherent ability to generate anthropomorphic interaction behavior. Our findings provide evidence that LLMs indeed autonomously adapt such behavior; yet these mechanisms do not necessarily align with user needs and mind perception theory as LLM CAs make only moderate use of experience cues in human-like tasks, and agency cues are even reduced for these tasks. In terms of user evaluations, we demonstrate that human-like tasks significantly reduce trust — users tend to be skeptical about machines performing these tasks, consistent with the concept of algorithm aversion (Castelo et al., 2019; Jussupow et al., 2024). This negative effect on trust is mitigated when the CA is perceived to have experience and agency. From a broader perspective, our research informs the ongoing debate on the extent to which LLM CAs can provide synthetic humanness, while also elucidating the requirements for user-oriented anthropomorphic design of LLM CAs.

3.2 Theoretical Foundations and Development of Hypotheses

3.2.1 Theories of Mind Perception and Anthropomorphism

Mind perception research revolves around the question of whether other entities, of human or nonhuman origin, are perceived to have a mind (Gray et al., 2007, 2012; Waytz et al., 2010a). The perception of a mind in another entity affects how individuals interact with that entity (Gray and Wegner, 2012; Yam et al., 2021). When the other entity is of nonhuman nature, the process of attributing human-like cognitive and emotional capacities to that entity is named *anthropomorphism* (Epley and Waytz, 2010; Epley et al., 2007; Waytz et al., 2010a). Research on mind perception and anthropomorphism argues that human and nonhuman entities are perceived along two dimensions, namely *experience* and *agency* (Gray et al., 2007; Waytz et al., 2010a, 2014; Yam et al., 2021). Agency reflects the cognitive dimension of mind perception (Gray et al., 2007; Waytz et al., 2010a), and refers to the perceived capacity to engage in reasoned action, strategic planning, and goal-directed behavior, thus attributing

to an entity the ability to form and communicate preferences, beliefs, and explicit knowledge (Epley and Waytz, 2010). Experience reflects the emotional and social dimension of mind perception (Gray et al., 2007; Waytz et al., 2010a) and refers to the perceived capacity to understand and convey emotions and basic psychological states (Epley and Waytz, 2010).

While other theoretical perspectives, such as the *stereotype content model* (Fiske et al., 2002) and *dehumanization* theory (Haslam, 2006), provide valuable insights into how social groups are perceived, they are less suitable for our study because they primarily address human-to-human perception, categorization, and moral judgment. By contrast, our research focuses on how people attribute mental capacities to non-human agents (LLMs), making mind perception theory a better conceptual fit.[1] Anthropomorphic cues can create *social presence*, which influences trust and engagement (Gefen and Straub, 2004; Nowak and Biocca, 2003; Short et al., 1976). We therefore treat social-presence effects as captured within our task-dependent mind perception framework.

Consistent with mind perception theory, attributions of agency and experience vary by entity: technical partners such as robots are typically granted moderate agency but little experience, babies and pets are granted high experience but low agency, and only adult humans are granted high levels of both (Bigman and Gray, 2018; Gray et al., 2007; Yam et al., 2021). Comparative work further demonstrates that humans receive the highest attributions overall, and that anthropomorphic robots (e.g., Google Home, Robot Kaspar) are attributed more agency and experience than non-anthropomorphic machines (Xu and Sar, 2018).

Mind perception research has primarily examined how features of an agent affect attributions of agency and experience, while comparatively less work examines how the interaction context and the task influence these attributions. At the same time, human–computer interaction research in the field of CAs has indicated that the nature of the interaction context (Castelo et al., 2019; Glikson and Woolley, 2020; Hong et al., 2014, 2007; Kaplan et al., 2023) plays a critical role in shaping user perceptions and behaviors. By integrating insights from mind perception theory with findings from CA literature, our study seeks to bridge these two areas. Specifically, we investigate how the task type — whether predominantly human-like or computer-like — affects

[1] A more detailed discussion of these alternative theoretical perspectives is provided in Appendix C.4.

the perception of agency and experience in interactions with LLM CAs. This approach offers a more comprehensive understanding of how both the nature of the interaction partner and the task context jointly shape mind perception.

3.2.2 Human- and Computer-like Tasks of LLM CAs

The versatility of text-based tasks that can be performed with the help of LLM CAs is unprecedented (Susarla et al., 2023). These tasks span a wide spectrum, including code or text generation and analysis, general reasoning, engaging in dialogues for various purposes (e.g., customer service, shopping advice, medical consultations, general conversations), controlling home devices, or providing factual information (Deng and Lin, 2023; Kocoń et al., 2023; Qin et al., 2023).

Previous studies have classified tasks performed by algorithms as objective — requiring rational, logical analysis — and subjective – requiring intuition and personal interpretation (Castelo et al., 2019; Dietvorst et al., 2015, 2016). However, this dichotomy may not fully capture the complexity of tasks where human-like qualities are necessary. Research on algorithmic decision-making instead suggests distinguishing computer-like from human-like tasks (Kordzadeh and Ghasemaghaei, 2022; Lee, 2018). This aligns with the dimensional view of mind perception. Human-like tasks hinge on experience and can involve varying degrees of agency. Computer-like tasks primarily draw on agency, while placing minimal demands on experience.

Human-like tasks are defined by the need for a CA to substitute for a human interaction partner (Doyle, 1999; Lankton et al., 2015; Norman, 1994). These tasks often involve nuanced conversations that require mutual exchange of complex ideas, information, and emotions (Doyle, 1999; Seeger et al., 2021; Waytz et al., 2014). Many of these tasks are subjective, open to interpretation and informed by personal experiences, gut feelings, and intuition (Castelo et al., 2019; Logg, 2017). However, human-like tasks may involve both analytical and emotional components. For example, offering medical advice includes an objective component that involves logical assessments of facts (Castelo et al., 2019; Clark et al., 2025), but also requires subjective elements, such as offering empathetic feedback (Seeger et al., 2021) and ensuring that patients feel their unique circumstances are considered (Longoni et al., 2019). Similarly, conducting a moral or jurisdictive judgment requires an objective analysis of facts and reasoning (Castelo et al., 2019), but the ultimate decision on what is morally or

legally right or wrong depends on personal normative interpretation (Awad et al., 2018; Gogoll and Müller, 2017).

Computer-like tasks are those where the CA does not substitute for a human interaction partner but is rather seen as a technical tool (Lankton et al., 2015). These tasks typically fall within the objective domain, focusing on solving problems based on quantifiable and measurable facts (Castelo et al., 2019; Logg et al., 2019). LLM CAs are proficient in handling a wide range of such tasks, including data analysis, software coding, problem-solving, and system control.

User perceptions may evolve and the border between human-like and computer-like might be fuzzy. What is perceived to be a traditionally more human-like task and what may be perceived as more computer-like may change as LLM CAs become more advanced. In order to still be able to make causal claims by assigning participants to either human-like or computer-like treatments, we therefore operationalize task type empirically: in a pre-study participants rate tasks on a scale to be more human-like or computer-like, and we use the most prototypical tasks in subsequent studies.

3.2.3 Research Model Development

Our research model integrates three complementary perspectives to provide a comprehensive understanding of mind perception toward LLM CAs across different task types. First, we demonstrate the phenomenon by providing evidence that users choose different anthropomorphic CA designs based on the specific nature of the task at hand. Next, we investigate the reasons behind these choices and compare user needs across different tasks, identifying the specific demands placed on LLM CAs through the lens of mind perception theory, particularly in terms of agency and experience. Finally, we analyze real user interactions with LLM CAs, providing insights into how mind perception is reflected in practical or jurisdictive use and how it is automatically embedded in current technology.

3.2.3.1 The Phenomenon: Task Dependent Choice of Anthropomorphic CA Design

Extant IS research repeatedly finds that anthropomorphic CA design — via visual, auditory, and language cues — tends to have positive effects and raises user trust (e.g. Gao et al., 2023; Han et al., 2023; Lee et al., 2020; Schanke et al., 2021; Schuetzler et al., 2021; Seymour et al., 2025). Most evidence comes from customer-service (Blut et al.,

2021; Cheng et al., 2022) and online shopping advice (Hess et al., 2009; Morana et al., 2020; Qiu and Benbasat, 2009; Yuan and Dennis, 2019), i.e., human-like tasks that mix information processing with empathy and personalization. There is, however, a huge potential for LLM CA adoption in the domain of computer-like tasks such as coding support and data analysis (Susarla et al., 2023).

While previous studies on human-like tasks support positive effects of anthropomorphic design, we argue that it may not yield similar benefits for computer-like tasks, where users may appreciate CAs for their programmed technical superiority in terms of rationality, reliability, and objectivity. Here, anthropomorphic design may signal human fallibility rather than algorithmic precision. Evidence from high-stakes technical settings aligns with this view: in professional environments (i.e., nuclear power plants, air navigation) users rate decision support systems for such computer-like tasks as more skilled and knowledgeable than human experts (Dzindolet et al., 2001; Skitka et al., 1999) and added anthropomorphism does not increase user trust (de Visser et al., 2016; Gruber et al., 2018).

Prior research has highlighted different preferences for experience and agency in specific contexts (Adam et al., 2022; Appel et al., 2020; Wiese et al., 2022): preferences shift across sales stages; "emotional" robots (high experience/high agency) are favored for social tasks, "unemotional" robots (low experience/high agency) for arithmetic tasks; and higher experience can evoke eeriness, mitigated in nursing. However, the influence of task-context has not been systematically explored for human-like versus computer-like tasks and mind perception was manipulated only by description of the robots.

In summary, we argue that anthropomorphic design is valuable to users when an LLM CA is supposed to perform a human-like task, because such design signals the emotional and cognitive capacities considered necessary for performing these tasks. By contrast, when an LLM CA is expected to perform a computer-like task, anthropomorphic design conflicts with the idea of precision and rationality required in such contexts.

Hypothesis 1 *For human-like tasks, users choose anthropomorphic CAs over non-anthropomorphic ones; for computer-like tasks, the opposite is true (to be tested in Study 1).*

3.2.3.2 User Needs

In a next step, we seek to understand and account for differing user needs that explain the choice phenomenon for anthropomorphic design across task types by building upon research on mind perception. In their early applications, robots or bots have traditionally taken on tasks that relate to the computer-like task domain, including classification, navigation, object recognition and manipulation (e.g. De Santis et al., 2008; Sheridan, 2016). Similarly, mind perception studies have often examined robots while they perform computer-like tasks, such as object identification and pick-up (Xu and Sar, 2018; Yam et al., 2021) or autonomous movement (Gray and Wegner, 2012; Xu and Sar, 2018). In recent years, specific research has explored the effect of mind perception on the acceptance and perception of social robots in the context of nursing and eldercare (Appel et al., 2020; Stafford et al., 2014). We argue that robot-based findings on agency and experience extend to LLM CAs performing computer-like tasks. Accordingly, such CAs should exhibit moderate agency to signal rational competence and low experience to fit the task.

As described in Section 3.2.2, human-like tasks require nuanced exchanges involving emotions, intuition, and logical assessment, making users value agents that display emotional understanding (experience) and independent reasoning (agency). Adult humans are the benchmark for both agency and experience (Gray et al., 2007; Xu and Sar, 2018). Accordingly, LLM CAs performing human-like tasks should convey high levels of both. Experience cues enable meaningful interaction by signaling emotional and experiential capacity, while agency cues signal the cognitive competence required for the task. Figure 3.1 positions the two task types along these dimensions.

Hypothesis 2 *For human-like tasks, users desire more experience-based (H2a) and more agency-based (H2b) anthropomorphism compared to computer-like tasks (to be tested in Study 2).*

3.2.3.3 Autonomous Design Adaptation across Task Types

Next, we seek to investigate the actual behavior of LLM CAs in terms of anthropomorphic design across task types. Generative text AI utilizes LLMs to read, analyze and generate text content (Susarla et al., 2023). The transformer architecture enabled effective long-range context modeling, underpinning today's LLMs, which are specialized via fine-tuning (Bender et al., 2021; Vaswani et al., 2017). LLMs, such as GPT,

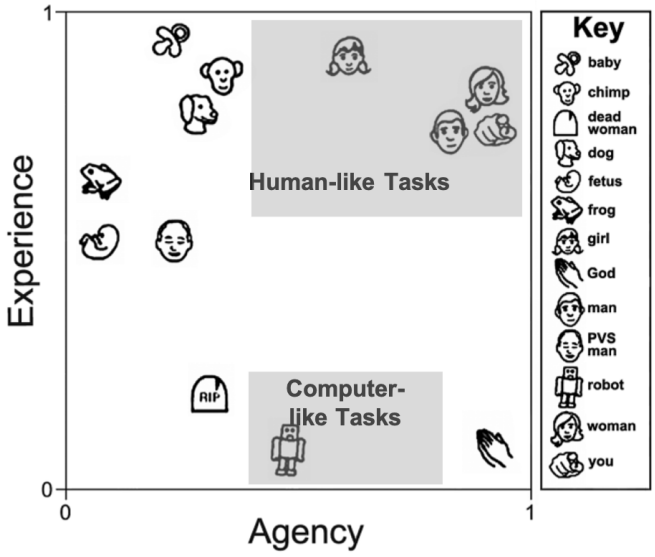


Figure 3.1: Dimensions of mind perception by Gray et al. (2007) with the proposed location of LLM CA task types.

Gemini, Claude and DeepSeek are trained on massive web-scale corpora (Susarla et al., 2023), e.g., CommonCrawl (Brown et al., 2020). As statistical pattern learners, LLM CAs can reproduce and amplify dataset biases (Bender et al., 2021; Lucy and Bamman, 2021; Lund and Wang, 2023).

Analogously, we examine whether LLM CAs exhibit an autonomous, task-dependent anthropomorphic bias along the agency and experience dimensions. Because LLMs generate text by statistically mirroring domain-specific corpora, their outputs likely reflect each domain's anthropomorphic density: human-like domains (e.g., customer reviews, health forums) contain abundant experience cues and some agency cues, whereas computer-like domains (e.g., scientific papers, coding forums) offer moderate agency cues and few experience cues. Consequently, we expect more anthropomorphic language — especially experience cues, and also agency cues — in human-like tasks than in computer-like tasks.

Adaptation of language used in the communication in human–computer interaction may be bidirectional. Prior research shows that perception of the AI's intelligence and anthropomorphism influences users' language style (Wang et al., 2021). Mind perception may not only influence how users engage with AI but also shape how AI responds, reinforcing a dynamic feedback loop in AI-user interaction.

Hypothesis 3 *LLM CAs autonomously adapt their anthropomorphic style to the task type, using more experience cues (H3a) and more agency cues (H3b) in human-like-tasks than in computer-like tasks (to be tested in Study 3).*

Prior work shows that increasing experience and agency cues raises the corresponding attributions to artificial agents (Bigman and Gray, 2018; Gray and Wegner, 2012); therefore, if LLM CAs use more of these cues in human-like tasks, users should report higher perceived experience and agency.

Hypothesis 4 *Human-like tasks as opposed to computer-like tasks have a positive indirect effect on perceived experience, via increased use of experience cues by the LLM CA (H4a), and perceived agency via increased use of agency cues (H4b, to be tested in Study 3).*

Finally, a key question is how LLM CA's autonomous, task-dependent adaptation of anthropomorphic language cues affects user evaluations. Research on CAs and other technological agents has reported that individuals rely on and trust technology more for tasks that require objective and rational analysis of facts as opposed to tasks that also entail emotional intelligence (Bigman and Gray, 2018; Castelo et al., 2019; Logg et al., 2019). This provides a basis to expect that trust in LLM CAs is higher for computer-like than for human-like tasks. Thus, we hypothesize:

Hypothesis 5 *User trust in LLM CAs is higher for computer-like tasks than for human-like tasks (to be tested in Study 3).*

At the same time, we expect that despite this overall negative effect of human-like versus computer-like tasks on user trust, the increased perception of experience and agency in human-like tasks can benefit trust. Agency and experience can make the LLM CA appear more trustworthy for human-like tasks because these tasks require the LLM CA to demonstrate empathy, intuition, goal-directed behavior, and mindfulness. By contrast, computer-like tasks require a CA to demonstrate technical competence reflected in rational, measurable, and predictable behavior. When talking about personal health issues, for example, CAs that appear to care about the unique circumstances of the user whilst showing empathy are perceived as more competent and trustworthy than CAs that simply list measures and statistics to the user's individual problem (Longoni et al., 2019). Extant research on human–CA interactions in the context of customer service and online shopping has corroborated this point,

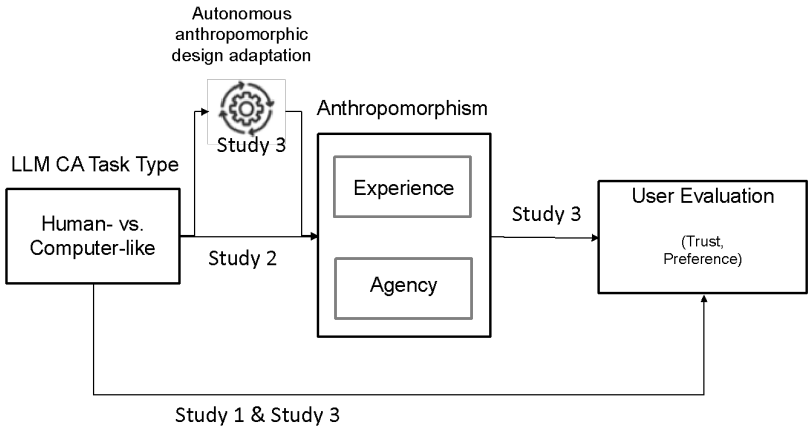


Figure 3.2: Overview of research model and studies.

suggesting that when CAs are anthropomorphized, they tend to be evaluated more positively (Morana et al., 2020; Schanke et al., 2021; Schuetzler et al., 2021).

Extensions of H5: *Human-like tasks as opposed to computer-like tasks have a positive indirect effect on user trust through increased perception of experience (H5a) and agency (H5b).*

3.3 Research Design

To empirically evaluate our research model, we conducted three experimental studies, each focusing on one specific aspect of our overarching research model (Figure 3.2) to reveal the underlying mechanisms of mind perception toward LLM CAs (Maruping et al., 2025). Study 1 demonstrates the phenomenon of task-dependent choices for anthropomorphic design in a laboratory experiment. Study 2 investigates task-dependent choices by examining the specific requirements for experience and agency across human- and computer-like tasks, using an online experiment. Finally, Study 3 shifts to a use perspective by examining actual interactions with an LLM CA in a classroom-experiment to explain how the agent autonomously adapts its anthropomorphic language cues across human- and computer-like tasks and how users evaluate this adaptation. Together, these studies offer complementary insights that enhance our understanding of mind perception toward LLM CAs, emphasizing the interplay between perceived characteristics, user needs, and user interactions across different task types.

3.3.1 Study 1: The Phenomenon on User Choice

In Study 1, we investigate users' task-dependent preferences for anthropomorphic design toward CAs as suggested by Hypothesis 1.[2] We conduct a laboratory scenario experiment with a 2 (within-subjects factor: low vs. high anthropomorphic design) × 2 (between-subjects factor: human-like vs. computer-like task) factorial design.

3.3.1.1 Procedure and Measures

Participants were invited via a standard subject pool in a university-owned research lab. An a priori G*Power analysis ($\omega = 0.3, \alpha = 0.05, 1 - \beta = 0.85$) indicated a required sample of 100 (Erdfelder et al., 1996; Faul et al., 2007). To allow for exclusions, 120 participants were invited. Seven failed an attention check and were excluded, yielding a sample size of $N = 113$ ($M_{age} = 22.53$, SD = 2.58; 38 female, 33.6%). Participants received a show-up fee of €5 and the study lasted approximately 13 minutes (M = 10.37, SD = 3.42).

Upon arrival in the lab, participants read and agreed to the experiment's terms and conditions. Afterward, we provided participants with a general description of CAs to ensure that everyone recognized that CAs are computer programs and not human interaction partners as well as to provide a common understanding of the technology. Each participant was then randomly assigned to one of the two task types (computer- vs. human-like). The CA performing the human-like task was a health agent that listens to users' mental health related problems and provides advice to handle such problems. The CA performing a computer-like task was a smart home agent that monitors and controls connected lighting and heating systems and provides usage statistics. These two CA types were adapted from Seeger et al. (2018, 2021), who identify them as among the most human-like and most computer-like types with respect to their tasks. Next, participants were provided with an interactive demonstration of two versions of the CA for their assigned task — one with low and one with high anthropomorphic design — before selecting the CA version that they preferred to perform the respective task. For the manipulation of the low and high anthropomorphic CA designs, we relied on previous studies that manipulate the

[2] Study 1 draws on an experimental paradigm and a subset of data previously used and reported by Seeger (2021). While the underlying experimental design and part of the data overlap, the present dissertation pursues a different research question and adopts a distinct analytical focus.

human-likeness of text-based CAs through verbal and non-verbal cues (Diederich et al., 2019a,b; Feine et al., 2019; Morana et al., 2020; Seeger et al., 2018). We did not manipulate any human identity cues (e.g. adding a human name or image) to make the CA manipulation comparable to publicly available LLM CA tools (e.g. ChatGPT, Google Gemini), which are used in Study 3. Anthropomorphic verbal cues manipulate the linguistic behavior of the CA and are reflected in the use of self-references (i.e., "I," "me"), salutations (i.e., "hello!"), and emotional expressions (i.e., apologies, concerns) by the high anthropomorphic CA. Anthropomorphic nonverbal cues are behavioral cues that are not purely linguistic and are reflected in the use of emoticons by the high anthropomorphic CA. An example is provided in Appendix C.1.1, Table C.1.

The order of the demonstration of each of the two CA versions (low vs. high anthropomorphic design) was randomized. In each version of the CA, participants navigated through a dialogue between a user and a CA. Participants had to press the space bar to trigger the next text message in the dialogue. Such interactive video vignettes create a more engaging and realistic scenario than static screenshots and have successfully been applied to research on interactions with anthropomorphic technological agents (Dennis et al., 2020; Mozafari et al., 2020; Nørskov et al., 2020; Stolte, 1994). After each CA demo (low and high anthropomorphic design), participants answered questionnaire items measuring perceived anthropomorphism as a manipulation check. Then participants had to decide which of the two versions they preferred for the specific task type. The experiment ended with a questionnaire including measures for task type manipulation check and demographic data.

3.3.1.2 Results

To verify the effectiveness of our task manipulation (between-subject), we asked participants to indicate if the presented task of the CA was more typical to be performed by a human or by a computer (1 = "very human-like", 7 = "very computer-like"). We adapted this approach of categorizing elements along a human–computer continuum from Touré-Tillery and McGill (2015). An independent sample t-test indicated a significant group difference, $t(111) = -6.14, p < 0.001$, confirming that participants in the human-like task condition perceived the tasks as more human-like ($M = 3.17, SD = 0.23$) than participants in the computer-like task condition ($M = 5.09, SD = 0.21$).

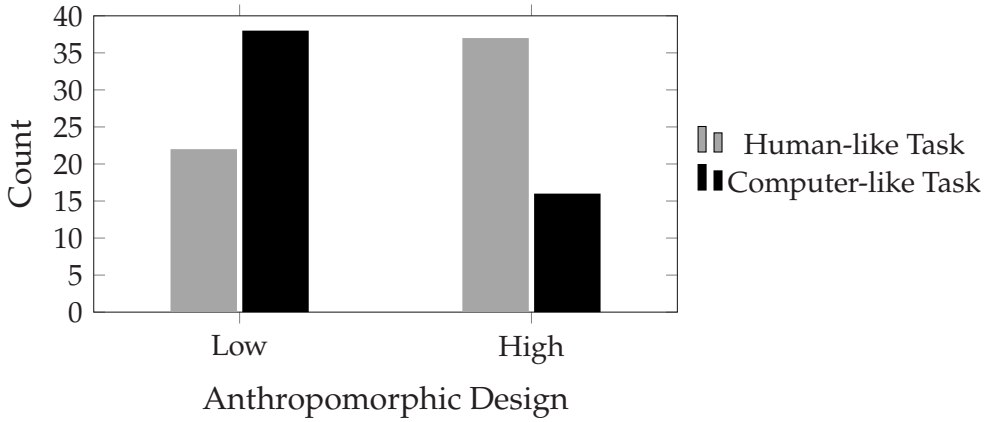


Figure 3.3: User choice between low versus high anthropomorphic CA design in human-like versus computer-like tasks.

Second, to verify the anthropomorphic design manipulation and confirm that participants noticed the difference between the CA versions, we compared perceived anthropomorphism ratings for the low- versus high-anthropomorphic CA. Participants were asked to indicate the extent to which they believed the CA appeared to have a "mind of its own", "intentions", "free will", "consciousness", "desires", "beliefs" and the "ability to experience emotions" on a seven-point Likert scale. This is an established measure of perceived anthropomorphism (Gray et al., 2012; Waytz et al., 2014, 2010c). These items were averaged into a composite ($\alpha = 0.94$). A paired t-test confirmed the manipulation. The high-anthropomorphism design was rated more anthropomorphic ($M_{high} = 3.01, SD = 0.14$) than the low design ($M_{low} = 2.68, SD = 0.13, t(112) = -3.91, p < 0.001$).

Table 3.1 presents how often participants in the computer- vs. human-like treatment group selected the low or high anthropomorphic CA design. Analyzing these data with Pearson's χ^2 test indicated that participants in the computer-like treatment group preferred the low anthropomorphic CA design (70.37%), as opposed to participants in the human-like treatment group who preferred the high anthropomorphic CA design (62.71%), $\chi^2(1) = 12.39, p < 0.001$, as depicted in Figure 3.3. The preference decision provides support for Hypothesis 1.

Table 3.1: User choice between low versus high anthropomorphic CA design in human-like versus computer-like tasks.

			Anthropomorphic Design	
			Low	High
Task Type	Human-like	Observed Frequency	22	37
		Expected Frequency	31.3	27.7
	Computer-like	Observed Frequency	38	16
		Expected Frequency	28.7	25.3
		Pearson $\chi^2(1) = 12.3904, p < 0.001$		

3.3.1.3 Discussion

Study 1 shows that user needs and evaluations regarding anthropomorphic CA design vary by task (H1). While prior IS work often finds positive effects of anthropomorphism, it largely examines human-like contexts (e.g. Gao et al., 2023; Han et al., 2023; Schanke et al., 2021; Schuetzler et al., 2021). Our results indicate that for computer-like tasks, users prefer less anthropomorphic designs. These findings motivate Study 2, which examines how needs differ across tasks, and Study 3, which examines how users evaluate LLM CAs' autonomous task-dependent adaptation.

3.3.2 Study 2: User Needs

Study 2 tests whether task-dependent preferences for anthropomorphic CAs are reflected in differing needs for agency and experience across human-like vs. computer-like tasks and CA vs. human partners.[3]

3.3.2.1 Pre-Study on Task Type

In Study 1, we relied on one human-like and one computer-like task from previous work. In this study, we improved the robustness of our findings by increasing the number of tasks to two per task type. Given that task classifications may change over time (Seeger et al., 2021), we sought to ensure that the classification of tasks as human-like or computer-like was up to date. Therefore, we conducted a pretest to

[3] Preregistered at AsPredicted #165013, ethics certificate at German Association for Experimental Economic Research e.V. #ZSoRFiHx.

establish our manipulation of human- and computer-like CA tasks and to consider additional possible CA tasks that LLM-based CAs can support before the main experiment. In a separate Prolific study, participants rated whether each of 12 tasks was more typically performed by a computer or by a human (1 = "very computer-like", 7 = "very human-like"; see Appendix C.1.2). Of 200 invited participants, two failed language checks and one failed an attention check. Thus, the final sample comprised $N = 197$ (133 female; $M_{age} = 43.92$, SD = 13.96). The mean completion time was 3.02 minutes (SD = 1.32). Table 3.2 provides an overview of the ratings across all tasks.

A paired-samples t-test comparing the averages across human-like versus computer-like tasks indicated a significant difference, $t(196) = -38.34, p < 0.001$, with lower ratings for computer-like tasks ($M = 2.72, SD = 0.06$) than for human-like tasks ($M = 5.66, SD = 0.05$). This pattern held even for the closest pair: the top computer-like task (Data Analysis and Interpretation; $M = 3.38, SD = 0.12$) was rated lower than the lowest human-like task (Shopping Recommendation; $M = 4.78, SD = 0.11$; $t(196) = -8.33, p < 0.001$).

Based on this pretest, we selected two computer-like and two human-like tasks to be used in the main Study 2. For the computer-like tasks to be used in the main study, we chose the coding task (Task 4) and the information search task (Task 5), as these two are well-suited to be performed with publicly available versions of LLM CAs. Tasks that involve calculations (Tasks 2, 3, 6) may result in errors with these LLM CAs and Task 1 requires integration and application with existing smart home systems, which cannot be achieved with the public version of the LLM CA. For the human-like task context, we selected the two most human-like tasks (Task 11 and 12).

3.3.2.2 Main Study 2

We conducted an online experiment through Prolific using a 2 (within-subjects factor: human-like vs. computer-like task) x 2 (between-subjects factor: LLM CA or human interaction partner) factorial design. An a priori G*Power analysis assuming a small effect ($f = 0.143$) and power = 0.90 indicated a minimum $N = 132$. To allow for exclusions, we invited 160 participants; three failed the language check and four failed the attention check, yielding a final sample of $N = 153$ (118 female).

Procedure and Measures. Participants were randomly assigned to complete the tasks either with the help of an LLM CA (treatment group) or with a human interaction partner (control group). Each participant was confronted with all four task scenarios

Table 3.2: Human- vs. computer-likeness ratings for all candidate tasks, pretest.

Task	Task Description	Task Type	N	Mean	Std. dev.	Min	Max
Task1	Monitor Smart Home System	computer-like	197	1.99	1.31	1	7
Task2	Statistical Tests/Predictions	computer-like	197	2.23	1.14	1	7
Task3	Solve Mathematical Problem	computer-like	197	2.87	1.59	1	7
Task4	Coding	computer-like	197	2.90	1.65	1	7
Task5	Information Search	computer-like	197	2.96	1.70	1	7
Task6	Data Analysis and Interpretation	computer-like	197	3.38	1.74	1	7
Task7	Shopping Recommendations	human-like	197	4.78	1.48	1	7
Task8	Customer Service	human-like	197	5.09	1.58	1	7
Task9	Teaching a Language	human-like	197	5.27	1.24	1	7
Task10	Diagnose Physical Health Condition	human-like	197	5.97	1.15	2	7
Task11	Determine Moral Responsibility	human-like	197	6.41	1.01	1	7
Task12	Diagnose Mental Health Condition	human-like	197	6.43	0.80	2	7

(within-subject factor). After a general introduction to the experiment, participants in the LLM CA condition were provided with a basic explanation and examples of LLM CAs to ensure they understood this technology. Subsequently, the four task scenarios were presented to each participant in random order. Participants were instructed to imagine completing each task with the help of an LLM CA (treatment group) or with a human interaction partner (control group). After reading each task scenario, participants completed a questionnaire to identify their task-dependent need for agency and experience.

We used established scales from the mind perception literature to measure the two dimensions of agency and experience (Bigman and Gray, 2018; Gray et al., 2007; Gray and Wegner, 2012; Yam et al., 2021). Accordingly, experience and agency were assessed using four items each, rated on a seven-point Likert scale (e.g., for experience: "...can feel pleasure"; for agency: "...can plan actions"). After participants evaluated all four task scenarios, we administered an attention-check question followed by questions to assess our control variables (gender, age, experience with LLM CAs, coding experience). All measurement items used in this study are listed in Appendix C.2.

3.3.2.3 Results

A factor analysis confirmed the reliability of the mind perception scales drawn from the literature, with Cronbach's α equal to 0.95 for experience, 0.75 for agency. The composite reliability (CR) and the average variance extracted (AVE) were greater than

the recommended 0.7 and 0.5 thresholds, respectively (Fornell and Larcker, 1981). The scale items, Cronbach's α, CR, and AVE values for each construct are detailed in Appendix C.3.1.

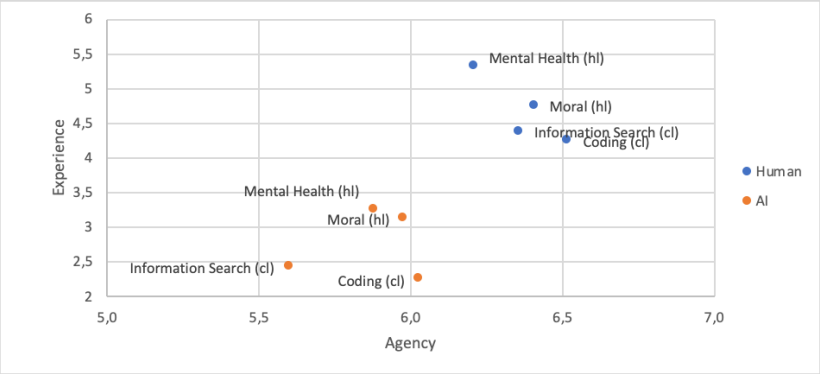
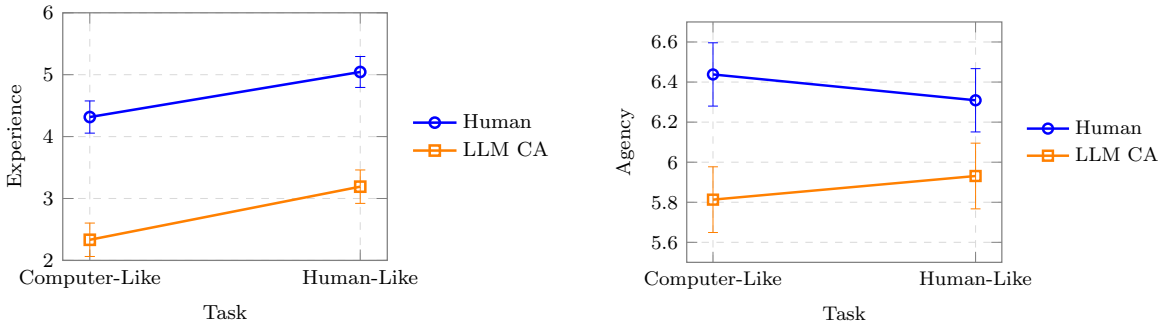


Figure 3.4: Descriptives for agency and experience.

Figure 3.4 illustrates the descriptives for agency and experience across the within-subjects (task type) and between-subjects (interaction partner) manipulations. To test whether required experience and agency differ depending on the task type, we conducted two mixed-model ANOVAs with interaction partner (LLM CA or human) as the between-subjects factor and task type (computer-like and human-like task) as the within-subject factor. In Appendix C.3.2, we additionally report the results of two ANCOVAs that included all control variables. Since the results were not affected by any controls, we focus on the interpretation of the mixed ANOVA findings.

A significant main effect of task type on experience was observed, $F(1, 151) = 65.42, p < 0.001, \eta 2 = 0.302$. Specifically, average experience was significantly higher for human-like tasks (M = 4.14, SD = 1.45) compared to computer-like tasks (M = 3.36, SD = 1.60). Furthermore, there was also a significant main effect of interaction partner on experience, $F(1, 151) = 133.41, p < 0.001, \eta 2 = 0.469$. Experience was significantly higher for human interaction partners (M = 4.68, SD = 0.1) than for LLM CAs (M = 2.76, SD = 0.1). However, there was no significant interaction effect between task type and interaction partner, $F(1, 151) = 0.51, p = 0.47, \eta 2 = 0.005$. Figure 3.5a illustrates the relationships of the factors. As hypothesized in Hypothesis 2a, independent of the type of interaction partner, the required experience is significantly higher when human-like tasks are performed.

A mixed ANOVA on agency revealed that there was no significant main effect of task type on agency, $F(1, 151) = 0.02, p = 0.9, \eta 2 = 0.00$. Average agency was not



(a) Effect of task type on experience needs.

(b) Effect of task type on agency needs.

Figure 3.5: Required experience and agency by task type and interaction partner.

significantly different for human-like tasks (M = 6.13, SD = 0.73) and computer-like tasks (M = 6.14, SD = 0.79). However, there was a significant main effect of interaction partner on agency $F(1, 151) = 22.43, p < 0.001, \eta 2 = 0.13$. Agency was significantly higher for human interaction partners (M = 6.37, SD = 0.05) than for LLM CAs (M = 5.87, SD = 0.07). Additionally, there was also a significant interaction effect between task type and interaction partner $F(1, 151) = 7.05, p = 0.009, \eta 2 = 0.05$, as illustrated in Figure 3.5b. A paired t-test on the subsample with between-subject factor LLM CA was not significant, $t(73) = -1.6, p = 0.11$ ($M_{cl} = 5.81, SD_{cl} = 0.09$; $M_{hl} = 5.93, SD_{hl} = 0.09$). A paired t-test on the subsample with between-subject factor human interaction partner was significant, $t(78) = 2.23, p = 0.03$ ($M_{cl} = 6.44, SD_{cl} = 0.07$; $M_{hl} = 6.31, SD_{hl} = 0.07$). Task type had no main effect on agency, indicating that the type of task (human-like or computer-like) did not have a significant impact on the required agency. Thus, Hypothesis 2b is not supported.

3.3.2.4 Discussion

Study 2 provides important insights to understand why individuals evaluate anthropomorphic design differently across task types. We reveal that individuals seek significantly more experience in an interaction partner when asked to perform a human-like task with its help, supporting Hypothesis 2a. This finding suggests that anthropomorphic designs that signal emotional capabilities correspond to users' need for experience in this task context. Contrary to Hypothesis 2b, we found no significant effect on users' need for agency. Instead, the level of required agency is independent of the type of task that users seek to perform with a LLM CA. Nevertheless, it is important to note that in accordance with mind perception literature (Gray et al., 2007; Gray and Wegner, 2012; Xu and Sar, 2018; Yam et al., 2021), both the required

agency and experience are significantly higher when performing tasks with a human interaction partner compared to a LLM CA. This finding indicates that people do not need or desire a LLM CA to possess the same level of agency and experience as a human. Instead, task-dependent needs shape their expectations regarding the experience dimension of anthropomorphism.

Building upon our understanding of task-dependent preferences and the need for anthropomorphism across different task types, our focus now shifts to LLM CAs' behavior in real-world use. In Study 3, we investigate whether LLM CAs autonomously adjust their anthropomorphic behavior across human- and computer-like tasks and how users evaluate such adjustment depending on task type. This allows us to assess whether real-world LLM CA behavior aligns with the preferences and needs identified in Study 2.

3.3.3 Study 3: LLM CA Behavior and User Evaluations

In Study 3, we aim to analyze whether LLM CAs autonomously adapt their anthropomorphic behavior as suggested in Hypotheses 3a and 3b, and to evaluate users' perceptions and evaluations of this anthropomorphic adaptation (H4, H4a, H4b, H5, H5a, H5b).[4]

Participants were recruited from three classes in business informatics (bachelor's and master's levels) at two universities. Since our experimental computer-like task involved coding-related tasks, we deliberately chose this sampling approach to ensure that participants possessed a basic understanding of such tasks. In total, 204 students participated in the experimental study. Of these, we excluded four participants due to server issues, 10 participants who failed the language check, and 11 participants who failed the attention check, which was a randomly placed question asking participants to select "undecided". The sample size underlying our data analysis was therefore $N = 179$. The average age of the participants was 20.21 (SD = 4.23), and 42 (23.5%) participants were female. The experimental study took place during a scheduled lecture session to ensure that participants had sufficient time and were motivated to complete the study. On average, the study took 74.88 minutes (SD = 14.16) to be completed. As an additional incentive to ensure ongoing motivation, every eighth

[4] Preregistered at AsPredicted #146851, ethics certificate at German Association for Experimental Economic Research e.V. #IdxreJIE

person won €50 for complete participation (random draw) and the five best responses also won €50 each. The best responses were identified after the sessions by two researchers who assessed the quality of task completion for all experimental tasks.

3.3.3.1 ChatGPT Integration

In this study, we asked participants to complete four different tasks (two computer-like tasks and two human-like tasks) with the assistance of ChatGPT. For this purpose, we integrated GPT-3.5-turbo through an iframe into our survey tool. This setup allowed participants to converse with ChatGPT in an integrated chat window without leaving the experimental platform, enabling us to collect all chat data for analysis. We consciously chose not to manipulate any behavior of ChatGPT as we aimed to analyze its autonomous adaptation to different task contexts. For each task, a new session with ChatGPT was initiated, ensuring that there was no information from previous sessions available or stored in memory. To encourage participants to engage in prolonged interactions with ChatGPT, they were required to use at least six prompts to complete each task. On average, each participant sent 7.53 prompts per task. In total, we analyzed chat data based on 5,394 user prompts.

Due to our focus on prolonged interactions to complete tasks with ChatGPT, we utilized GPT-3.5-turbo with a maximum token limit of 16,385, as the token limit for GPT-4 at the time of this study was 8,192 tokens. This ensured that ChatGPT retained context for each task. Retaining context within a session increases the total number of tokens handled by the model. Each prompt from previous interactions adds to the overall context considered by the AI model when generating responses to subsequent prompts within the same session. Consequently, the total number of tokens processed by the model accumulates with each new prompt. To maintain context retention during task completion with ChatGPT, we opted to implement GPT-3.5-turbo.

3.3.3.2 Procedure and Measures

Following a general briefing on the experimental procedure, including a short introduction to LLM CAs, participants were instructed to complete four tasks using the integrated version of ChatGPT. The order of the tasks was randomized. While a general description of the task context was provided within the experimental online tool, specific task instructions for interacting with ChatGPT were given to

each participant as a printed handout. This approach ensured that participants used their own words when completing the tasks. Consistent with Study 2, the two computer-like tasks involved coding-related and information search activities, while the two human-like tasks comprised a mental health discussion and a moral discussion. Detailed descriptions of each task can be found in Appendix C.1.3.

Following the completion of each task, participants were asked to complete a questionnaire to measure agency, experience, trust, eeriness, and human likeness for that particular task. In addition, participants were prompted to write an open-text reflection on their user experience regarding both the computer-like and human-like tasks with ChatGPT. They were explicitly instructed to consider what they liked or felt was lacking in the interaction. Finally, the study concluded with a questionnaire regarding demographics and previous experience with LLM CAs and coding in general.

We utilized the same established scales to measure agency, experience, and task human likeness as in Study 2. Additionally, we used established scales to measure trust (McKnight and Choudhury, 2002) and the control variable eeriness (Gray and Wegner, 2012; Ho and MacDorman, 2010). All measurement items deployed in this study are listed in Appendix C.2.

3.3.3.3 Text Analysis

We used the Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2001), a text analysis software, to examine the extent to which the task-dependent responses of ChatGPT reflected anthropomorphic behavior in terms of agency and experience. LIWC is a state-of-the-art psycholinguistic software that has more recently been applied to text analysis in IS studies (Gnewuch et al., 2024; Kumar et al., 2022). LIWC uses default psycholinguistic dictionaries to analyze the linguistic features of text data by calculating percentages of words belonging to different categories. In our analysis of ChatGPT's chat data, we operationalized experience and agency using established psycholinguistic categories of LIWC that reflect emotional and analytical language cues. Detailed analysis can be found in Appendix C.3.4.

For each task, we analyzed the dialogue text produced by ChatGPT with respect to experience and agency. Accordingly, the resulting scores for experience and agency isolate the anthropomorphic cues employed by ChatGPT.

3.3.3.4 Results

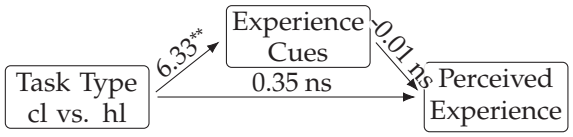
We conducted a factor analysis to assess the reliability of the measured constructs. Due to a high correlation between one item of our agency scale and the competence-based trust dimension, we had to exclude this item from our analysis. Additionally, one item of the benevolence-based trust dimension was removed due to factor loadings smaller than 0.5. The scale items, along with Cronbach's α, CR, and AVE values for each construct, are detailed in Appendix C.3.1.

To verify the effectiveness of our task manipulation, we conducted a paired t-test on the task human-likeness ratings that range from 1 (very computer-like) to 7 (very human-like) which is the same as in Study 1. The t-test indicated a significant difference, $t(178) = 15.71, p < 0.001$, confirming that participants perceived the human-like tasks as more human-like (M = 5.33, SD = 0.11) while they perceived the computer-like task as more computer-like (M = 3.32, SD = 0.09).

To test whether ChatGPT exhibits task-dependent autonomous anthropomorphic adaptation in its conversational behavior (H3a, H3b), we conducted two paired t-tests comparing ChatGPT's use of experience and agency in the interactions across task types. The first t-test indicated a significant difference, $t(178) = 43.89, p < 0.001$, confirming that in the human-like task dialogues, more experience cues (M = 12.6, SD = 0.19) were used than in the computer-like task dialogues (M = 3.19, SD = 0.11), thus supporting Hypothesis 3a. The second t-test indicated a significant difference, $t(178) = -24.10, p < 0.001$, demonstrating that, contrary to Hypothesis 3b, in the human-like tasks dialogues, fewer agency cues (M = 70.42, SD = 0.79) were used than in the computer-like task dialogues (M = 89.17, SD = 0.39). Thus, Hypothesis 3b cannot be confirmed; instead, we find a significant difference in the opposite direction.

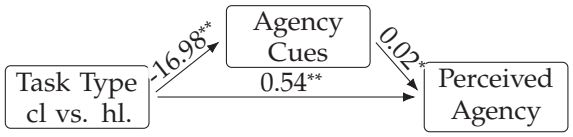
Taken together, these findings indicate that the anthropomorphic style of the conversation varies by task type, with more experience cues in human-like tasks and, unexpectedly, fewer agency cues. We further investigated whether this task effect carries over to users' mind perception, accounting for potential indirect effects via experience and agency cues (H4a, H4b). To do so, we conducted mediation analyses using the SPSS MEMORE macro (Montoya and Hayes, 2017), which allows estimation of mediation models for two-instance within-subjects design using a bootstrapping approach (Judd et al., 2001).

The first analysis examined whether experience cues in ChatGPT's dialogue text mediated the effect of task type on perceived experience, using 5,000 bootstrap samples (Figure 3.6a). Although task type significantly affected ChatGPT's use of experience cues (path a), the indirect effect was not significant ($ab = -0.05$, SE = 0.40, 95% CI $[-0.92,\ 0.62]$). The total effect of task type on perceived experience was positive and significant ($\beta = 0.30$, SE = 0.06, $t(178) = 4.85$, $p < 0.001$). Thus, although task type affects perceived experience, we find no support for Hypothesis 4a. The increased use of experience cues (H3a) did not mediate the effect of task type on perceived experience. When accounting for the indirect path via experience cues, task type had no direct effect on perceived experience ($\beta = 0.35$, SE = 0.31, $t(176) = 1.14$, $p = 0.26$).



- Experience: a*b path coefficient = -0.05, *SE* = 0.40, 95%CI = [-0.92, 0.62]
- Total Effect: 0.30 **

(a) Effect of Task Type on Experience Perception.



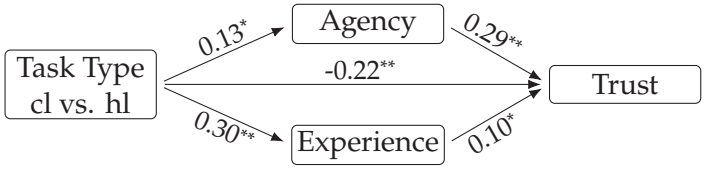
- Agency: a*b path coefficient = -0.42*, *SE* = 0.21, 95%CI = [-1.01, -0.19]
- Total Effect: 0.13 *

(b) Effect of Task Type on Agency Perception.

Figure 3.6: Effect of computer-like (cl) vs. human-like (hl) task type on mind perception.

The second analysis examined the effect of task type on perceived agency via ChatGPT's use of agency cues, based on 5,000 bootstrap samples (Figure 3.6b). Results revealed that the indirect effect through agency cues was negative and significant ($ab = -0.42$, SE = 0.21, 95% CI $[-1.02,\ -0.18]$). Thus, there is a significant indirect effect of task type on perceived agency through agency cues, but this effect is in the *opposite* direction to that hypothesized (H4b). Specifically, agency cues had a positive association with perceived agency (path b: $b = 0.025$, SE = 0.008, $t(176) = 3.24$, $p = 0.0014$), while the reduced use of agency cues in human-like tasks produced an overall negative indirect effect. Yet, the mediation is only partial because accounting for the indirect path via agency cues, task type still had a positive direct effect on perceived agency ($\beta = 0.54$, SE = 0.14, $t(176) = 3.81$, $p = 0.0002$). Accordingly, while perceived agency decreases in human-like tasks because of reduced usage of agency cues, human-like tasks retain a direct positive effect on perceived agency that is not accounted for by agency cues. This direct effect is substantial and offsets the negative indirect effect, such that the total effect of task type on perceived agency remained positive and significant ($\beta = 0.125$, SE = 0.06, $t(178) = 2.01$, $p = 0.046$).

Having examined how task type shapes anthropomorphic interaction style and perceived anthropomorphism, we move our focus to users' evaluations, specifically trust in the LLM CA (H5, H5a, H5b). We analyzed the effect of task type on user trust through agency and experience, using 5,000 bootstrapped samples. Results revealed that the indirect effect through agency was positive and significant ($ab = 0.036$, SE = 0.02, 95% CI [0.001, 0.07]), which supports Hypothesis 5b. The indirect effect through experience was positive and marginally significant ($ab = 0.029$, SE = 0.016, 95% CI [0.0006, 0.076]), which partially supports Hypothesis 5a. The direct effect of task type on trust was negative and significant ($\beta = -0.22$, SE = 0.04, $t(174) = -5.53$, $p < 0.001$). Additionally, the total effect of task type on trust was negative and significant ($\beta = -0.16$, SE = 0.04, $t(178) = -3.71$, $p < 0.001$), supporting Hypothesis 5. Figure 3.7 illustrates these results.



- Agency: a*b path coefficient = 0.04, *SE* = 0.02, 95 % CI = [0.001, 0.076]
- Experience: a*b path coefficient = 0.03, *SE* = 0.02, 95 % CI = [-0.001, 0.062]
- Total Effect: -0.16 **

Figure 3.7: Effect of computer-like (cl) vs. human-like (hl) task type on user trust.

In summary, we find that people trust LLM CAs more to perform computer-like tasks than human-like tasks (H5). For human-like tasks, however, this negative effect is attenuated through perceptions of agency (H5b) and experience (H5a). To further unpack these effects, we estimated the same mediation model for the two trust sub-dimensions (competence-based trust and goodwill-based trust). The additional mediation models are provided in Appendix C.3.3. They reveal that perceived agency, but not perceived experience, has a significant positive indirect effect on competence-based trust, whereas both perceived agency and perceived experience have significant positive indirect effects on goodwill-based trust.

In addition to the reported analyses, we conducted further analyses with regard to the control variables gender, experience, dispositional anthropomorphism, and eeriness. These analyses did not affect the interpretation of our results (Figure C.2).

3.3.3.5 Discussion

Study 3 presents several important findings. First, our analysis reveals that the LLM CA's anthropomorphic style adaptation is significantly influenced by the type of task being performed. Specifically, we observe that in contexts resembling human tasks, the use of experience cues is significantly higher compared to contexts resembling computer tasks. Contrary to expectations, the use of agency cues is significantly lower in human-like task contexts than in computer-like task contexts. This suggests that the language style of ChatGPT becomes less analytical in human-like task contexts. This outcome contradicts H3b and prior mind perception literature, which would predict that human-like tasks elicit more agency cues than computer-like tasks. Our findings suggest that the training data for computer-like tasks contain more analytical language than those for human-like tasks. However, other aspects of agency such as forming opinions and preferences may not be adequately captured by the analytical language score. Further investigation using qualitative data offers the potential to verify this.

3.4 General Discussion

Table 3.3 summarizes our empirical findings. We demonstrate that users choose anthropomorphized LLM CAs only for human-like tasks (Study 1), in particular they have a higher need for experience in human-like compared to computer-like tasks (Study 2). In contrast, the users' needs for agency remain consistent across task-types, challenging expectations from mind perception literature (Gray et al., 2007; Gray and Wegner, 2012; Xu and Sar, 2018; Yam et al., 2021). We further analyzed how the behavior of LLM CAs complements these user needs in real-world use scenarios (Study 3). We discovered that task-type-specific automatic adaptations align to a certain extent: LLM CAs use more experience cues when solving human-like tasks but fewer agency cues. Notably, only agency cues have a significant positive effect on user perception. User trust in LLM CAs is generally higher for computer-like tasks than for human-like tasks, although this effect is mitigated by the increased mind perception for LLM CAs in human-like tasks.

Table 3.3: Overview of empirical findings

Study	Hypothesis	Result
Study 1	**H1**: Task-dependent choice for anthropomorphism in CAs	**Confirmed**: anthropomorphic CAs chosen for hl tasks
Study 2	**H2**: Need for agency and experience	*Experience*: **confirmed**, higher need for hl tasks *Agency*: **not confirmed**, no difference between hl and cl tasks
Study 3	**H3**: Task-dependent anthropomorphic design adaptation	*Experience*: **confirmed**, in hl tasks more usage; *Agency*: **contradicted**, in hl tasks less usage
	H4: Task-dependent anthropomorphism perception mediated through cues	*Experience*: **not confirmed**, no effect on perception; *Agency*: **contradicted**, positive effect of cues on perceived agency, but less usage of agency cues in hl tasks
	H5: Task-dependent user trust	*Task Dependent User Trust*: **confirmed**, more trust for cl tasks; *Experience & Agency*: **confirmed**, higher perception in hl tasks mitigates lower trust

Note: hl = human-like, cl = computer-like

3.4.1 Theoretical Contributions

Our three studies offer new insights into task-dependent differences in the dimensions of mind perception — agency and experience. While prior work in mind perception has primarily focused on the nature of the entity being evaluated, comparing the perception of machines versus human decision makers (Epley et al., 2007; Gray et al., 2007, 2012; Lee et al., 2020; Yam et al., 2021), few studies have highlighted context-dependent preferences for experience and agency in AI (Appel et al., 2020; Wiese et al., 2022). In contrast, we have systematically examined task-dependent mind perception in interactions with LLM CAs, focusing both on user perceptions and how these agents autonomously adapt their anthropomorphic behavior across a range of human-like and computer-like tasks. Our results reveal that task type is a significant antecedent of mind perception, independent of the anthropomorphic design features of the agent. Human-like tasks have a positive direct effect on users' agency perception. The effect on experience appears more complex and is likely to be mediated through additional factors. This result challenges the perception that dialogue behavior throughout the interaction alone determines whether a CA is

perceived as human-like. It suggests that task types should become an integral part of mind perception theory.

Furthermore, our study extends IS research on human–CA interaction and trust in technology. The context of the interaction is pivotal in shaping user trust in artificial intelligence (Castelo et al., 2019; Glikson and Woolley, 2020; Kaplan et al., 2023). Extant research has shown that anthropomorphism impacts user trust and that different aspects of anthropomorphism — like experience or warmth and agency or competency — affect trust in distinct ways (Blut et al., 2021; Cheng et al., 2022; Waytz et al., 2014). Our findings provide a deeper understanding by analyzing both of these factors within one cohesive theoretical framework of mind perception theory. In particular in human-like tasks compared to computer-like tasks, users prefer more human-likeness in CAs specifically with regard to emotional capability (i.e., experience) but not agency. We hypothesized that discrepancies between these task-specific user expectations and LLM CA behavior result in poorer user evaluations and diminished trust, whereas fulfilling these needs can promote development of trust. Our findings confirm this relationship: in human-like tasks, where user trust is initially lower, a stronger perception of anthropomorphism across both dimensions effectively counteracts this trust deficit.

Our framework allows for a systematic analysis of ChatGPT's behavior as a prominent LLM CA and the fastest-growing consumer application (Hu, 2023; Reuters, 2023). We find that task-dependent user needs for experience and agency are reflected to some extent this real-world application. The LLM CA automatically adapts its anthropomorphic features based on the task at hand. However, this automatic adaptation is neither optimally aligned to the user's preferences nor does it have a positive impact on mind perception. We postulated that the increased mind perception that we found in human-like tasks would largely be explained by the usage of more experience and agency cues. Considering that LLM CAs digest vast amounts of human-produced dialogue data on the internet, we hypothesized that, in alignment with mind perception theory, they employ both agency and experience cues more frequently in human-like tasks — as humans do. Yet, our results indicate that this is not the case. Instead, CAs reduce the usage of agency cues in human-like tasks. This may be a crucial limitation when trying to build trust, as agency cues could have a significant positive effect on agency perception, which in turn has a positive effect on user trust. Furthermore, while LLM CAs do use more experience cues in human-like tasks, these cues do not have an effect on users' perception of

experience. Therefore, it is important to identify experience cues that address user expectations more effectively.

An important observation in the anthropomorphism literature is that anthropomorphic features are often manipulated deliberately through specific design cues (Feine et al., 2019; Gnewuch et al., 2024; Schanke et al., 2021; Schuetzler et al., 2021; Seeger et al., 2021). Our research demonstrates, however, that LLMs adjust their behavior automatically depending on the context, which can influence perceptions of anthropomorphism beyond deliberate manipulations. It is therefore crucial for researchers to be aware of such automatic processes and to account for them in their experimental designs.

Finally, our research contributes to the rising field of algorithm aversion. Some work in this field differentiates between objective and subjective tasks (Castelo et al., 2019; Dietvorst et al., 2015, 2016; Jussupow et al., 2024). Reconsidering this categorization may be beneficial to capture the full nature of the tasks performed by LLM CAs, which often combine objective and subjective elements. Human-like tasks typically emphasize subjective aspects requiring emotional abilities (experience) but also involve objective components requiring cognitive skills (agency) to varying degrees. In contrast, computer-like tasks mostly contain objective elements that require agency. This distinction clarifies task effects on experience and agency perception, allowing us to pinpoint more precisely when and why AI aversion arises. Our study raises the prospect of "spillover" effects. A distinct feature of LLM CAs is their versatility, allowing them to perform both computer-like and human-like tasks. This may enable positive experiences in computer-like tasks — where trust is already higher — to mitigate aversion in human-like tasks (Castelo et al., 2019). In other words, trust and familiarity built with a CA in computer-like tasks can reduce aversion to using it in human-like tasks, fostering broader acceptance over time. Effective automatic adaptation by CAs supports this process by addressing evolving task-dependent needs and sustaining trust across applications.

3.4.2 Practical and Societal Implications

Our findings carry important implications for both the design and societal understanding of CAs. First, we demonstrate that users do not require CAs to exhibit the same level of agency or experience as human interaction partners. This suggests that striving to mimic human mental capacities as closely as possible in CAs is

not always the best approach. Tying into the concept of the uncanny valley, the discomfort that too much focus on anthropomorphism can evoke seems to provide limitations (Gray and Wegner, 2012; Waytz et al., 2014). However, with advances in robotics and immersive technologies dissolving the boundaries between physical and digital presence, societal perceptions of the uncanny may evolve. Furthermore, if, as self-humanization theory suggests (Haslam, 2006), humans anchor their identity in traits like morality and autonomy, then the design of such systems must not only consider functionality, but also calls for societal discourse on how closely we want machines to mirror such traits.

From a practical perspective, our results provide guidance for improving the design of CAs, especially in contexts where trust is a key consideration. To mitigate lower trust in CAs for human-like tasks, designers should focus on fostering trust by stimulating users' mind perception of the CA. Users state a higher need for experience in human-like tasks, but the impact of the CA's experience cues is limited.

Our findings on the adaptation of experience and agency cues suggest that there is substantial room for design improvements in current LLM CAs. Advanced technologies, such as virtual reality (VR) or three-dimensional display technologies (3D displays), may enable more immersive and impactful experience cues by incorporating nonverbal signals, such as gestures, facial expressions, and shared gazes, that can better convey emotional capabilities (Cassell et al., 1999; Ghazali et al., 2018). Embodying conversational agents with human-like avatars in 3D environments has been demonstrated to deepen connections with users, enriching interactions through additional social and nonverbal cues (Barlow et al., 2004; Holzwarth et al., 2006). Similarly, audio output leveraging tonal variations offers another avenue for enhancing the emotional depth of interactions (Heßler et al., 2023). However, implementing these advanced cues may require greater technological effort compared to agency cues, which are more readily conveyed through language and 2D graphics.

The limited effectiveness of observed experience cues, coupled with the significant positive impact of agency cues on mind perception and consequently user trust, underscores the need to maintain strong agency capabilities regardless of the task context. Automatic adaptation to domain-specific data that results in reduced agency for human-like tasks may backfire and instead require purposeful design interventions to ensure that agency remains high. Issues around misdirected automatic adaptation of LLM CAs could be amplified by AI systems potentially learning from each other's output, leading to a self-reinforcing spiral of AI becoming more human-like in

tone with regards to experience (Alemohammad et al., 2023; Briesch et al., 2024; Martínez et al., 2023). In contrast, recent models (e.g., GPT-4o) capable of delivering explanations alongside their solutions (OpenAI, 2024; Wei et al., 2022) may heighten perceptions of anthropomorphism by mirroring human analytical behavior and signaling agency.

Practitioners should be aware that not only conscious design decisions — such as giving the CA a name, picture, or emoticons — determine anthropomorphism. LLM CAs now independently learn to use agency and experience cues from the vast amounts of interaction data they are trained on, thus moving beyond purely manual design choices. Such CAs adjust their behavior based on context, learning the very mechanisms that we attempt to investigate causally through experiments. In addition to training LLMs on the basis of human-generated content, tuning and alignment procedures further shape model behavior and can reinforce human-like interaction patterns. Recent studies demonstrate that alignment processes can lead to sycophantic and socially desirable responses, reflecting anthropomorphic traits (Sharma et al., 2023). Together, these processes likely contribute to the task-dependent emergence of agency and experience cues. The way in which they adapt raises critical implications for user adoption and acceptance in practical applications. Ideally, LLM CAs could perfectly adapt to user needs, but our findings suggest the opposite — particularly regarding the improper use of agency cues. This suggests that either current LLM CA capabilities to adapt are not yet advanced enough or that insights from literature, such as those regarding user needs, need to be explicitly integrated into these systems rather than relying solely on automatic adaptation.

In both society and research, there is growing concern about the blurring boundaries between human and machine perception (Bender et al., 2021; Susarla et al., 2023). This issue extends beyond design choices and includes the selection of tasks assigned to CAs, as demonstrated by our findings. Therefore, design constraints or enhancements are not solely responsible for determining how CAs are perceived. The increasing capacity of CAs to perform tasks traditionally associated with humans raises questions about how society navigates the potential overlap in decision-making roles between humans and machines.

3.4.3 Limitations and Future Research

Despite offering valuable insights into user preferences for anthropomorphism in CAs, our study bears some limitations that open avenues for future research.

First, while our study identifies user preferences for agency and experience cues in CAs, it remains unclear what drives these preferences. Whether our findings on lowered needs for agency and experience compared to a human interaction partner stem from the users' appraisal of what is actually technologically possible or their real ideal wish remains unanswered by our paper. The difference we observe may be due to individuals not wanting AI to reach the same level of capability as humans, preserving the notion of humans occupying a unique and special role. Alternatively, it could reflect users' underlying belief that AI is not yet capable of fully replicating human abilities, which influences the expectations individuals set for the technology. Future research is deemed necessary to distinguish these influences.

Second, our sample consisted of bachelor and master students from business informatics classes that presumably had more knowledge than the average population on the subject. Additionally, our study did not account for potential cultural effects, which could play a significant role in how users perceive agency and experience cues in CAs. Subsequent studies should aim to include participants from various backgrounds to explore user perceptions more comprehensively across different demographics and cultures.

Third, we analyzed agency and experience cues in the LLM CAs using LIWC, which is well-suited for assessing linguistic features but may have limitations in fully capturing the complexity of how users perceive these dimensions in text-based interactions. For example, LIWC focuses on predefined word categories, which might overlook subtle contextual or stylistic aspects that influence user perceptions. Future research could explore complementary methods, such as fine-tuned natural language processing models, to deeper analyze textual cues for agency and experience.

3.5 Conclusion

Our study contributes to the understanding of task-dependent anthropomorphic design in CAs by integrating mind perception theory into human–computer interaction research. By differentiating agency and experience as distinct dimensions, we

reveal how task contexts shape user expectations and trust. Our findings highlight a misalignment between user preferences and autonomous adaptations of CAs for human-like tasks. While individuals desire sustained high levels of agency coupled with increased experience, LLMs reduce agency cues and employ ineffective experience cues. As LLM CAs become increasingly integrated into diverse domains, our insights emphasize the need for deliberate, user-centered design to ensure these systems effectively support user trust and engagement. Our results contribute to a better understanding of how LLM CAs can navigate the complexities of human-like interaction more effectively, thereby refining mind perception theory and opening avenues for future work on more adaptive and context-sensitive design.

A Appendix for Chapter 1

A.1 Instructions

[Welcome and Instructions]
Welcome! Thank you for your participation in a behavioral economics study conducted by the Karlsruhe Institute of Technology (KIT), one of the largest research universities in Germany.

Please complete the study in a **quiet place where you will not be distracted**. Ideally, you should not take long breaks during the study, but rather complete it without interruption.

Please **DO NOT** use the back button on your browser while completing the survey.

The study takes approx. **5 to 10 minutes** to complete.

Important: Participants who have not read the instructions, or randomly marked answers may be disqualified from payment.
Comprehension questions are used to verify that the instructions have been read.

[Consent form]

[Sociodemographics (Office of Management and Budget (OMB), 2024; Statistisches Bundesamt (Destatis), 2022; U.S. Census Bureau, 2023, 2024)]

1. **What sex are you?**
2. **How old are you?**
3. **What is your ethnicity?** *[For US-sample]*
4. **What is the highest diploma/degree or level of school you have completed?**

[Decision Consequence]
****In RealCons Treatments****

Your decisions have real consequences!

As in all behavioral economic studies at KIT, **all the facts described in the study are true.**

At the end of the study, the computer randomly selects about one in ten participants.

The decisions made by these selected participants in the study are implemented exactly as described. Your decisions in this study are therefore not hypothetical, they can have real-world consequences.

Therefore, make your decision carefully.

****In HypoCons Treatments****

Hypothetical Decision Scenarios!

As in all behavioral economic studies at KIT, **all the facts described in the study are true.**
Your decisions in this study exclusively concern hypothetical scenarios.
Nevertheless, your decisions are essential to research. Therefore, please decide carefully.

On the following pages we present the work of two reputable charities. Please read the information carefully. You will need it in the further course of the study.

[Donation Information (AMF) — English Version]
****In RealCons, Gain Treatment****

Against Malaria Foundation

In the following, a donation of **5 dollars** to the **Against Malaria Foundation** will be made by us in your name.

With this donation, a **child can be saved from malaria, from which it might otherwise die.**

Fighting Malaria

- Each year, more than **600,000 people die from malaria.**
- **More than 70% of them are children under the age of 5.**
- **Malaria can be prevented:** Anti-malaria nets are an effective form of protection.

[Image of child receiving help]

Against Malaria Foundation

... distributes long-lasting insecticidal nets (LLINs) in malaria endemic countries.

Recipients of nets hang and sleep under them so they are not bitten by malaria-carrying mosquitoes.

- **Nets save lives!**
- **Providing one net costs ca. $5.**

****In Loss Treatment****

Against Malaria Foundation

There is a donation voucher in your name worth **5 dollars** to the **Against Malaria Foundation**. Upon completion of this study, we will redeem this donation voucher on your behalf, and the corresponding amount will be donated to the organization in question.

With this donation, a **child can be saved from malaria, from which it might otherwise die.**

Subsequent Information identical as above

****In HypoCons Treatment****

Against Malaria Foundation

Donations to the Against Malaria Foundation protect **children from malaria that could otherwise kill them.**

Subsequent Information identical as above

[Donation Information (HKI)- English Version]

****In RealCons, Gain Treatment****

Helen Keller International

You can also actively intervene and donate **7 dollars** to **Helen Keller International** instead.

With this donation, **seven children** will **receive vitamin A who might otherwise die from a deficiency**.

Fighting Vitamin A Deficiency

- Vitamin A deficiency makes children susceptible to infections and can lead to death.
- Each year, more than **200,000 children's deaths** are attributed to vitamin A deficiency.
- Providing vitamin A supplements **saves children's lives!**

[Image of child receiving help]

Helen Keller International

... distributes long-lasting vitamin A supplements.

In areas where vitamin A deficiency is a public health problem, children aged 6 months to 5 years receive a high dose of vitamin A.

- **Vitamin A saves lives!**
- **Vitamin A for a child under 5 costs ca. $1.**

****In Loss Treatment****

Helen Keller International

Additionally, there is a donation voucher worth **7 dollars** to **Helen Keller International**.

This donation will provide Vitamin A to 7 children who might otherwise be at risk of dying from a deficiency.

Subsequent Information identical as above

****In HypoCons Treatment****

Helen Keller International

With this donation, **seven children will receive vitamin A who might otherwise die from a deficiency.**

Subsequent Information identical as above

Please note:
According to the independent initiative GiveWell, which evaluates charities, both programs are among the top donation opportunities.
Selected are donation organizations that are particularly efficient, whose impact is particularly well documented, that work particularly transparently, that require additional donations and that meet other criteria.

[Donation and Delegation Option]

****RealCons****

Your donation

On the following screens you can influence which donation will be made.
Alternatively, you can delegate to ***[another participant in this study/an artificial intelligence (AI)]***.
Then you will not be confronted with the situation and you will also not be informed which donation will be made in the end.
Instead, *[another participant will be randomly drawn and their behavior will be implemented./an AI will then determine which donation is made.]*

****HypoCons****

A donation

Imagine you could decide **which of these two charities should receive a donation.**

Alternatively, you could delegate to ***[another participant in this study/an artificial intelligence (AI)]***.
Then you would not be confronted with the situation any further and would also not be informed which donation would have been made in the end.
Instead, *[another participant would then be randomly selected and their behavior implemented./an artificial intelligence would then determine which donation to make.]*

[Comprehension questions]
Comprehension question on basic instructions. Participants that answered incorrectly more than twice were disqualified.

[Decision – RealCons]
****Gain Treatments****
Your donation
On the next screen, **20 seconds** will count down

- If you do **nothing**, the donation to the **Against Malaria Foundation (option A)** will be made.
- You can also **actively intervene** and donate to **Helen Keller International (option B)** instead

---

****Loss Treatments****
Your donation
On the next screen, **20 seconds** will count down

- If you do **nothing**, the donation voucher to **Helen Keller International (option B)** will be **destroyed**.
- You can also **actively intervene** and **destroy** the donation voucher to the **Against Malaria Foundation (option A)** instead.

If you prefer, you can also delegate to *[another participant in this study/an artificial intelligence (AI)]* instead.
Then you will not be confronted with the situation and you will also **not be informed** which donation will be made in the end.

Instead, another participant is randomly drawn and their behavior is implemented/Instead, an AI will then determine which donation is made.

If you want to delegate to *[another participant/the AI]*, click the button.

Otherwise, click "Next" to proceed to the donation options.

****If "Next" (No Delegation), Gain****
Which donation should be made?

Option A:	**Option B:**
$5 to the **Against Malaria Foundation**.	**$7** to **Helen Keller International**.

Remaining time: *20s*

****If "Next" (No Delegation), Loss****
Which donation should be destroyed?

Option A:	**Option B:**
Destroy **$7**-donation voucher to **Helen Keller International**.	Destroy **$5**-donation voucher to **Against Malaria Foundation**.

Remaining time: *20s*

****If Button (Delegation)****
You have delegated the decision to *[another participant/the AI]* in this study.

[Delegation Decision – HypoCons]
****Human Treatment****
How would you decide? *[random order]*
In the situation described, would you **delegate to another participant** or **make the decision yourself**?
Please note that this decision is purely **hypothetical** and **will not be implemented** over the course of this study.

- I would **decide myself** which of the two charities would receive the donation.
- I would **delegate** the decision about which of the two charities receives the donation **to another participant**.

****AI Treatment****
How would you decide? *[random order]*
In the situation described, would you **delegate to an artificial intelligence** or **make the decision yourself**?
Please note that this decision is purely **hypothetical** and **will not be implemented** over the course of this study.

- I would **decide myself** which of the two charities would receive the donation.
- I would **delegate** the decision about which of the two charities receives the donation **to an artificial intelligence**.

[Follow-Up Questions]
[Decision Justification for RealCons/HypoCons]
Why *[did/would]* you *[/not]* delegate the decision? Multiple answers possible

****If Delegated****

- The *[other participant/AI] [will/would]* make a better decision.
- The decision *[was/would be]* too difficult or I *[didn't/wouldn't]* have a clear preference.
- I *[had/would have]* too little information about the decision.
- I *[wanted/would want]* to hand over responsibility for the decision.
- I *[wanted/would want]* to keep it as simple as possible and not have to deal with the decision any further.
- Other (please specify): _____

Corresponding opposite reasons provided if the decision was not delegated.

[Responsibility for RealCons/HypoCons]
Please indicate to what extent you agree or disagree with each of the following statements.
5-point scale: 1 = strongly disagree, 5 = strongly agree

- I **would like to be fully responsible** for the decision, whatever the outcome.

- I *[feel/would feel] responsible* for the outcome of this decision.
- I *have/would have* a moral obligation to make such a decision.

How confident are you that you have made the right decision about whether to delegate or make the decision to donate yourself?
How difficult do you find the decision between the two donation options?
How important are the following criteria to you when making a donation?
5-point scales

- **Cost-effectiveness** of the donation, i.e. how much donation money is needed to save a life.
- **Number of people affected**, i.e., how many people are fatally threatened by the issue being addressed (e.g., disease or hunger).

[Rating of AI's capability for moral decision-making]
The following questions are about your assessment of the capabilities of artificial intelligence (AI).
5-point scale: 1 = strongly disagree, 5 = strongly agree

- In a situation as described in this study, an artificial intelligence (AI) can make a better decision between two donations than I can.
- I have full confidence that an AI can make a high-quality decision between two donations in a situation like this.
- AI can make good moral decisions.

[Mind Perception Scale (Bigman and Gray, 2018)]
To what extent do you think an AI can/is ...
5-point scale: 1 = Not at all, 5 = Extremely

****Experience****
- ... sensitive to pain?
- ... experience happiness?
- ... experience fear?
- ... experience compassion?
- ... experience empathy?
- ... experience guilt?

****Agency****
- ... communicate with others?
- ... able of thinking?
- ... plans its actions?
- ... is intelligent?
- ... has foresight?
- ... is able to think things through?

A.2 Robustness Check: No Effect of Burden-Disclaimer

In the original study, participants in human treatments were informed that in case of delegation *"another participant will be randomly drawn and their behavior will be implemented"* in the instructions and on the decision screen (see A.1). Our intention was to convey that no additional decision burden would be imposed on the selected delegate. However, the phrasing may have been perceived as ambiguous.

To address this concern, we conducted an additional study (U.S. representative sample via Prolific)[1]. We employed the two real-consequence treatments from the main studies (Human delegate, AI delegate) and added a *No-Burden Human* treatment. This treatment was identical to the original Human delegate condition, but included the following disclaimer both in the instructions as well as on the decision screen, visually highlighted in red font: *"The selected participant will be drawn from those* ***who have already made a decision****. Their choice will be implemented* ***without requiring any further action or awareness on their part****"*. Since live filtering was not technically possible, we applied the same exclusion criteria as in the main studies retrospectively: participants who failed the comprehension question, failed the attention check, or completed the study too quickly (Leiner, 2019) were excluded. After applying these criteria, the final sample consisted of $N = 592$ participants.

As shown in Figure A.1, delegation rates in the No-Burden Human condition (11.6%) were nearly identical to the Standard Human condition (11%). A chi-square test confirmed that this difference was not statistically significant ($p = 0.846$). Delegation to AI (18.6%) remained substantially higher, indicating that the increased delegation to AI we find in our studies cannot be explained by concerns about imposing a burden on a human delegate.

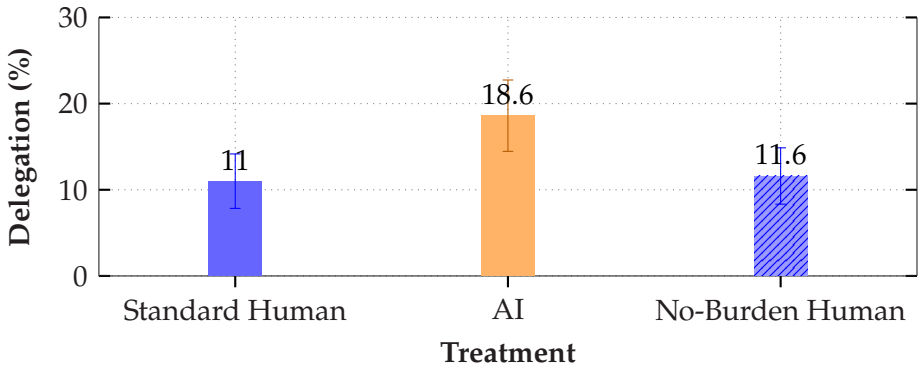


Figure A.1: Delegation rates across Standard Human, No-Burden Human, and AI treatment. Explicitly clarifying that the human delegate would not bear any additional burden did not affect delegation rates. ($\chi^2(1, N = 398) = 0.0377$, $p = 0.846$).

[1] Preregistration at AsPredicted.org: https://aspredicted.org/79sb-jb38.pdf

A.3 Additional Results & Statistical Analyses

Table A.1: Logistic regression results for delegation behavior by delegate type and framing.

	OR	Std. Error
Delegate (AI vs. Human)	4.551***	1.763
Framing (Loss vs. Gain)	2.077	0.875
Delegate × Framing	0.433	0.216
Constant	0.047***	0.016

Note: LR $\chi^2(3) = 24.01$, Pseudo $R^2 = 0.0412$, $N = 800$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table A.2: Logistic regression results for delegation behavior by delegate and decision impact.

	OR	Std. Error
Delegate (AI vs. Human)	2.233***	0.226
Decision Impact (HypoCons vs. RealCons)	1.375**	0.145
Delegate × Decision Impact	0.601***	0.084
Constant	0.190***	0.015

Note: LR $\chi^2(3) = 75.15$, $p < 0.001$, Pseudo $R^2 = 0.0144$, $N = 4,839$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table A.3: Regression analysis of responsibility ratings on a 5-point scale by delegation decision and delegate type.

	β	Std. Error
Delegate (AI vs. Human)	0.163***	0.027
Delegation Decision (Yes vs. No)	-0.755***	0.043
Interaction (AI × Delegation = Yes)	-0.241***	0.057
Constant	3.966***	0.018

Note: $F(3, 4839) = 342.06$, $p < 0.001$, $R^2 = 0.1750$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table A.4: Regression analysis of capability ratings on a 5-point scale as a function of delegate type and decision impact.

	β	Std. Error
Delegate (AI vs. Human)	0.331***	0.042
Decision Impact (HypoCons vs. RealCons)	-0.008	0.042
Interaction (AI × HypoCons)	-0.163**	0.060
Constant	2.475***	0.030

Note: $F(3, 4839) = 28.56$, $p < 0.001$, $R^2 = 0.0174$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table A.5: Regression analysis of capability ratings on a 5-point scale as a function of delegate type and delegation behavior.

	β	Std. Error
Delegate (AI vs. Human)	0.012	0.032
Delegation Behavior (Yes vs. No)	0.348***	0.051
Interaction (AI × Delegation = Yes)	0.732***	0.068
Constant	2.407***	0.022

Note: $F(3, 4839) = 238.19$, $p < 0.001$, $R^2 = 0.1287$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table A.6: Open-text responses from delegating participants by delegate type (Study 1).

Reason for Delegation	Delegation to Human (%)	Delegation to AI (%)
Better decision	**3.70%**	**44.12%**
Decision difficult or uncertain / no clear preference	51.85%	32.35%
Too little information	18.52%	10.29%
Hand over responsibility	**44.44%**	**19.12%**

Table A.7: Logistic regression results for reasons to delegate.

Reason	Delegate	Decision Impact	Interaction	**Significance Notes**
Delegate makes better decision	0.51** (0.19)	-0.44 (0.22)	0.28 (0.27)	Delegation to AI is justified by "better decisions" more often, especially in RealCons. HypoCons reduces this justification.
Decision difficult or unclear preference	0.52** (0.19)	0.81*** (0.20)	-0.70** (0.25)	Difficulty drives justification for AI delegation, particularly in RealCons scenarios. HypoCons reduces this reasoning for AI.
Insufficient information	-0.42 (0.22)	0.47* (0.22)	-0.11 (0.29)	HypoCons increases this justification, while AI is slightly less likely to elicit it compared to humans.
Hand over responsibility	0.06 (0.24)	0.56* (0.24)	-0.46 (0.32)	Responsibility-shifting is justified more often in HypoCons, regardless of delegate type.
Simplify and avoid dealing	0.28 (0.25)	0.34 (0.26)	0.12 (0.11)	No significant differences. Less prominent justification overall.

Note: The findings from the multiple-choice question after the decision shed light on how participants rationalize their delegation decisions, rather than uncovering the true motivational drivers. These results support the hypothesis that delegation to AI is justified more frequently by perceived capability (e.g., "better decisions") and decision difficulty, particularly in RealCons conditions. Responsibility-shifting appears more prominent in HypoCons scenarios, which might reflect greater willingness to admit to this justification when the decision lacks real consequences.

* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table A.8: Distribution of moral relevance ratings: "How morally significant do you find the situation?"

Moral Relevance	Frequency	Percent	Cumulative Percent
... very significant	147	18.38%	18.38%
... significant	397	49.62%	68.00%
... slightly significant	169	21.12%	89.12%
... insignificant	64	8.00%	97.12%
... morally very insignificant	23	2.88%	100.00%
Total	800	100.00%	

Table A.9: Donation decisions by condition in Study 1.

Positive framing (gain)		
Option	**Freq.**	**Percent**
Option A: €5 to the Against Malaria Foundation	157	43.98
Option B: €7 to Helen Keller International	200	56.02
Total	357	100.00

Negative framing (loss)		
Option	**Freq.**	**Percent**
Option A: Destroy €7 donation voucher to Helen Keller International	142	40.80
Option B: Destroy €5 donation voucher to the Against Malaria Foundation	206	59.20
Total	348	100.00

Table A.10: Donation decisions by condition in Study 2, restricted to the real-consequence condition.

German representative sample		
Option	**Freq.**	**Percent**
Option A: €5 to the Against Malaria Foundation	204	59.30
Option B: €7 to Helen Keller International	140	40.70
Total	344	100.00

U.S. representative sample		
Option	**Freq.**	**Percent**
Option A: $5 to the Against Malaria Foundation	971	64.60
Option B: $7 to Helen Keller International	532	35.40
Total	1,503	100.00

Table A.11: Stated reasons for choosing the Against Malaria Foundation (AMF) in open-text form in Study 1.

Theme	Brief description	Example (EN translation; participant ID)
Severity/urgency & mortality	Malaria perceived as acute and more lethal (often citing higher annual death tolls).	"Malaria is deadly; vitamin A deficiency is not necessarily. Malaria is more widespread." (ID 130)
Durability & reusability of nets	Nets seen as one-off, long-lasting, reusable; can protect multiple sleepers.	"The net can be used multiple times and is not a consumable product. So perhaps it can also save lives in the long term." (ID 175)
Concreteness & familiarity	Problem/solution felt more tangible or better understood; personal experience common.	"I am aware of the problem with malaria and I know that mosquito nets help." (ID 95)
Direct, visible impact ("save a life")	Clear line from donation to concrete protection of a child / life saved.	"Because a specific human life would be saved, I chose it." (ID 487)
Cost-effectiveness & numbers (e.g., GiveWell)	Perceived efficiency and references to rankings/ratios.	"On the GiveWell website, the Malaria Project currently had higher costs per life saved than the other charity. Additional donations thus would theoretically help more with the realization of this project than the other donation. However, the decision was not easy, as both projects are important." (ID 178)
Skepticism about Vitamin-A route	View that vitamin A can be obtained via diet or is less critical/immediate.	"I believe that vitamin A can also be consumed in ways other than through supplements . . . " (ID 58)
Other (simplicity, autonomy, fairness)	Preference to decide oneself; belief others will fund vitamin A, etc.	"I did not delegate the decision because I wanted to decide myself." (ID 168) "I assumed more people would donate to Option B because it's the larger amount, so I chose A." (662)

Notes: Translations by the authors; lightly edited for brevity. Multiple themes can co-occur. Participants frequently weighed several considerations simultaneously (e.g., severity/urgency vs. breadth of beneficiaries; durability/reusability vs. compliance concerns).

Table A.12: Stated reasons for choosing Helen Keller International (HKI) in open-text form in Study 1.

Theme	**Brief description**	**Example (EN translation; participant ID)**
More beneficiaries ("7 > 1") & higher amount (€7 vs. €5)	Preference to help more children with a single donation and/or to send the larger amount.	"With €7 I can help seven children; with nets for €5 I can only help one person." (ID 337)
Concerns about net usage/compliance	Nets protect mainly at night, may be unused/misused/stolen/break; protection not assured.	"...I was also skeptical about how effective a mosquito net is if it only provides protection from bites while you are sleeping..." (ID 77)
Broader health benefits / basic nutrition	Vitamin A strengthens immunity and prevents multiple illnesses (cause-oriented support).	"Supplementing with vitamins can prevent several diseases ..." (ID 106)
Cost-effectiveness (lives per €)	HKI perceived to save more lives per euro in the presented setup.	"... the estimated cost-per-life-saved ratio is lower for Helen Keller." (ID 370)
Implementation reliability/ease	Supplement delivery seen as simpler or more reliable than correct net installation/use.	"... A one-time treatment can help. I'm not sure mosquito nets are feasible in recipients' everyday lives." (ID 202)
Other (balancing attention, personal ties, autonomy)	Malaria already well known/funded; desire to back the other cause; personal trust/experience.	"I assumed more people would choose the better-known cause (malaria) and wanted to support the other organization..." (ID 349)

Notes: Translations by the authors; lightly edited for brevity. Multiple themes can co-occur. Participants frequently weighed several considerations simultaneously (e.g., severity/urgency vs. breadth of beneficiaries; durability/reusability vs. compliance concerns).

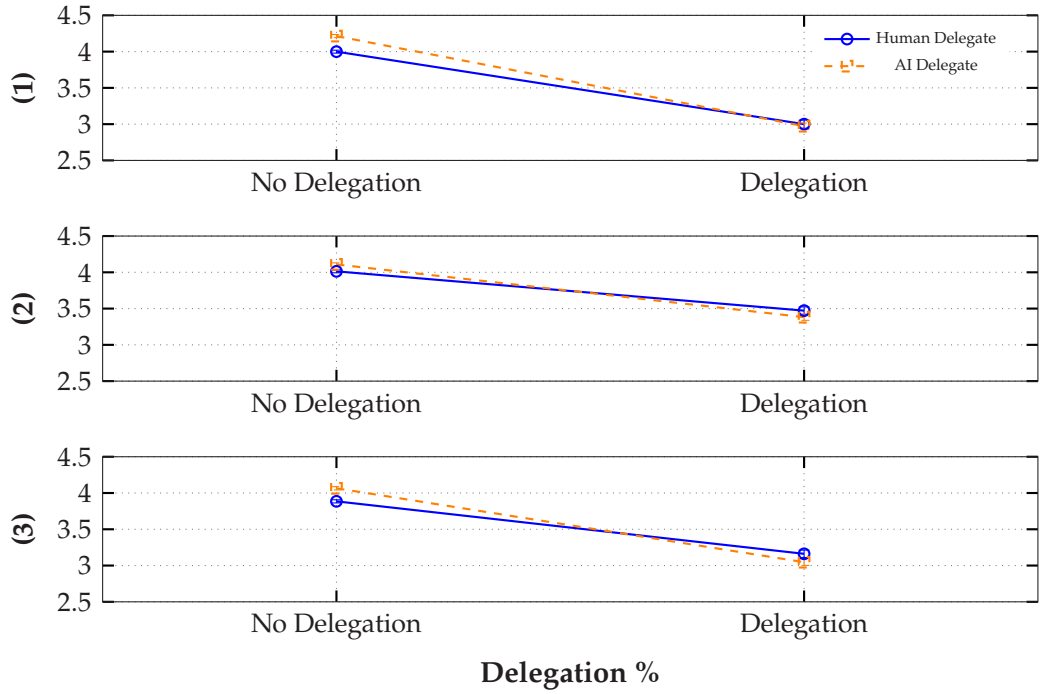


Figure A.2: Interaction plots for responsibility measures (1)–(3), showing effects of delegation (No vs. Yes) and delegate type (Human vs. AI). (01) "I would like to be fully responsible for the decision, whatever the outcome.", (02) "I feel responsible for the outcome of this decision.", and (03) "I have a moral obligation to make such a decision myself." Error bars represent 95% confidence intervals.

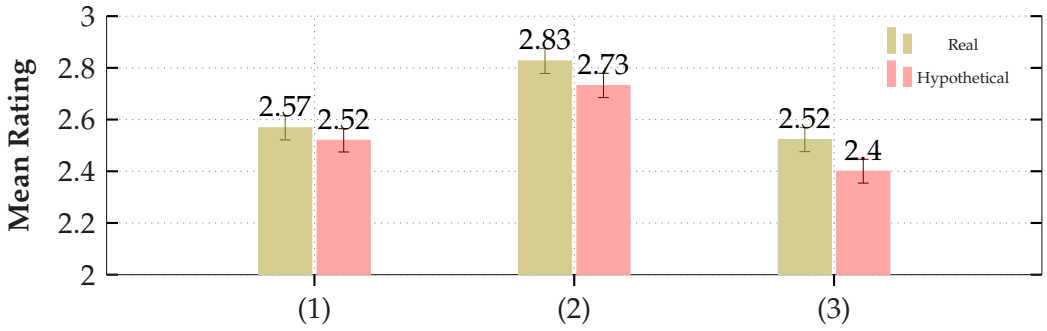
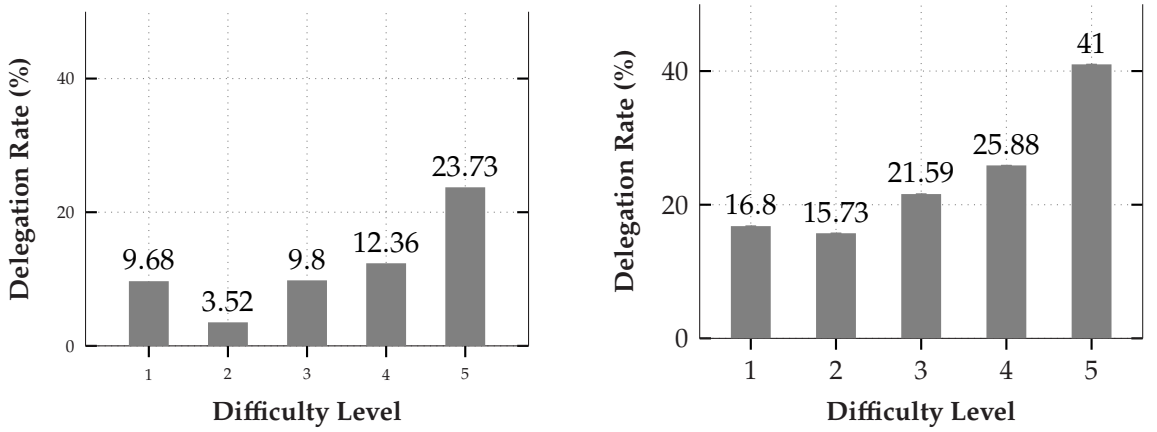


Figure A.3: Comparison of AI's capability ratings by decision impact on a five-point scale (1 = strongly disagree to 5 = strongly agree) for three questions: (1) "In a situation as described in this study, an artificial intelligence (AI) can make a better decision between two donations than I can" (t-test, $p = 0.0720$); (2) "I have full confidence that an AI can make a high-quality decision between two donations in a situation like this" ($p = 0.0029$); and (3) "AI can make good moral decisions." ($p = 0.0001$). Error bars represent 95% confidence intervals.



(a) Study 1: Delegation increases as the difficulty level rises from 9.68% to 23.73%

(b) Study 2: Delegation increases as the difficulty level rises from 16.8% to 41.00%

Figure A.4: Delegation rates for each level of decision difficulty on a 5-point scale for both samples. Difficulty is significantly higher for delegators than non-delegators (Sample 1 $\chi^2(4, N = 800) = 26.16, p < 0.001$, Sample 2 $\chi^2(4, N = 4,843) = 136.58, p < 0.001$).

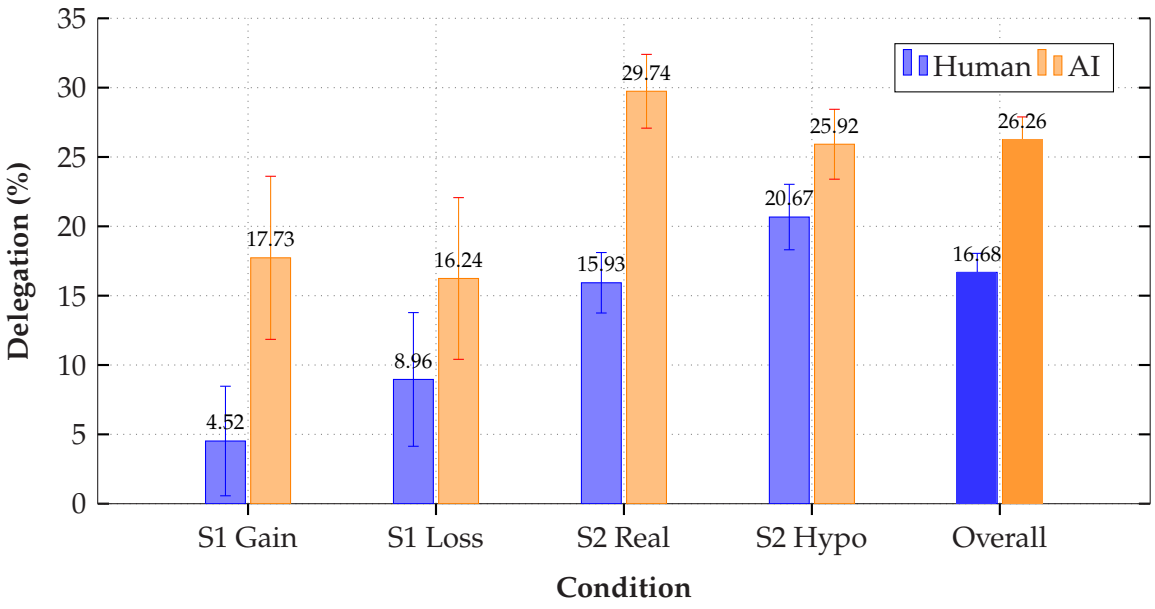


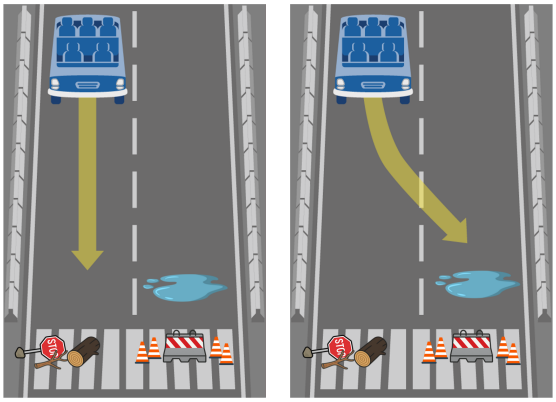
Figure A.5: Delegation rates by condition and delegate type. Bars show means (%) with 95% confidence intervals. Overall rates are weighted across all situations and both studies.

B Appendix for Chapter 2

B.1 Materials and Measures

B.1.1 Instructions

Neutral scenario example



In this case, the self-driving car with sudden brake failure will continue ahead.
This will result in a collision with . . .

- A traffic sign on the left side (side impact)
- A branch on the left side (side impact)
- A tree trunk in front (front impact)

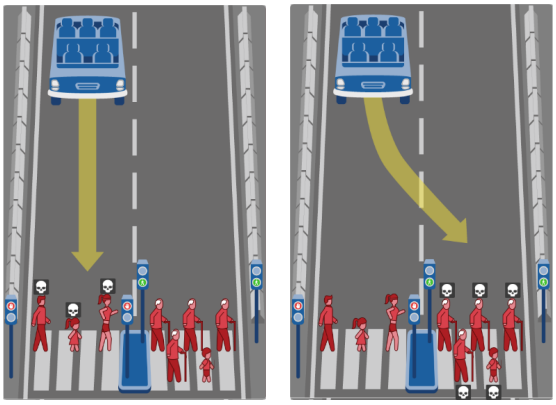
Note that this side of the street is dry.

In this case, the self-driving car with sudden brake failure will swerve and drive onto the other lane.
This will result in a collision with . . .

- Two pylons on the left side (side impact)
- A concrete column in front (front impact)
- Two pylons on the right side (side impact)

Note that this side of the street is wet, which reduces braking efficiency.

Moral scenario example



In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead.
This will result in . . .
Dead:

- One man
- One girl
- One female athlete

Note that the affected pedestrians are flouting the law by crossing on the red signal.

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane.
This will result in . . .
Dead:

- Four old men
- One boy

Note that the affected pedestrians are abiding by the law by crossing on the green signal.

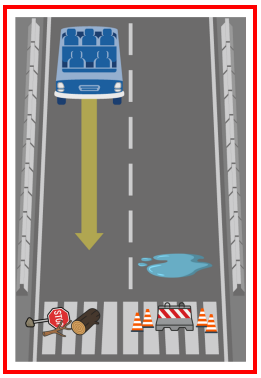
Figure B.1: Example of matched neutral and moral scenario. The moral scenario images were generated using the scenario creation feature of the Moral Machine online experimental platform (Awad et al., 2018) and are used with permission. The matched neutral scenario images were created by the authors using the Moral Machine road background and AI-generated object images.

Neutral scenario example

The AI's Decision

The AI of the autonomous vehicle decides whether the vehicle should stay on course or take evasive action and move into the other lane. The AI considers factors such as the number of obstacles, the characteristics of the obstacles (e.g. size and hardness), and the road conditions affecting brake efficiency.

In this case, the AI decides to continue ahead. **The car collides with three obstacles: one traffic sign (side impact), one branch (side impact), one tree trunk (front impact).**



Moral scenario example

The AI's Decision

The AI of the autonomous vehicle decides whether the vehicle should stay on course or take evasive action and move into the other lane. The AI considers factors such as the number of affected individuals, the characteristics of the persons (e.g. age and gender), and whether they are obeying traffic laws.

In this case, the AI decides to continue ahead.

Three individuals are killed: one man, one girl, one female athlete.

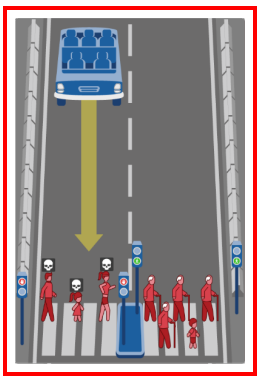
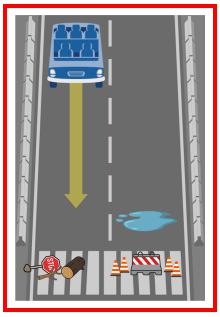


Figure B.2: Example decision screen for matched moral and neutral scenario with decision, outcomes, and short information on the AI's decision logic.

Neutral scenario example



The AI's approach: The AI's decision-making process involves evaluating factors, such as the road conditions affecting braking (e.g. dry or wet), the number of objects on each lane and their characteristics, as well as their position and resulting impact angle (front vs. side impact). Based on these factors, the AI assigns a value for the severity of the expected property damage on each lane and makes the decision: the car will drive toward the lane with the lower assigned value, resulting in a crash with the objects on that lane, while the lane with the objects on the other side will be avoided. Factors that lower a lane's value increase the likelihood that the car will proceed in that direction. Factors that increase a lane's value encourage avoiding that lane.

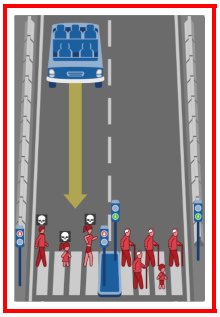
Factors in this scenario:
On the left lane, the autonomous vehicle would collide with three obstacles: a traffic sign and a branch on the left side (side impact), and a tree trunk in front (front impact). The road on this side is dry, meaning braking efficiency is not affected, which reduces the assigned risk value. The tree trunk poses the largest risk on this lane due to its large size and solidity. This is further amplified by its central position. The traffic sign slightly increases the assigned value further while the branch—being lightweight and movable—reduced the value somewhat.
On the right lane, the vehicle would collide with five obstacles: a concrete column in front (front impact), and two pylons on each side (side impact). The road on this lane is wet, which negatively affects braking efficiency, increasing the assigned risk. The concrete column is the primary contributor to the high assigned value due to its solid structure and central position. While the pylons are less hazardous and help lower the overall value slightly despite their number, they cannot offset the effect of the centrally positioned concrete column coupled with the unfavorable road conditions.

Ultimately, the AI chose to continue ahead, hitting the obstacles on the left lane. The decision was driven by the extremely high impact value of the concrete column on the right lane, which outweighed the combined contributions of the tree trunk and traffic sign on the left, despite the dry road conditions on that lane.

If an additional tree trunk were located in the middle of the left lane, **the AI would still have maintained its original decision**. Although two tree trunk would then have a higher impact than before, the risk of the concrete column in the middle is still rated substantially higher.

If the tree trunk on the left lane and the concrete column on the right lane switched sides, **the AI would change its decision**. This is because the left lane would then contain the concrete column, which contributes a much higher risk than the tree trunk now located on the right, outweighing other factors. As a result, the total assigned damage value would be higher for the left lane, making the right lane the more favorable choice despite its wet road conditions.

Moral scenario example



The AI's approach: The AI's decision-making process involves evaluating factors, such as whether the car continues on its current path or requires intervention to swerve, whether pedestrians are crossing legally (e.g., at a green light) or illegally (e.g., at a red light), the number of individuals on each lane, and their characteristics. Based on these factors, the AI assigns a value to the collective entities on each lane and makes the decision: the car will drive toward the lane with the lower assigned value, resulting in the death of the individuals on that lane, while the individuals on the other lane will be spared. Factors that lower a lane's value increase the likelihood that the car will proceed in that direction. Factors that increase a lane's value encourage avoiding that side and thus saving the individuals on it.

Factors in this scenario:
On the left lane, the autonomous vehicle would collide with three individuals: one man, one girl, and one female athlete. This group is crossing illegally against a red light, which reduces their assigned value. Sparing them would require intervention, which slightly decreases this group's value further. The girl makes the most significant positive contribution to this group's value. The man and the female athlete also slightly increase the assigned value. However, the effect of illegal crossing partially offsets these individual characteristics. Thus, overall, the left lane is assigned a low to medium value.
On the right lane, the autonomous vehicle would collide with five individuals: four elderly men and one boy. This group is crossing legally with a green light, which significantly increases their assigned value. No intervention is required to spare this group, which slightly increases their favorability in the AI's decision further. The high number of individuals on this lane contributes strongly to the group's total value. While the elderly men reduce the assigned value slightly, this effect is outweighed by the presence of the boy. Overall, the right lane is thus assigned a high value.

Ultimately, the AI chose to continue ahead, sparing the group on the right and sacrificing the group on the left. The decision was driven by the legal crossing behavior and higher number of individuals on the right, which outweighed the stronger individual contributions on the left lane.
If an additional girl had been present on the left lane, **the AI would still have maintained its original decision**. Although her presence would have significantly increased the assigned value of the left lane and slightly reduced the difference in the number of individuals between the lanes, the group would still be penalized for illegal crossing and for requiring intervention to be spared. These factors would continue to outweigh the individual contributions, and the right lane would remain more valuable overall.

If the boy from the right lane had been on the left lane instead, **the AI would have changed its decision**. This is because both lanes would then contain the same number of individuals, but the group on the left would consist entirely of individuals with positive value contributions: the girl, the boy, the man, and the female athlete. In contrast, the group on the right would consist only of elderly men, who substantially lower the lane's assigned value. In this case, the higher individual value of the left-lane group would outweigh the penalties from illegal crossing and required intervention, prompting the AI to swerve and spare them.

Figure B.3: Example explanation for matched moral and neutral scenario. Explanations follow established post-hoc XAI formats and include: (1) a brief description of the AI's decision logic, example features considered, and the value-calculation process; (2) a local feature-attribution summary indicating which scenario elements mattered most and in which direction, together with the resulting decision; and (3) counterfactual "what-if" statements showing how plausible single-feature changes would affect the decision, including one change that leaves the decision unchanged and one that flips it.

B.1.2 Scales

B.1.2.1 Affect Scales

Table B.1: Item wording for STAI (anticipated state anxiety), MCI (epistemic curiosity), and dissonance discomfort.

Scale	Wording
STAI, state anxiety (Marteau and Bekker, 1992; Spielberger et al., 1971)	I feel calm.; I am tense.; I feel upset.; (I am relaxed.); I feel content.; I am worried.
MCI, state curiosity (Naylor, 1981)	I want to know more.; I want things to make sense.; I am speculating about what is happening.; I feel inquisitive.; I want to explore possibilities.; (I am feeling puzzled.)
Dissonance discomfort (Elliot and Devine, 1994; Matz and Wood, 2005)	I feel uncomfortable.; I feel uneasy.; I feel bothered.; I feel tense.; I feel concerned.

Notes. All items use the same stem shown in the survey: "At the thought of reading the explanation for the AI decision, right now, at this moment . . . " In Study 1, STAI and the selected items of the MCI used a 4-point response scale ("Not at all", "Somewhat", "Moderately", "Very much"), while dissonance discomfort used a 7-point scale anchored by "Not at all" and "Very much". In Study 2, MCI and dissonance discomfort used a 7-point scale anchored by "Not at all" and "Very much". Item 4 of STAI had to be excluded due to almost perfect correlations with Items 1 and 5 (Heywood case). Item 6 of the MCI had to be excluded to reach acceptable model fit.

B.1.2.2 Motivations

Table B.2: Item wording for accuracy motivation and defense motivation.

Measure	Wording
Accuracy	To what extent have you been thinking about different factors in the scenario? In this scenario, it is important to me to understand which decision is best based on the facts, regardless of the decision I made. I wanted to obtain the most accurate possible understanding of the situation.
Defense	The possibility that the explanation would challenge my view/evaluation of the situation influenced my decision about whether I wanted to look at it. To what extent did you think about your own original decision in the situation when evaluating the AI's decision? For me, sticking with my own judgment in this scenario was more important than following the AI's decision.

Notes. Items use a 7-point response scale anchored by "Not at all" and "Very much".

B.1.2.3 Scales for Controls

Table B.3: Item wording for the Miller Behavioral Style Scale (MBSS; Miller, 1987).

Scenario	
Dentist scenario	1. I would ask the dentist exactly what he was going to do. 2. I would take a tranquilliser or have a drink before going. 3. I would try to think about pleasant memories. 4. I would want the dentist to tell me when I would feel pain. 5. I would try to sleep. 6. I would watch all the dentist's movements and listen for the sound of the drill. 7. I would watch the flow of water from my mouth to see if it contained blood. 8. I would do mental puzzles in my mind.
Lay-off scenario	1. I would talk to my fellow workers to see if they knew anything about what the supervisor's evaluation of me said. 2. I would review the list of duties for my present job and try to figure out if I had fulfilled them all. 3. I would go to the movies to take my mind off things. 4. I would try to remember any arguments or disagreements I might have had with the supervisor that would have lowered his opinion of me. 5. I would push all thoughts of being laid off out of my mind. 6. I would tell my spouse that I would rather not discuss my chances of being laid off. 7. I would try to think which employees in my department the supervisor might have thought had done the worst job. 8. I would continue doing my work as if nothing special was happening.

Notes. Participants tick all statements that might apply to them.

Table B.4: Item wording for algorithm aversion, affinity for technology, moral conviction, utility, and attitude toward AI.

Measure	Wording
Algorithm aversion	Which decision maker would you choose?
Affinity for technology (Franke et al., 2019)	It is enough for me that a technical system works; I don't care how or why. I try to understand how a technical system exactly works. I predominantly deal with technical systems because I have to. I try to make full use of the capabilities of a technical system. I enjoy spending time becoming acquainted with a new technical system. It is enough for me to know the basic functions of a technical system. When I have a new technical system in front of me, I try it out intensively. I like testing the functions of new technical systems. I like to occupy myself in greater detail with technical systems. It is important that you pay attention to this study. Please tick 'Completely disagree'.
Moral conviction (Skitka, 2010; Skitka et al., 2005)	... connected to your core moral beliefs or convictions? ... based on fundamental questions of right and wrong? ... based on moral principles?
Utility (Davis, 1989)	... would help me to better understand and evaluate the AI's decision. ... would improve my decision-making in similar situations.
Attitude toward AI (Grassini, 2023)	I believe that AI will improve my professional or personal projects. I think AI technology is positive for humanity. I think I will use AI technology in the future. I believe that AI will improve my life.

Notes. Algorithm aversion uses a 7-point response scale anchored by "Definitely a human decision maker" and "Definitely an AI decision maker". Affinity for technology uses a 7-point Likert response ("Completely disagree" – "Completely agree"). The final statement is an attention check. Moral conviction and utility items use a 7-point response scale anchored by "Not at all" and "Very much". Attitude toward AI items use a 7-point agreement scale anchored by "Not at all" and "Completely agree".

B.1.3 AI Model

In this study, we developed an AI model to make binary decisions based on scenarios derived from the Moral Machine experiment (Awad et al., 2018), a dataset that is publicly accessible. Our focus was on only pedestrians' scenarios from the USA, reflecting a subset of the data (N=17,850,148). We further refined this dataset by selecting only those scenarios involving pedestrians on both sides, and excluding entries with missing values, resulting in 15,723,644 observations. Because of limited computing power we trimmed the dataset to 10,000,000 which is still more than enough data to train our algorithm. Given the binary nature of the decision-making process in our study, we required an algorithm capable of comparing two options directly. To this end, we employed a ranking algorithm, specifically a gradient boosting tree algorithm implemented via the XGBoost package (Chen and Guestrin, 2016). Ranking algorithms, akin to the logic employed by search engines, prioritize options

based on a relevance score, even when limited to comparing only two entries at a time. The final model assigned a value to the group of pedestrians on each side of the street based on behavior data of real humans. Finally, we employed *SHAP* (*SH*apley *A*dditive ex*P*lanations) values to interpret AI decision-making (Lundberg and Lee, 2017). We used SHAP values to elucidate the influence of each feature on the model's decisions. They allow us to compare both scenarios and build understandable explanations.

B.2 Additional Analyses

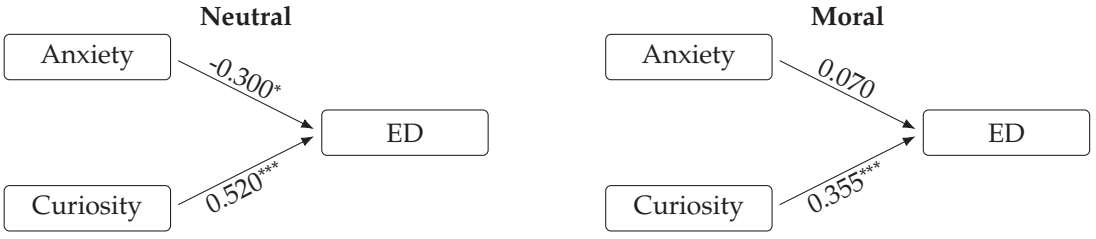
B.2.1 Study 1

B.2.1.1 Context × Affect

Table B.5: Logistic regression results for explanation demand with context × emotion interaction (item-consistent with SEM), Study 1.

	OR	Std. Error
Context (Moral vs. Neutral)	0.609	0.694
STAI (Neutral)	0.515*	0.159
Context × STAI (Moral)	2.199*	0.869
MCI (Neutral)	2.823***	0.698
Context × MCI (Moral)	0.674	0.232
Constant	1.362	1.120

Note: Robust standard errors. Wald $\chi^2(5) = 28.32$, Pseudo $R^2 = 0.0912$, $N = 393$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.



- Anxiety: $b_1 = -0.300$, 95% CI [−0.549; −0.050], $p = 0.019$
- Curiosity: $b_2 = 0.520$, 95% CI [0.303; 0.737], $p < 0.001$

- Anxiety: $b_1 = 0.070$, 95% CI [−0.163; 0.302], $p = 0.558$
- Curiosity: $b_2 = 0.355$, 95% CI [0.139; 0.570], $p = 0.001$

Cross-group differences in path coefficients (neutral−moral): $\Delta b_{\text{Anxiety}} = -0.369$, $p < 0.001$; $\Delta b_{\text{Curiosity}} = 0.166$, $p = 0.084$.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p \leq .001$.

Figure B.4: Multi-group paths predicting explanation demand (neutral vs. moral decision context), Study 1. Values are unstandardized probit coefficients (WLSMV).

B.2.1.2 Context × Decision Congruence Interaction

Table B.6: Logit regression results for explanation demand depending on context, decision congruence, and their interaction, Study 1.

	Coefficient	Std. Error
Context (Moral vs. Neutral)	-1.1359*	0.4581
Decision congruence	-0.6807	0.4414
Context × Decision congruence	1.5077*	0.6048
Constant	2.3321***	0.3312

Note: LR $\chi^2(3) = 7.0700$, $p = 0.0697$, Pseudo $R^2 = 0.0227$, $N = 393.0000$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table B.7: Predicted explanation demand from a logistic model with a context × decision congruence interaction, Study 1.

	Neutral context	**Moral context**
Incongruent decision	0.912	0.768
Congruent decision	0.839	0.883
Moral − Neutral (simple effects)		
Incongruent	−14.4 pp, $p = .021$	
Congruent	+4.4 pp, $p = .358$	
Congruence effects within context		
Congruent vs. Incongruent in Neutral	−7.2 pp, $p = .128$	
Congruent vs. Incongruent in Moral	+11.5 pp, $p = .066$	
Difference-in-differences	+18.8 pp, $p = .017$	

Note. "pp" denotes percentage points. The context × congruence interaction is statistically significant ($p = .013$). Marginal predictions are reported for each cell.

Table B.8: Exploratory OLS regressions, Study 1: state curiosity (MCI) and state anxiety (STAI) depending on context, decision congruence, and their interaction.

	MCI		STAI	
	Coefficient	Std. Error	Coefficient	Std. Error
Context (moral vs. neutral)	-0.433^{**}	0.133	0.606^{***}	0.133
Decision congruence (DC= 1)	-0.512^{***}	0.116	−0.053	0.116
Context × DC	0.565^{**}	0.173	−0.044	0.174
Constant	3.058^{***}	0.076	1.745^{***}	0.077

Note. OLS regressions. Two-sided tests. Curiosity: $F(3, 389)=8.03$, $p < 0.001$, $R^2=0.0583$, Adj. $R^2=0.0511$, $N=393$. STAI: $F(3, 389)=15.78$, $p < 0.001$, $R^2=0.1085$, Adj. $R^2=0.1016$, $N=393$.
$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

B.2.1.3 Robustness Check

Table B.9: Exploratory moderated mediation, Study 1: Wald tests for moderation by decision congruence (DC) of the *a*- and *b*-paths in the controlled latent CB-SEM (ED as binary outcome).

Effect	Wald χ^2	p	Interpretation (English)
Anxiety×DC → ED	0.635	0.425	No evidence that decision congruence moderates the anxiety–ED link.
Curiosity×DC → ED	8.286	0.004	Curiosity predicts ED more strongly under agreement with the AI decision.
Moderation (DC), both *b*-paths (joint)	8.772	0.013	At least one *b*-path differs by decision congruence.
Slope anxiety → ED | DC= 0	1.202	0.273	Anxiety slope is not reliably different from zero under incongruence.
Slope anxiety → ED | DC= 1	1.572	0.210	Anxiety slope is not reliably different from zero under congruence.
Slope curiosity → ED | DC= 0	8.372	0.004	Curiosity positively predicts ED under incongruence.
Slope curiosity → ED | DC= 1	15.940	6.54×10^{-5}	Curiosity positively predicts ED under congruence.
Context×DC → anxiety	4.247	0.039	The context effect on anxiety depends on decision congruence.
Context×DC → Curiosity	5.313	0.021	The context effect on curiosity depends on decision congruence.
Context×DC, both mediators (joint)	8.889	0.012	At least one *a*-path is moderated by decision congruence.
DC → anxiety (main effect)	1.464	0.226	No overall anxiety difference by congruence.
DC → Curiosity (main effect)	2.883	0.090	Trend-level overall difference in curiosity by congruence.
DC → ED (direct)	0.203	0.652	No direct effect of congruence on ED.
Controls → anxiety (joint)	5.362	0.252	Controls do not jointly predict anxiety.
Controls → Curiosity (joint)	13.296	0.010	Controls jointly predict curiosity.
Controls → ED (joint)	2.480	0.648	Controls do not jointly predict ED.
Residual covariance anxiety↔Curiosity	17.012	3.72×10^{-5}	Positive residual covariance between anxiety and curiosity.
ATI → Curiosity (main effect)	7.773	0.005	Higher ATI is associated with higher curiosity.
ATI → anxiety (main effect)	1.071	0.301	ATI is not reliably associated with anxiety.
ATI → ED (direct)	0.750	0.387	ATI is not directly associated with ED.

Note. DC indicates decision congruence with the AI decision (DC= 0 incongruence, DC= 1 congruence). Anxiety and curiosity denote latent factors measured by STAI and MCI items, respectively. Wald tests are two-sided.

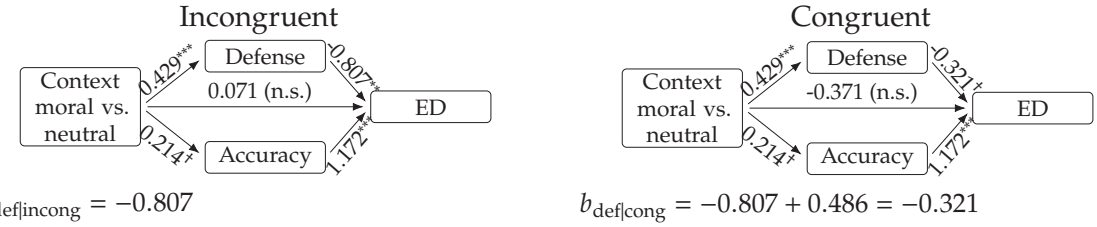
Table B.10: Exploratory robustness check, Study 1: standardized key path estimates from the latent mediation model without vs. with the preregistered control block (including DC interactions).

	Without controls		With controls	
Effect	Est.	p	Est.	p
Context → anxiety	0.349***	< .001	0.495***	< .001
Context → Curiosity	−0.132*	.015	−0.369***	< .001
Context → ED (c')	0.017	.829	−0.137	.261
Anxiety → ED (b_1)	−0.095	.233	−0.073	.273
Curiosity → ED (b_2)	0.410***	< .001	0.178**	.004
Context×DC → anxiety	–	–	−0.546*	.039
Context×DC → Curiosity	–	–	0.451*	.021
Anxiety×DC → ED	–	–	−0.086	.425
Curiosity×DC → ED	–	–	0.268**	.004
Indirect effects				
ind_anxiety (overall)	−0.033	.239	–	–
ind_Curiosity (overall)	−0.054*	.026	–	–
ind_Anxiety | DC= 0	–	–	−0.036	.272
ind_Anxiety | DC= 1	–	–	0.008	.775
ind_Curiosity | DC= 0	–	–	−0.066**	.008
ind_Curiosity | DC= 1	–	–	0.037	.594
Total (overall)	−0.070	.381	–	–
Total indirect (overall)	−0.087*	.013	–	–
Total | DC= 0	–	–	−0.238	.065
Total | DC= 1	–	–	−0.092	.547

Note. DC = decision congruence. Controls include the full preregistered control block; the qualitative pattern remains unchanged (no direct context effect; curiosity robustly predicts explanation demand). Significance levels * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

B.2.2 Study 2

B.2.2.1 Motivations as Mechanism



$b_{\text{def|incong}} = -0.807$

$b_{\text{def|cong}} = -0.807 + 0.486 = -0.321$

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Figure B.5: Reduced gSEM (R3b), Study 2. Decision congruence moderates the effect of defense on ED. Other paths are identical across panels. Indirect effects are reported in Table B.11.

Table B.11: Exploratory gSEM with indirect effects of context on explanation demand via defense and accuracy, Study 2.

Indirect effect	Estimate	SE	z	p	95% CI
$a_{\text{def}}b_{\text{def\|incong.}}$	-0.346	0.133	-2.61	0.009	[−0.606, −0.086]
$a_{\text{def}}b_{\text{def\|cong.}}$	-0.138	0.089	-1.55	0.122	[−0.312, 0.037]
$a_{\text{acc}}b_{\text{acc}}$	0.250	0.146	1.72	0.086	[−0.035, 0.536]

B.2.2.2 Regression Models for Explanation Demand

Table B.12: Logit regression results for explanation demand including emotions with context moderation, Study 2.

	Coefficient	Std. Error
Context (moral vs. neutral)	-1.535	1.001
Decision congruence (congruent)	-0.622	0.441
Context × Decision congruence	0.652	0.620
Dissonance discomfort	-0.156	0.126
Context × Dissonance	0.014	0.165
Curiosity	0.481***	0.110
Context × Curiosity	0.317	0.178
Constant	0.205	0.666

Note: LR $\chi^2(7) = 70.310$, $p < 0.001$, Pseudo $R^2 = 0.172$, $N = 492$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table B.13: Logit regression results for explanation demand including emotions with decision congruence moderation, Study 2.

	Coefficient	Std. Error
Context (moral vs. neutral)	0.293	0.495
Decision congruence (congruent)	-2.074*	0.970
Context × Decision congruence	-0.168	0.637
Dissonance discomfort	-0.335**	0.126
Decision congruence × Dissonance	0.346*	0.168
Curiosity	0.503***	0.149
Decision congruence × Curiosity	0.146	0.184
Constant	0.705	0.853

Note: LR $\chi^2(7) = 73.100$, $p < 0.001$, Pseudo $R^2 = 0.178$, $N = 492$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table B.14: Logit regression results for explanation demand including emotions and motivations, Study 2.

	Coefficient	Std. Error
Context (moral vs. neutral)	0.606	0.519
Decision congruence (congruent)	-3.125*	1.429
Context × Decision congruence	-1.021	0.710
Dissonance discomfort	-0.366**	0.131
Decision congruence × Dissonance	0.347	0.182
Curiosity	0.250	0.201
Decision congruence × Curiosity	0.107	0.239
Defense motivation	-0.792**	0.281
Decision congruence × Defense	0.366	0.323
Accuracy motivation	1.098***	0.162
Constant	-0.067	1.192

Note: LR $\chi^2(10) = 137.740$, $p < 0.001$, Pseudo $R^2 = 0.336$, $N = 492$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

Table B.15: Logit regression results for explanation demand including motivations, controls, and moral conviction, Study 2.

	Coefficient	Std. Error
Context (moral vs. neutral)	0.513	0.547
Decision congruence (congruent vs. incongruent)	-1.982	1.401
Context × Decision congruence	-0.661	0.698
Accuracy motivation	0.973***	0.156
Defense motivation	-0.850***	0.261
Decision congruence × Defense	0.448	0.280
Utility beliefs	0.608***	0.124
Algorithm aversion	0.029	0.110
Attitude toward AI	-0.138	0.128
Gender (male)	-0.462	0.343
Gender (other)	-1.919	1.145
Age	0.010	0.010
Moral conviction	-0.051	0.101
Constant	-1.574	1.418

Note: LR $\chi^2(13) = 150.460$, $p < 0.001$, Pseudo $R^2 = 0.368$, $N = 491$.

Gender category "prefer not to answer" is omitted due to perfect prediction (one observation dropped).

* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

B.2.2.3 Analysis of Dissonance and Curiosity

Table B.16: Preregistered exploratory OLS regressions, Study 2: curiosity and dissonance depending on context, decision congruence, and their interaction.

	Curiosity		Dissonance	
	Coefficient	Std. Error	Coefficient	Std. Error
Context (moral vs. neutral)	−0.367	0.203	1.505***	0.250
DC (congruent vs. incongruent)	−1.110***	0.198	−0.947***	0.244
Context × DC	0.800**	0.281	0.256	0.345
Constant	5.770***	0.145	2.922***	0.178

Note. OLS regressions. Two-sided tests. For curiosity, $F(3, 488)=11.320$, $p < 0.001$, $R^2=0.065$, adj. $R^2=0.059$, $N=492$. For dissonance, $F(3, 488)=38.530$, $p < 0.001$, $R^2=0.192$, adj. $R^2=0.187$, $N=492$.

$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

B.2.2.4 Moral Conviction and Defense Motivation

Table B.17: OLS regression results for defense motivation depending on context and moral conviction.

	Coefficient	Std. Error
Context (moral vs. neutral)	0.041	0.107
Moral conviction	0.238***	0.031
Constant	3.581***	0.146

Note: Robust SEs. $F(2, 489) = 37.010$, $p < 0.001$, $R^2 = 0.155$, $N = 492$.
* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$.

C Appendix for Chapter 3

C.1 Material

C.1.1 Human- and Computer-like Task & Anthropomorphic LLM CA Design Cues

Task Description:
In the following, you will observe three exemplary dialogs with a conversational agent. You have the role of a potential user who wants to get to know different variants of the conversational agent in order to then decide on one of the variants. For this task, it is important that you understand the user's needs and select the best variant of the conversational agent based on them.

A: Health Agent:
Recently, you have felt a lot of stress in your everyday life and have often felt left alone with it. You want to do something about it, as you know that this can have a negative impact on your health and quality of life in the long term. Through a search on the Internet, you came across the "health agent", who regularly talks to you about your mental and physical well-being and offers you advice if you have any problems.

You would like to use this conversational agent and now have the opportunity to look at different versions of the health agent and choose the one you like best afterwards. Please take a close look at the following dialogs with the conversation agent.

Put yourself in the role of the user (dialog box with blue background) who is talking to the health agent (dialog box with light grey background). By **clicking the space bar**, you can call up the next part of the dialog. After you have seen a variant of the agent, you will be asked to make some assessments of the conversation agent. After you have seen all the variants, you will be asked to make your preference decision. Please take your time to watch the entire dialog.

B: SmartHome agent:
A year ago, you installed a SmartHome system in your home to intelligently control the heating and lighting system. The aim and purpose of this retrofit was to be able to conveniently control the temperature and light from anywhere in order to optimize your consumption. To make it even easier to interact with your SmartHome system, the system provider now offers a conversation-based SmartHome agent. This agent can control the "smart" systems in your home, make consumption-optimizing recommendations and provide information on past consumption. You would like to use this conversation agent and now have the opportunity to look at different variants of the SmartHome

agent and select the variant that you like best afterwards. Please take a close look at the following dialogs with the conversation agent. [. . .]

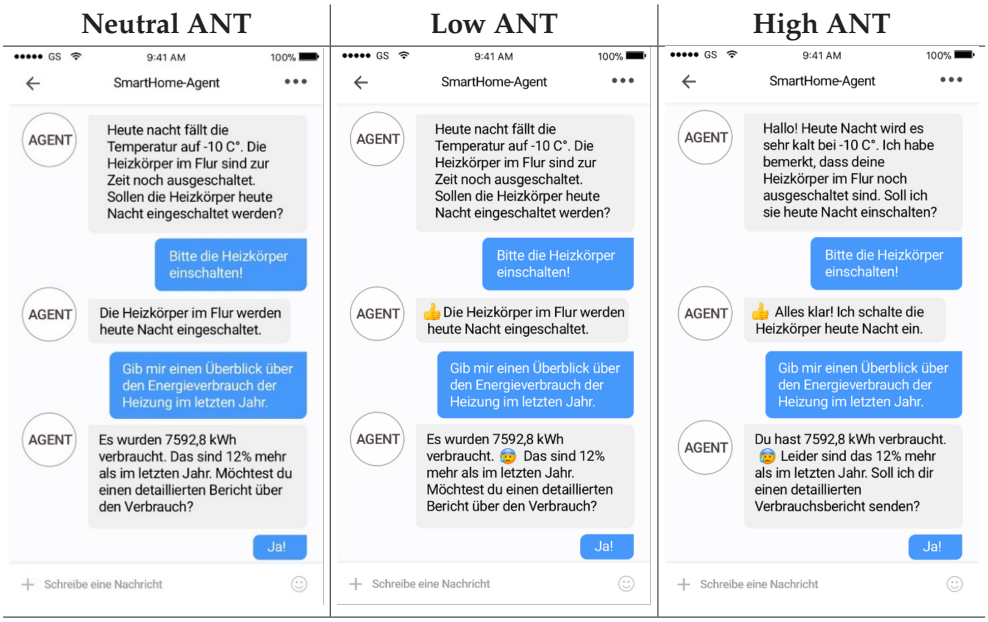


Table C.1: Example scenario vignette for the Smarthome agent.

Social Presence and Efficiency (randomized order)

- The agent conveys a sense of human contact personality warmth | togetherness | human sensitivity.
- The user only had to interact briefly with the agent to perform the task.
- The interaction with the agent was less time-consuming for the user.
- The agent tried to perform the task quickly.

Preference Decision:
Please decide now. Which variant of the SmartHome agent / health agent best meets your requirements for a social interaction partner? Please rank all three versions according to your preference (1= highest preference; 3= lowest preference)

- I trust new technologies until I have a contrary experience.
- I usually trust new technologies unless they give me a reason to the contrary.
- If in doubt, I would tend to trust a new technology.

Please rate the following statements.

- Technologies, devices or machines (e.g. computers, cars, televisions) have intentions.
- A fish has free will.
- A mountain has free will.
- A television set has feelings.
- A robot has its own consciousness.
- Cows have intentions.
- A car has free will.

- The ocean has its own consciousness.
- A computer has a mind of its own.
- A big cat has feelings.
- The environment has feelings.
- An insect has a mind of its own.
- A tree has a mind of its own.
- The wind has intentions.
- A reptile has its own consciousness.

C.1.2 Task Human-Likeness – Pretest Study

Please rate the following list of tasks with regard to whether you think each task is more computer-like (1) or human-like (7):

- Solve a mathematical problem.
- Analyze and interpret data to extract meaningful insights.
- Guidance and advice for making purchase decisions
- Supporting code writing for a software program.
- Providing solutions for customer service inquiries.
- Search and access relevant information.
- Use statistical tests to forecast future outcomes (e.g., stocks, elections, performance)
- Determine the moral responsibility for an accident that occurred.
- Teaching a language.
- Diagnose a physical health condition.
- Diagnose a mental health condition.
- Monitor a smart home system to ensure performance and problem detection.

C.1.3 Tasks Solved with ChatGPT – Study 3

Moral Dilemma Task

Please imagine you are tasked with assessing responsibility in the following scenario. You intend to use ChatGPT to discuss potential assignments of responsibility. Please familiarize yourself with the scenario.

The Scenario

A state-of-the-art self-driving car (RoboCar) is traveling on a two-lane country road with a passenger named Kim. The car is following the speed limit and driving in the right lane. As the RoboCar approaches a curve, it detects a man running on the right lane at a short distance ahead. The RoboCar attempts to brake, but the brakes are locked, and it cannot stop in time. If the RoboCar continues on its current path, it will hit and kill the man. However, the RoboCar recognizes that the left lane is empty, and there is no oncoming traffic. It has the option to switch to the left lane, which would prevent the accident and spare the man's life.

1. The RoboCar decides not to change lanes.
2. Kim, the passenger, has the ability to intervene and change lanes.
3. If Kim changes lanes, nobody will be hit.
4. If Kim doesn't change lanes, the man will be hit and killed.
5. Kim decides not to change lanes.

Health Task

Please imagine that you are Toni in the following scenario. You intend to use ChatGPT to discuss the reasons for your symptoms and explore potential solu-tions. Please familiarize yourself with the scenario:

The Scenario

Toni is 22 years old and has just moved to a new city to pursue a master's degree. Toni lives alone and finds the studies demanding, leaving little time to make social connections. The university offers various events and initiatives to meet new people, but Toni hasn't taken advantage of them yet due to the need for study time. Achieving a strong degree is vital to Toni, as it would enhance job prospects and aid in repaying a substantial student loan. Recently, Toni has been waking up in the middle of the night, experiencing rapid heartbeat, sweating, and nausea. These sudden symptoms are not related to nightmares and disappear quickly. Now, Toni is becoming concerned about what this could develop into.

SQL Task

SQL Introduction

```
SELECT name, age, salary FROM employee WHERE age > 40;
```

The above statement will select all values in the Name, Age, and Salary columns from the Employee table where the age is greater than 40. You can use following operators in the select statement:

=	equal to	<=	less than or equal to
>	greater than	<> or !=	not equal to
<	less than	LIKE	character/string comparison
>=	greater than or equal to		

Python Task

Please read this short introduction to Python coding. Afterwards, you will be asked to answer six questions about the Python project provided at the end of this page with the help of ChatGPT.

Introduction

Python is a high-level, general-purpose programming language known for its simplicity and readability. It's widely used for web development, data science, artificial intelligence, automation, and more.

Some Python Basics:

Please take a look at the following Python coding project. After reading it, please take the task sheet titled "Python Task" and respond to all questions with the help of ChatGPT.

Python Coding Project:

```
def check_guess(guess, answer):
    global score
    still_guessing = True
    attempt = 0
    while still_guessing and attempt < 3:
        print("Your Score is "+ str(score))
```

C.2 Scales and Variables Measured

Table C.2: Mind perception items, 7-point scale (Bigman and Gray, 2018; Gray et al., 2007; Gray and Wegner, 2012).

To what extent do you think an AI can/is . . .	
Experience	. . . can feel pleasure . . . can have desires . . . can express a personality . . . can be happy
Agency	. . . can think . . . can communicate with others . . . can remember things . . . can plan actions

Notes. Items are presented in random order within each dimension.

Table C.3: Items measuring knowledge of AI technologies and software coding skills.

Knowledge of AI technologies	Please rate your familiarity and knowledge of AI technologies. Please choose only one of the following:
	No Knowledge Basic Knowledge Proficient Knowledge Expert Knowledge
Software coding skills	Please rate your familiarity and proficiency in software coding skills. (. . .)

Table C.4: Trust items, 7-point scale (McKnight and Choudhury, 2002).

Measure	**Wording**
Goodwill trust (benevolence & integrity)	I believe that ChatGPT/the agent would act in my best interest. If I required help, ChatGPT/the agent would do its best to help me. ChatGPT/the agent is interested in my well-being, not just its own. ChatGPT/the agent is truthful in its dealings with me. I would characterize ChatGPT/the agent as honest. ChatGPT/the agent would keep its commitments.
Competence trust	ChatGPT/the agent is competent and effective in helping with completing the task. ChatGPT/the agent performs its role of assisting with task completion very well. Overall, ChatGPT/the agent is capable and proficient.

Table C.5: Eeriness items (Gray and Wegner, 2012; Ho and MacDorman, 2010)

Measure	**Wording**
Eeriness index	ChatGPT seemed | reassuring | eerie
Humanness index	ChatGPT seemed | natural | artificial
Uncanniness	ChatGPT seemed | comforting | unnerving

C.3 Additional Analyses

C.3.1 Findings from Factor Analysis for Composite Reliability

Table C.6: Rotated component matrix, Study 2.

Item	Experience	Agency
Agency-1	0.205	0.758
Agency-2	0.269	0.627
Agency-3	-0.082	0.849
Agency-4	-0.017	0.853
Experience-1	0.929	0.121
Experience-2	0.934	0.098
Experience-3	0.909	0.069
Experience-4	0.928	0.092

Note. Extraction method: principal component analysis. Rotation method: varimax with Kaiser normalization. Rotation converged in 3 iterations.

Table C.7: Convergent and discriminant validity, Study 2.

Latent construct	Composite reliability	Cronbach's alpha	AVE	1	2
1. Agency	0.858	0.750	0.604	0.777	—
2. Experience	0.960	0.949	0.856	0.201	0.925

Table C.8: Rotated component matrix, Study 3. Factor loadings and cross-loadings of the measures.

	Anthro-pomorphism	Trusting belief: Benevolence	Trusting belief: Integrity	Trusting belief: Competence	Dispositional trust
ANT_1	**0.873**	0.097	0.003	0.024	0.007
ANT_2	**0.706**	0.048	0.036	0.01	0.172
ANT_3	**0.862**	0.018	0.006	-0.106	-0.057
ANT_4	**0.895**	0.016	0.039	-0.01	-0.057
ANT_5	**0.825**	0.022	0.065	0.101	0.073
ANT_6	**0.875**	-0.05	-0.075	0.019	0.029
ANT_7	**0.826**	0.026	0.117	-0.057	0.058
TB_BEN_1	-0.025	**0.697**	0.431	0.251	0.209
TB_BEN_2	-0.039	**0.559**	0.469	0.418	0.093
TB_BEN_3	0.08	**0.901**	0.122	0.159	0.074
TB_INT_1	0.031	0.329	**0.79**	0.295	0.062
TB_INT_2	-0.003	0.36	**0.811**	0.231	0.067
TB_INT_3	0.071	0.087	**0.743**	0.39	0.072
TB_CMP_1	-0.031	0.17	0.281	**0.877**	0.066
TB_CMP_2	0.006	0.275	0.291	**0.828**	0.082
TB_CMP_3	0.018	0.264	0.335	**0.826**	0.084
DT_1	0.078	0.078	0.081	0.02	**0.888**
DT_2	0.062	0.069	0.145	-0.066	**0.896**
DT_3	-0.068	0.111	-0.054	0.243	**0.745**

Note. Extraction method: principal component analysis. Rotation method: equamax with Kaiser normalization. Rotation converged in 6 iterations.

Table C.9: Construct attributes (e.g., reliability and related measurement statistics).

Latent construct	Composite reliability	Cronbach's alpha	AVE	1	2	3	4	5
1. Anthropomorphism	0.94	0.93	0.70	0.84				
2. Trusting Belief: Benevolence	0.77	0.83	0.54	0.05	0.73			
3. Trusting Belief: Integrity	0.82	0.87	0.61	0.07	0.69*	0.78		
4. Trusting Belief: Competence	0.88	0.93	0.71	0.01	0.63*	0.66*	0.84	
5. Dispositional Trust	0.88	0.82	0.72	0.07	0.27*	0.20*	0.19*	0.85

Note. Square roots of AVE are shown on the diagonal. The lower triangle reports correlations between constructs. $^{*}p < .01$ (two-tailed).

C.3.2 ANOVA Agency and Experience with Control Variables

Table C.10: ANOVA for experience including control variables, Study 2.

Within-subjects factor (Task)

Effect	*F*	df	*p*
Task (Main Effect)	5.692	1,147	0.018
Task * Gender	0.515	1,147	0.474
Task * Age	2.518	1,147	0.115
Task * AI Experience	0.636	1,147	0.426
Task * Coding Experience	2.361	1,147	0.127
Task * Partner	0.683	1,147	0.410

Between-subjects factor (Interaction partner)

Effect	*F*	df	*p*
Intercept	89.775	1	< .001
SEX	2.254	1	0.135
Age	0.218	1	0.641
AIExp	1.702	1	0.194
CodingExp	0.981	1	0.324
Partner	129.685	1	< .001

Table C.11: ANOVA for agency including control variables, Study 2.

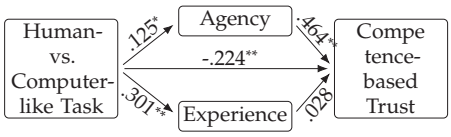
Within-subjects factor (Task)

Effect	*F*	df	*p*
Task (Main Effect)	2.265	1	0.134
Task * Gender	2.667	1	0.105
Task * Age	3.531	1	0.062
Task * AI Experience	1.413	1	0.236
Task * Coding Experience	0.382	1	0.538
Task * Partner	5.379	1	0.022

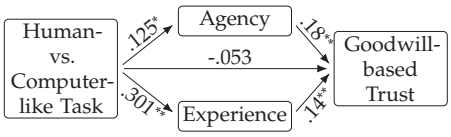
Between-subjects factor (Interaction partner)

Effect	*F*	df	*p*
Intercept	396.044	1	< .001
SEX	0.034	1	0.853
Age	2.696	1	0.103
AIExp	0.483	1	0.488
CodingExp	3.454	1	0.065
Partner	25.742	1	< .001

C.3.3 Mediation Models



(a) Agency: a*b coeff. =.06, *SE* =0.03, 95% CI =[.001, .126]
Experience: a*b coeff. =.01, *SE* =0.02, 95% CI =[-.034, .049], ns
Total Effect: -.442 **



(b) Agency: a*b coeff. =.02, *SE* =0.01, 95% CI =[.000, .052]
Experience: a*b coeff. =.04, *SE* =0.02, 95% CI =[.006, .078]
Total Effect: .012

Figure C.1: Mediation models for competence-based and goodwill-based trust, Study 3.

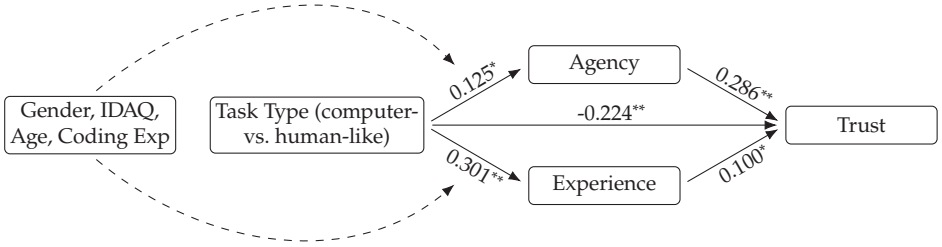


Figure C.2: Moderated mediation with gender, IDAQ, age, and coding experience, Study 3.
Note. Experience: a*b path coefficient = 0.029, *SE* = 0.016, 95% CI = [-0.002, 0.062].
Agency: a*b path coefficient = 0.036, *SE* = 0.020, 95% CI = [0.002, 0.078].
The moderation analysis shows that age, gender, IDAQ and coding experience have no significant influence on the mediation paths. The main effects remain stable, indicating that mediation is robust to individual differences.

(a) a*b coeff. =-.31, *SE* =.28, 95% CI =[-.83, 0.25]
Total Effect: .3010 **

(b) a*b coeff. =-.3, *SE* =.14, 95% CI =[-.63, -.05]
Total Effect: .125 *

Figure C.3: Mediation with ChatGPT's and users' language cues, Study 3.

C.3.4 LIWC Analysis

Table C.12: LIWC word count for human-like and computer-like tasks, Study 3.

	Human-like Tasks	Computer-like Tasks
ChatGPT	1411.36 (SD=375.73)	894.66 (SD=327.83)
User	241.78 (SD=96.14)	236.24 (SD=154.03)

Experience: Experience encompasses the emotional and social aspects of anthropomorphism (Gray et al. 2007, Waytz et al. 2010a) and is operationalized through the two default LIWC categories: affective processes (e.g., happy, sad, sweet) and prosocial behavior (e.g., help, thanks) words. Both categories have been demonstrated to correlate with the experience dimension of mind perception (Schweitzer and Waytz 2021). Consistent with the construction of LIWC summary variables (Pennebaker et al. 2014), we employed an additive model to calculate our experience score. This computation yields a percentage of all experience-related (affective and prosocial) words in the text. A higher experience score indicates a greater utilization of experience-related language cues.

Agency: We operationalize agency through the LIWC-based default variable—analytical thinking. Agency reflects the cognitive dimension of anthropomorphism, demonstrating the ability to think, plan and be rational (Gray et al. 2007, Waytz et al. 2010a). The LIWC analytical thinking score measures whether individuals write in an analytic and logical style based on the use of function words (Pennebaker et al. 2014). Low analytical thinking scores reflect more intuitive and personal language (Jordan et al. 2019), while high scores reflect advanced cognitive abilities (Jordan and Pennebaker 2017, Pennebaker et al. 2014).

C.4 Additional Theories

	Description	Why it is less applicable
Stereotype Content Model (SCM)	SCM proposes that warmth (Friendliness, Trustworthiness) and competence (Capability, Assertiveness) are two fundamental dimensions of social perception that reliably differentiate societal group stereotypes across cultures. The model distinguishes four broad stereotype categories, each associated with distinct emotional and behavioral responses (Cuddy et al., 2008; Fiske, 2018): • High warmth and high competence elicit admiration and ative facilitation (e.g helping) • high warmth and low competence elicit pity and passive facilitation (e.g. allowing privileges) • Low warmth and high competence elicit envy and passive harm (e.g. exclusion) • Low warmth and low competence elicit contempt and active harm (e.g. discrimination) Warmth and competence are influenced by perceived social structure: status predicts competence, competition predicts low warmth (Fiske et al., 2002).	SCM primarily focuses on perception and categorization of humans and social groups, while our study examines perceptions of artificial agents (Large Language Models). While SCM has been applied to human-machine interactions, including studies on robots and chatbots (Krügel et al., 2022; Seiler and Schär, 2021), it originates from research on stereotypes and social perception in human groups. There are conceptual overlaps between SCM and MPT with warmth and experience capturing emotional capacity and competence and agency capturing cognitive capacity (Cuddy et al., 2008; Gray et al., 2007). Although SCM could thus be considered a relevant alternative, its behavioral consequences – helping, privileges exclusion, and discrimination – are fundamentally different from the mechanisms we study. Furthermore, mind perception theory has been developed to understand when and why people attribute mental states to others, specifically including non-human entities, and has been more widely applied in AI research, particularly in analyzing anthropomorphism in conversational agents and robots, making it the more suitable theoretical foundation for our study.(Gray et al., 2007; Kawai et al., 2023; Waytz et al., 2010b).
Dehumanization	Dehumanization refers to the denial of human-like mental attributes to individuals or social groups (Haslam, 2006). Haslam (2006) distinguishes two forms: mechanistic dehumanization (denying emotional capacity and warmth, treating entities as cold, robotic) and animalistic dehumanization (denying cognitive capacity, treating entities as primitive or irrational). Dehumanization is linked to intergroup biases, moral disengagement, and justification of harm, with empirical evidence highlighting its role in discrimination, violence, and exclusion (Haslam and Loughnan, 2014; Kelman, 2017). The theory is frequently ap-plied in contexts of ethnic conflict, political propaganda, and the denial of moral consideration to outgroups. As artificial agents are often seen as high in agency but low in experience (Gray and Wegner, 2012), this could be judged as a form of mechanistic dehumanization.	Traditional dehumanization research focuses on interpersonal and intergroup contexts, where humans dehumanize other humans. In contrast, our study investigates how non-human agents (LLMs) are anthropomorphized, not dehumanized, since LLMs were never considered fully human. Dehumanization is further especially relevant to moral judgement and prosocial scenarios, where some research has applied it to AI (Heßler et al., 2022; Swiderska and Küster, 2020). However, in this study we do not focus on moral judgement about the LLM. Mind Perception Theory (MPT) provides a framework better tailored to our context, explaining how users attribute different levels of mind perception without implying prior human status.

List of Publications

Hüholt, N. and Szech, N. (2026). Trusting machines with morality — delegating moral decisions to AI. *European Economic Review*, 184:105255.

Bibliography

Acemoglu, D. (2024). The simple macroeconomics of AI. Technical Report 32487, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w32487.

Acemoglu, D. and Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160.

Adam, D. (2024). Lethal AI weapons are here: How can we control them? *Nature*, 629(ePub):521–523.

Adam, M., Roethke, K., and Benlian, A. (2022). Human vs. automated sales agents: How and why customer responses shift across sales stages. *Information Systems Research*, 34(3):1148–1168.

Agrawal, A., Gans, J., and Goldfarb, A. (2019). Economic policy for Artificial Intelligence. *Innovation Policy and the Economy*, 19(1):139–159.

Agrawal, N. and Maheswaran, D. (2005). Motivated reasoning in outcome-bias effects. *Journal of Consumer Research*, 31(4):798–805.

Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I., and Rieser, V. (2024). All too human? Mapping and mitigating the risk from anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):13–26.

Akrich, M. (1992). The de-scription of technical objects. In Bijker, W. E. and Law, J., editors, *Shaping Technology/Building Society: Studies in Sociotechnical Change*. MIT Press.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. (2023). Self-consuming generative models go mad. *arXiv preprint*, arXiv:2307.01850.

Allen, C., Wallach, W., and Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, 21(4):12–17.

Anderson, M. and Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15–26.

Anderson, M. and Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press.

Anguiano, D. (2025). AI lovers grieve loss of ChatGPT's old model: "like saying goodbye to someone I know". The Guardian. https://www.theguardian.com/technology/2025/aug/22/ai-chatgpt-new-model-grief. Accessed: 2026-01-30.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

Appel, M., Izydorczyk, D., Weber, S., Mara, M., and Lischetzke, T. (2020). The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior*, 102:274–286.

Apple (2024). Introducing Apple Intelligence for iPhone, iPad, and Mac. Apple Newsroom (Press Release). `https://www.apple.com/newsroom/2024/06/introducing-apple-intelligence-for-iphone-ipad-and-mac/`. Accessed: 2026-01-30.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85:183–189.

Argenton, C., Potters, J., and Yang, Y. (2023). Receiving credit: On delegation and responsibility. *European Economic Review*, 158:104522.

Arnd-Caddigan, M. (2015). Sherry Turkle: Alone together: Why we expect more from technology and less from each other. *Clinical Social Work Journal*, 43(2):247–248.

Attard-Frost, B., De los Ríos, A., and Walters, D. R. (2023). The ethics of AI business practices: a review of 47 AI ethics guidelines. *AI and Ethics*, 3(2):389–406.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–30.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729):59–64.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, 4(2):134–143.

Banner, M. (2025). Steam engines to artificial intelligence: Why developers and consultants need to rethink now. Forbes Technology Council (Council Post), Forbes. `https://www.forbes.com/councils/forbestechcouncil/2025/07/29/steam-engines-to-artificial-intelligence-why-developers-and-consultants-need-to-rethink-now/`.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and 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*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Barlow, A. K. J., Siddiqui, N. Q., and Mannion, M. (2004). Developments in information and communication technologies for retail marketing channels. *International Journal of Retail & Distribution Management*, 32(3):157–163.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Bartling, B., Engl, F., and Weber, R. A. (2014). Does willful ignorance deflect punishment? An experimental study. *European Economic Review*, 70:512–524.

Bartling, B., Fehr, E., and Özdemir, Y. (2023). Does market interaction erode moral values? *The Review of Economics and Statistics*, 105(1):226–235.

Bartling, B. and Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies*, 79(1):67–87.

Bauer, K., von Zahn, M., and Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, 34(4):1582–1602.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology*, 45:180–191.

Bick, A., Blandin, A., and Deming, D. J. (2026). The rapid adoption of generative AI. *Management Science*.

Bigman, Y. E. and Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181:21–34.

Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & Technology*, 31(4):543–556.

Björkegren, D. (2025). AI is transforming the economy — understanding its impact requires both data and imagination. *Nature*, 648:535–537. Comment.

Blut, M., Wang, C., Wünderlich, N., and Brock, C. (2021). Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, 49:632–658.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R. B., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L. E., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., and et al. (2021). On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Bonnefon, J.-F., Rahwan, I., and Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology*, 75:653–675.

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? *Journal of Econometrics*, 65(1):83–108.

Briesch, M., Sobania, D., and Rothlauf, F. (2024). Large language models suffer from their own output: An analysis of the self-consuming training loop. arXiv preprint. arXiv:2311.16822.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942.

Brynjolfsson, E. and McAfee, A. (2017). The business of artificial intelligence. *Harvard Business Review*, 7(1):1–2.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (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).

Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.

Calvino, F., Haerle, D., and Liu, S. (2025). Is generative AI a general purpose technology? Implications for productivity and policy. OECD Artificial Intelligence Papers 40, OECD Publishing.

Caplin, A. and Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *The Quarterly Journal of Economics*, 116(1):55–79.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Casal, J. E. and Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing? a study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, 2(3):100068.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, page 520–527, New York, NY, USA. Association for Computing Machinery.

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825.

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180080.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., and Wadman, K. (2025). How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research.

Chen, L., Chen, P., and Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8:75264–75278.

Chen, S., Shechter, D., and Chaiken, S. (1996). Getting at the truth or getting along: Accuracy-versus impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology*, 71(2):262.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA. ACM.

Cheng, X., Zhang, X., Cohen, J., and Mou, J. (2022). Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. *Information Processing & Management*, 59:102940.

Chugunova, M. and Sele, D. (2022). We and it: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics*, 99:101897.

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., and Zemmel, R. (2023). The economic potential of generative AI: The next productivity frontier. Technical report, McKinsey & Company. `https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier`.

Cint GmbH (2024). Academic Research. `https://de.cint.com/academic-research`. Accessed: 2024-12-04.

Citron, D. K. (2007). Technological due process. *Washington University Law Review*, 85:1249.

Citron, D. K. and Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89:1.

Clark, A., James, T., Vance, A., and Lowry, P. B. (2025). Rethinking intake: Exploring patient experience and AI-mediated interviews. In *AMCIS 2025 TREOs*, number 202 in AIS TREO Papers. Association for Information Systems.

Coffman, L. C. (2011). Intermediation reduces punishment (and reward). *American Economic Journal: Microeconomics*, 3(4):77–106.

Cohn, M., Pushkarna, M., Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z., and Heldreth, C. (2024). Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*, pages 54:1–54:15, New York, NY, USA. ACM. CHI 2024 Poster. `https://doi.org/10.1145/3613905.3650818`. Accessed: 2026-01-30.

Council of Europe (2024). Council of Europe framework convention on artificial intelligence and human rights, democracy and the rule of law. CM(2024)52-final. `https://rm.coe.int/1680afae3c`. Accessed: 2026-01-30. Adopted 17 May 2024; opened for signature 5 September 2024.

Covington, P., Adams, J., and Sargin, E. (2016). Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 191–198, New York, NY, USA. Association for Computing Machinery.

Crafts, N. (2021). Artificial intelligence as a general-purpose technology: an historical perspective. *Oxford Review of Economic Policy*, 37(3):521–536.

Crolic, C., Thomaz, F., Hadi, R., and Stephen, A. T. (2022). Blame the bot: Anthropomorphism and anger in customer–chatbot interactions. *Journal of Marketing*, 86(1):132–148.

Cuddy, A. J., Fiske, S. T., and Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. In *Advances in Experimental Social Psychology*, volume 40, pages 61–149. Academic Press.

Dafoe, A. (2018). AI governance: A research agenda. Technical report, Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, Oxford, UK. Version 1.0.

Dana, J., Cain, D. M., and Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100(2):193–201.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, 29:245–268.

Danaher, J. (2019). The rise of the robots and the crisis of moral patiency. *AI & Society*, 34(1):129–136.

Darley, J. M. and Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, 44(1):20.

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications.

Dattner, B., Chamorro-Premuzic, T., Buchband, R., and Schettler, L. (2019). The legal and ethical implications of using AI in hiring. *Harvard Business Review*, 25:1–7.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Quarterly*, 13(3):319–340.

Davuluri, P. (2024). New experiences coming to Copilot+ PCs and Windows 11. Microsoft Windows Experience Blog. `https://blogs.windows.com/windowsexperience/2024/10/01/new-experiences-coming-to-copilot-pcs-and-windows-11/`. Accessed: 2026-01-30.

De Santis, A., Siciliano, B., De Luca, A., and Bicchi, A. (2008). An atlas of physical human–robot interaction. *Mechanism and Machine Theory*, 43(3):253–270.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P., Krueger, F., and Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3):331–349.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper; The Wharton School Research Paper 24-013, Harvard Business School. Available at SSRN: https://ssrn.com/abstract=4573321.

Deng, J. and Lin, Y. (2023). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2):81–83.

Dennis, A. R., Kim, A., Rahimi, M., and Ayabakan, S. (2020). User reactions to covid-19 screening chatbots from reputable providers. *Journal of the American Medical Informatics Association*, 27(11):1727–1731.

Diederich, S., Brendel, A. B., Morana, S., and Kolbe, L. (2022). On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, 23(1):96–138.

Diederich, S., Janssen-Müller, M., Brendel, A. B., and Morana, S. (2019a). Emulating empathetic behavior in online service encounters with sentiment-adaptive responses: Insights from an experiment with a conversational agent. In *Proceedings of the International Conference on Information Systems (ICIS)*.

Diederich, S., Lichtenberg, S., Brendel, A. B., and Trang, S. (2019b). Promoting sustainable mobility beliefs with persuasive and anthropomorphic design: Insights from an experiment with a conversational agent. In *Proceedings of the International Conference on Information Systems (ICIS)*.

Dietvorst, B. J. and Bartels, D. M. (2022). Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies. *Journal of Consumer Psychology*, 32(3):406–424.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology. General*, 144(1):114–126.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.

Ditto, P. H. and Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63(4):568–584.

Ditto, P. H., Pizarro, D. A., and Tannenbaum, D. (2009). Motivated moral reasoning. *Psychology of learning and motivation*, 50:307–338.

Dong, M. and Bocian, K. (2024). Responsibility gaps and self-interest bias: People attribute moral responsibility to AI for their own but not others' transgressions. *Journal of Experimental Social Psychology*, 111:104584.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint. arXiv:1702.08608.

Doyle, P. (1999). When is a communicative agent a good idea. In *Proceedings Agents-99 Workshop on Communicative Agents*. Citeseer.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1):79–94.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., and Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology*, 13(3):147–164.

Eiband, M., Buschek, D., Kremer, A., and Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–6, New York, NY, USA. Association for Computing Machinery.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38.

Elgammal, A., Liu, B., Elhoseiny, M., and Mazzone, M. (2017). CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. arXiv preprint. `arXiv:1706.07068`.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5:40–60.

Elliot, A. J. and Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, 67(3):382.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint. arXiv:2303.10130.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 160–171. PMLR.

Epley, N. and Waytz, A. (2010). Mind perception. In *Handbook of Social Psychology*, pages 498–541. John Wiley & Sons, Inc., 5th edition.

Epley, N., Waytz, A., and Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4):864–886.

Erdfelder, E., Faul, F., and Buchner, A. (1996). G*Power: A general power analysis program. *Behavior research methods, instruments, & computers*, 28:1–11.

Espeland, W. N. and Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, 24(Volume 24, 1998):313–343.

EU AI Act (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union, L series, 2024/1689 (12 July 2024). `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`. Accessed: 2026-01-30. Full title: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, New York, NY.

European Commission (2026). Artificial intelligence in healthcare. `https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en`. Accessed: 2026-01-30.

European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88. `https://gdpr-info.eu/`. Cited at recital 71 (Erwägungsgrund 71).

Eurostat (2025). Use of artificial intelligence in enterprises. Eurostat Statistics Explained (European Commission). `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises`. Accessed: 2026-01-30.

Fahrenwaldt, A., tho Pesch, F., Fiedler, S., and Baumert, A. (2024). What's moral wiggle room? A theory specification. *Judgment and Decision Making*, 19:e17.

Falk, A., Neuber, T., and Szech, N. (2020). Diffusion of being pivotal and immoral outcomes. *The Review of Economic Studies*, 87(5):2205–2229.

Falk, A. and Szech, N. (2013). Morals and markets. *Science*, 340(6133):707–711.

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191.

Federspiel, F., Mitchell, R., Asokan, A., Umana, C., and McCoy, D. (2023). Threats by artificial intelligence to human health and human existence. *BMJ Global Health*, 8(5):e010435.

Feier, T., Gogoll, J., and Uhl, M. (2021). Hiding behind machines: When blame is shifted to artificial agents. Papers 2101.11465, arXiv.org.

Feine, J., Gnewuch, U., Morana, S., and Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, 132:138–161.

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

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford University Press.

Filippucci, F., Gal, P., Jona-Lasinio, C., Leandro, A., and Nicoletti, G. (2024a). The impact of Artificial Intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges. Technical Report 15, OECD Publishing, Paris.

Filippucci, F., Gal, P., and Schief, M. (2024b). Miracle or myth? Assessing the macroeconomic productivity gains from artificial intelligence. OECD Artificial Intelligence Papers 29, OECD Publishing.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2):67–73. PMID: 29755213.

Fiske, S. T., Cuddy, A. J. C., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902.

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2):e19.

Floridi, L. and Sanders, J. W. (2004). On the morality of artificial agents. *Mind and Machine*, 14(3):349–379.

Fornell, C. and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1):39.

Franke, T., Attig, C., and Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467.

Freisinger, E. and Schneider, S. (2025). Decoding decision delegation to artificial intelligence: A mixed-methods study on the preferences of decision-makers and decision-affected in surrogate decision contexts. *European Management Journal*, 43(6):958–969.

Frey, D. (1986). Recent research on selective exposure to information. *Advances in Experimental Social Psychology*, 19:41–80.

Friedman, B. and Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, Cambridge, MA.

Fritz, M. S. and MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3):233–239. PMID: 17444920.

Gao, Y., Rui, H., and Sun, S. (2023). The power of identity cues in text-based customer service: Evidence from Twitter. *Management Information Systems Quarterly*, 47(3):983–1014.

Garrett, R. K. (2009). Echo chambers online? Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, 14(2):265–285.

GDPR (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (EU)*, 679:2016.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., and Crawford, K. (2018). Datasheets for datasets. *CoRR*, abs/1803.09010.

Gefen, D. and Straub, D. W. (2004). Consumer trust in B2C e-commerce and the importance of social presence: experiments in e-products and e-services. *Omega*, 32(6):407–424.

Gerke, S., Minssen, T., and Cohen, G. (2020). Chapter 12 - ethical and legal challenges of artificial intelligence-driven healthcare. In Bohr, A. and Memarzadeh, K., editors, *Artificial Intelligence in Healthcare*, pages 295–336. Academic Press.

Gert, B. (2005). *Morality: Its Nature and Justification*. Oxford University Press.

Ghazali, A. S., Ham, J., Barakova, E. I., and Markopoulos, P. (2018). Effects of robot facial characteristics and gender in persuasive human-robot interaction. *Frontiers in Robotics and AI*, 5:73.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

GiveWell (2024). Top charities. `https://www.givewell.org/charities/top-charities`. Accessed: 2024-11-18.

Glickman, M. and Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9:345–359.

Glikson, E. and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.

Gnewuch, U., Morana, S., Hinz, O., Kellner, R., and Maedche, A. (2024). More than a bot? The impact of disclosing human involvement on customer interactions with hybrid service agents. *Information Systems Research*, 35(3):936–955.

Gnewuch, U. and Reinkemeier, F. (2025). Overcoming breakdowns in customer–chatbot interaction: Design and impact of collaborative repair strategies. *Management Information Systems Quarterly*, Forthcoming:1–33.

Gogoll, J. and Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics*, 23:681–700.

Gogoll, J. and Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, 74:97–103.

Goldman Sachs (2023). Generative AI could raise global GDP by 7%. Goldman Sachs Insights (article). `https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent`. Accessed: 2026-02-08.

Golman, R., Hagmann, D., and Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1):96–135.

Goodman, B. and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

Google Workspace (2024). Gemini in the side panel of Google Docs, Google Sheets, Google Slides, and Google Drive is rolling out now. Google Workspace Updates. `https://workspaceupdates.googleblog.com/2024/06/gemini-in-side-panel-of-google-docs-sheets-slides-drive.html`. Accessed: 2026-01-30.

Grassini, S. (2023). Development and validation of the AI attitude scale (AIAS-4): a brief measure of general attitude toward artificial intelligence. *Frontiers in psychology*, 14:1191628.

Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812):619.

Gray, K. and Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1):125–130.

Gray, K., Young, L., and Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, 23(2):101–124.

Greene, J. D. (2009). The cognitive neuroscience of moral judgment. *The Cognitive Neurosciences*, 4:1–48.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2):389–400.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American psychologist*, 35(7):603.

Grossman, Z. and van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Gruber, D., Aune, A., and Koutstaal, W. (2018). Can semi-anthropomorphism influence trust and compliance? Exploring image use in app interfaces. In *Proceedings of Technology, Mind, and Society*, volume 13, pages 1–6.

Gruetzemacher, R. and Whittlestone, J. (2022). The transformative potential of artificial intelligence. *Futures*, 135:102884.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42.

Gunning, D. and Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, 40(2):44–58.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

Hale, K. (2021). AI bias caused 80% of black mortgage applicants to be denied. Forbes. `https://www.forbes.com/sites/korihale/2021/09/02/ai-bias-caused-80-of-black-mortgage-applicants-to-be-denied/`. Accessed: 2026-02-08.

Hamman, J. R., Loewenstein, G., and Weber, R. A. (2010). Self-interest through delegation: An additional rationale for the principal-agent relationship. *American Economic Review*, 100(4):1826–46.

Han, E., Yin, D., and Zhang, H. (2023). Bots with feelings: Should AI agents express positive emotion in customer service? *Information Systems Research*, 34(3):1296–1311.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., and Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4):555.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, 10(3):252–264.

Haslam, N. and Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, 65(Volume 65, 2014):399–423.

Hertwig, R. and Engel, C. (2016). Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science*, 11(3):359–372. PMID: 27217249.

Hess, T. J., Fuller, M., and Campbell, D. E. (2009). Designing interfaces with social presence: Using vividness and extraversion to create social recommendation agents. *Journal of the Association for Information Systems*, 10(12):1.

Heßler, P., Pfeiffer, J., and Unfried, M. (2023). Conversational agents with voice: How social presence influences the user behavior in microlending decisions. In *ECIS 2023 Research Papers*.

Heßler, P. O., Pfeiffer, J., and Hafenbrädl, S. (2022). When self-humanization leads to algorithm aversion. *Business & Information Systems Engineering*, 64(3):275–292.

High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. Technical report, European Commission.

Ho, C. C. and MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices. *Computers in Human Behavior*, 26(6):1508–1518.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2019). Metrics for explainable AI: Challenges and prospects. arXiv preprint. arXiv:1812.04608.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-Ai performance. *Frontiers in Computer Science*, 5:1096257.

Holbrook, C., Holman, D., Clingo, J., and Wagner, A. R. (2024). Overtrust in AI recommendations about whether or not to kill: Evidence from two human-robot interaction studies. *Scientific Reports*, 14(1):19751.

Holmes, W., Miao, F., et al. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.

Holzwarth, M., Janiszewski, C., and Neumann, M. M. (2006). The influence of avatars on online consumer shopping behavior. *Journal of Marketing*, 70(4):19–36.

Hong, W., Chan, F. K. Y., Thong, J. Y. L., Chasalow, L. C., and Dhillon, G. (2014). A framework and guidelines for context-specific theorizing in information systems research. *Information Systems Research*, 25(1):111–136.

Hong, W., Thong, J. Y., and Tam, K. Y. (2007). How do web users respond to non-banner-ads animation? The effects of task type and user experience. *Journal of the American Society for Information Science and Technology*, 58(10):1467–1482.

Hu, K. (2023). ChatGPT sets record for fastest-growing user base — analyst note. Reuters. `https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/`. Accessed: 2026-02-08.

Hüholt, N. and Szech, N. (2026). Trusting machines with morality — delegating moral decisions to AI. *European Economic Review*, 184:105255.

International Energy Agency (2025). Energy and AI. Technical report, International Energy Agency (IEA), Paris. Licence: CC BY 4.0.

Ipsos (2025). Ipsos AI Monitor 2025. Technical report, Ipsos.

Jago, A. S. (2017). Algorithms and authenticity. *Academy of Management Discoveries*, 5(1):38–56.

Jauernig, J., Uhl, M., and Walkowitz, G. (2022). People prefer moral discretion to algorithms: Algorithm aversion beyond intransparency. *Philosophy & Technology*, 35(1):2.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Jonas, E., Schulz-Hardt, S., Frey, D., and Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology*, 80(4):557.

Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of Management Review*, 16(2):366–395.

Judd, C. M., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, 6(2):115.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *Proceedings of the 28th European Conference on Information Systems (ECIS)*, pages 15–17.

Jussupow, E., Benbasat, I., and Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *Management Information Systems Quarterly*, 48(4):1575–1590.

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., and Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, 12(6):518–527.

Kaplan, A. D., Kessler, T. T., Brill, J. C., and Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, 65(2):337–359.

Karlsson, N., Loewenstein, G., and Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty*, 38(2):95–115.

Kawai, Y., Miyake, T., Park, J., Shimaya, J., Takahashi, H., and Asada, M. (2023). Anthropomorphism-based causal and responsibility attributions to robots. *Scientific Reports*, 13(1):12234.

Kelman, H. C. (2017). Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. In *The criminology of war*, pages 145–181. Routledge.

Kirchkamp, O. and Strobel, C. (2019). Sharing responsibility with a machine. *Journal of Behavioral and Experimental Economics*, 80:25–33.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Klockmann, V., von Schenk, A., and Villeval, M. C. (2022). Artificial intelligence, ethics, and intergenerational responsibility. *Journal of Economic Behavior & Organization*, 203:284–317.

Knobloch-Westerwick, S. (2014). *Choice and preference in media use: Advances in selective exposure theory and research*. Routledge.

Köbis, N., Bonnefon, J.-F., and Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, 5(6):679–685.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., and Bielaniewicz, J. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Kordzadeh, N. and Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409.

Korinek, A. and Stiglitz, J. E. (2017). Artificial intelligence and its implications for income distribution and unemployment. Working Paper 24174, National Bureau of Economic Research.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2016). Accountable algorithms. Working Paper Fordham Law Legal Studies Research Paper No. 2765268, SSRN. Later published in University of Pennsylvania Law Review, Vol. 165 (2017).

Krügel, S., Ostermaier, A., and Uhl, M. (2022). Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. *Philosophy & Technology*, 35(1):17.

Krügel, S., Ostermaier, A., and Uhl, M. (2023a). Algorithms as partners in crime: A lesson in ethics by design. *Computers in Human Behavior*, 138:107483.

Krügel, S., Ostermaier, A., and Uhl, M. (2023b). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569.

Kruglanski, A. W. (1990). Lay epistemic theory in social-cognitive psychology. *Psychological inquiry*, 1(3):181–197.

Kshetri, N., Dwivedi, Y. K., Davenport, T. H., and Panteli, N. (2024). Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda. *International Journal of Information Management*, 75(6):102716.

Kumar, N., Qiu, L., and Kumar, S. (2022). A hashtag is worth a thousand words: An empirical investigation of social media strategies in trademarking hashtags. *Information Systems Research*, 33(4):1403–1427.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498.

Lankton, N. K., McKnight, D. H., and Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10):1.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., and Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684.

Lee, S., Lee, N., and Sah, Y.-J. (2020). Perceiving a mind in a chatbot: Effect of mind perception and social cues on co-presence, closeness, and intention to use. *International Journal of Human-Computer Interaction*, 36(10):930–940.

Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, 13(3):229–248.

Leiner, D. J. (2024). Sosci survey (version 3.5.02) [computer software]. `https://www.soscisurvey.de`. Accessed: 2024-12-04.

Lerman, C., Narod, S., Schulman, K., Hughes, C., Gomez-Caminero, A., Bonney, G., Gold, K., Trock, B., Main, D., Lynch, J., Fulmore, C., Snyder, C., Lemon, S. J., Theresa, Tonin, P., Lenoir, G., and Lynch, H. (1996). Brca1 testing in families with hereditary breast-ovarian cancer: A prospective study of patient decision making and outcomes. *JAMA*, 275(24):1885–1892.

Leydon, G. M., Boulton, M., Moynihan, C., Jones, A., Mossman, J., Boudioni, M., and McPherson, K. (2000). Cancer patients' information needs and information seeking behaviour: In depth interview study. *Bmj*, 320(7239):909–913.

Liao, Q. V., Gruen, D., and Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA. Association for Computing Machinery.

Lim, B. Y., Dey, A. K., and Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 2119–2128, New York, NY, USA. Association for Computing Machinery.

Lima, G., Grgić-Hlača, N., Jeong, J. K., and Cha, M. (2022). The conflict between explainable and accountable decision-making algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 2103–2113, New York, NY, USA. Association for Computing Machinery.

Lipsey, R. G., Carlaw, K. I., and Bekar, C. T. (2005). *Economic Transformations: General Purpose Technologies and Economic Growth*. Oxford University Press, Oxford, UK.

Lipton, Z. C. (2016). The mythos of model interpretability. arXiv preprint. arXiv:1606.03490.

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences*, 44(7):1585–1595.

Liu, B. and Sundar, S. S. (2018). Should machines express sympathy and empathy? Effects of empathic expressions by a computer agent on user perceptions and behavioral intentions. *Cyberpsychology, Behavior, and Social Networking*, 21(10):625–636.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American Economic Review*, 90(2):426–432.

Loewenstein, G. (2006). The pleasures and pains of information. *Science*, 312(5774):704–706.

Logg, J. M. (2017). Theory of machine: When do people rely on algorithms? Working Paper 17-086, Harvard Business School.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109.

Lucas, G. M., Gratch, J., King, A., and Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Lucy, L. and Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Lund, B. D. and Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News*, 40(3):26–29.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Maheswaran, D. and Chaiken, S. (1991). Promoting systematic processing in low-motivation settings: Effect of incongruent information on processing and judgment. *Journal of Personality and Social Psychology*, 61(1):13–25.

Mahmud, H., Islam, A. N., Ahmed, S. I., and Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390.

Makridakis, S. (2017). The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms. *Futures*, 90:46–60.

Malik, K. (2022). ChatGPT can tell jokes, even write articles. But only humans can detect its fluent bullshit. The Guardian. `https://www.theguardian.com/commentisfree/2022/dec/11/chatgpt-is-a-marvel-but-its-ability-to-lie-convincingly-is-its-greatest-danger-to-humankind`. Accessed: 2026-02-08.

Marteau, T. M. and Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State–Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, 31(3):301–306.

Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., and Sarkar, R. (2023). Towards understanding the interplay of generative artificial intelligence and the internet. In *International Workshop on Epistemic Uncertainty in Artificial Intelligence*, volume 14523, pages 59–73.

Maruping, L., Yin, D., Chen, A., Kankanhalli, A., Burton-Jones, A., and Brown, S. (2025). Editor's comments: Quantitative behavioral IS research—a look back and a look forward. *Management Information Systems Quarterly*, 49(1):iii–xviii.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Betts Lotufo, J., Rome, A., Shi, A., and Oak, S. (2025). Artificial intelligence index report 2025. Technical report, Stanford University, Institute for Human-Centered Artificial Intelligence (HAI).

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183.

Matz, D. C. and Wood, W. (2005). Cognitive dissonance in groups: The consequences of disagreement. *Journal of Personality and Social Psychology*, 88(1):22–37.

McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Breslau, J., Stein, B. D., et al. (2025). Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open*, 8(11):e2542281–e2542281.

McKnight, D. H. and Choudhury, V. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3):334–359.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Metz, C. and Satariano, A. (2020). An algorithm that grants freedom, or takes it away. The New York Times. `https://www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html`. Accessed: 2026-01-27.

Miller, S. M. (1987). Monitoring and blunting: validation of a questionnaire to assess styles of information seeking under threat. *Journal of Personality and Social Psychology*, 52(2):345.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, 55(5):3503–3568.

Mirbabaie, M., Brünker, F., Möllmann Frick, N. R. J., and Stieglitz, S. (2022). The rise of artificial intelligence – understanding the AI identity threat at the workplace. *Electronic Markets*, 32:73–99.

Misch, F., Park, B., Pizzinelli, C., and Sher, G. (2026). Artificial intelligence and productivity in Europe. CESifo Working Paper 12401, CESifo. SSRN eLibrary. Posted 20 Jan 2026.

Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*, 68(11):7793–7817.

Montoya, A. K. and Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, 22(1):6–27.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21.

Morana, S., Pfeiffer, J., and Adam, M. T. P. (2020). User assistance for intelligent systems. *Business & Information Systems Engineering*, 62:189–192.

Mozafari, N., Weiger, W. H., and Hammerschmidt, M. (2020). The chatbot disclosure dilemma: Desirable and undesirable effects of disclosing the non-human identity of chatbots. In *ICIS*, pages 1–18.

Mullen, E. and Skitka, L. J. (2006). Exploring the psychological underpinnings of the moral mandate effect: Motivated reasoning, group differentiation, or anger? *Journal of Personality and Social Psychology*, 90(4):629–643.

Nannini, L., Marchiori Manerba, M., and Beretta, I. (2024). Mapping the landscape of ethical considerations in explainable AI research. *Ethics and Information Technology*, 26:44.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*, pages 72–78. ACM.

National Institute of Standards and Technology (2023). Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology.

Naylor, F. D. (1981). A state-trait curiosity inventory. *Australian Psychologist*, 16(2):172–183.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Norman, D. A. (1994). How might people interact with agents? *Communications of the ACM*, 37(7):68–71.

Nougrères, A. B. (2023). Principles of transparency and explainability in the processing of personal data in artificial intelligence. Report of the Special Rapporteur on the right to privacy A/78/310, United Nations General Assembly.

Nowak, K. L. and Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 12(5):481–494.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Nørskov, S., Damholdt, M. F., Ulhøi, J. P., Jensen, M. B., Ess, C., and Seibt, J. (2020). Applicant fairness perceptions of a robot-mediated job interview: A video vignette-based experimental survey. *Frontiers in Robotics and AI*, 7:586263.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

OECD (2019). Recommendation of the council on artificial intelligence. OECD/LEGAL/0449. Adopted 22 May 2019; revised 8 November 2023. `https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449`. Accessed: 2026-01-30.

Office of Management and Budget (OMB) (2024). Revisions to OMB's statistical policy directive no. 15: Standards for maintaining, collecting, and presenting federal data on race and ethnicity. `https://www.federalregister.gov/d/2024-06469`. Accessed: 2024-12-04.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

OpenAI (2024). Introducing structured outputs in the API. `https://openai.com/index/introducing-structured-outputs-in-the-api/`. Accessed: 2025-06-25.

OpenAI (2025). Der Stand von KI für Unternehmen: The State of Enterprise AI (2025 Report). Technical report, OpenAI. https://openai.com/de-DE/index/the-state-of-enterprise-ai-2025-report/. Accessed: 2026-02-08.

Organisation for Economic Co-operation and Development (2025). Generative AI and the SME workforce: New survey evidence. Technical report, OECD Publishing, Paris.

Pandey, D. K. and Mishra, R. (2024). Towards sustainable agriculture: Harnessing AI for global food security. *Artificial Intelligence in Agriculture*, 12:72–84.

Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, Cambridge, MA.

Pazzanese, C. (2020). Ethical concerns mount as AI takes bigger decision-making role. https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/. Accessed: 2024-12-09.

Pennebaker, J., Francis, M., and Booth, R. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. LIWC.

Pentina, I., Hancock, P. A., and Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600.

Pew Research Center (2025). How Americans View Artificial Intelligence and Its Impact on People and Society. Technical report, Pew Research Center. https://www.pewresearch.org/global/2025/10/15/how-people-around-the-world-view-ai/.

Phang, J., Lampe, M., Ahmad, L., Agarwal, S., Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., and Maes, P. (2025). Investigating affective use and emotional well-being on ChatGPT. arXiv preprint. arXiv:2504.03888.

Phillips, P. J., Hahn, C., Fontana, P., Yates, A., Greene, K. K., Broniatowski, D., and Przybocki, M. A. (2021). Four principles of explainable artificial intelligence. NIST Interagency/Internal Report (NISTIR) 8312, National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.IR.8312.

Poszler, F. and Lange, B. (2024). The impact of intelligent decision-support systems on humans' ethical decision-making: A systematic literature review and an integrated framework. *Technological Forecasting and Social Change*, 204:123403.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Poushter, J., Fagan, M., and Corichi, M. (2025). How People Around the World View AI. Technical report, Pew Research Center.

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., and Taylor, S. H. (2017). "alexa is my new bff": Social roles, user satisfaction, and personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*, pages 2853–2859. ACM.

Pyszczynski, T. and Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In *Advances in Experimental Social Psychology*, volume 20, pages 297–340. Elsevier.

Qian, C. and Cong, X. (2023). Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 6(3):1.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver? In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Qiu, L. and Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems*, 25(4):145–182.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B., and Wellman, M. (2022). Machine behaviour. In *Machine Learning and the City*, pages 143–166. Wiley Online Library.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA. Association for Computing Machinery.

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28:31–38.

Reeves, B. and Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Reuters (2023). ChatGPT sets record for fastest-growing user base - analyst note. `https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/`. Accessed: 2025-02-03.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Rick, S. and Loewenstein, G. (2008). The role of emotion in economic behavior. In *Handbook of emotions, 3rd ed*, pages 138–156. The Guilford Press, New York, NY, US.

Rogha, M. (2023). Explain to decide: A human-centric review on the role of explainable artificial intelligence in AI-assisted decision making. arXiv preprint. arXiv:2312.11507.

Ropka, M. E., Wenzel, J., Phillips, E. K., Siadaty, M., and Philbrick, J. T. (2006). Uptake rates for breast cancer genetic testing: A systematic review. *Cancer Epidemiology, Biomarkers & Prevention*, 15(5):840–855.

Ross, S. I., Martinez, F., Houde, S., Muller, M., and Weisz, J. D. (2023). The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 491–514.

Rothenhäusler, D., Schweizer, N., and Szech, N. (2018). Guilt in voting and public good games. *European Economic Review*, 101:664–681.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

Rudin, C., Wang, C., and Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1):1.

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.

Savolainen, R. (2014). Emotions as motivators for information seeking: A conceptual analysis. *Library & Information Science Research*, 36(1):59–65.

Schanke, S., Burtch, G., and Ray, G. (2021). Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research*, 32(3):736–751.

Schoemann, A. M., Boulton, A. J., and Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, 8(4):379–386.

Schuetzler, R. M., Grimes, G. M., Giboney, J. S., and Rosser, H. K. (2021). Deciding whether and how to deploy chatbots. *Management Information Systems Quarterly Executive*, 20(1):1–15.

Schulz-Hardt, S., Frey, D., Lüthgens, C., and Moscovici, S. (2000). Biased information search in group decision making. *Journal of Personality and Social Psychology*, 78(4):655–669.

Schwab, K. (2016). *The Fourth Industrial Revolution*. World Economic Forum.

Sears, D. O. and Freedman, J. L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly*, 31(2):194–213.

Seeger, A.-M. (2021). Anthropomorphic conversational agents and user trust: three essays on the design and effect of anthropomorphism in human-computer interactions.

Seeger, A.-M., Pfeiffer, J., and Heinzl, A. (2018). Designing anthropomorphic conversational agents: Development and empirical evaluation of a design framework. In *39th International Conference on Information Systems, ICIS 2018*. Association for Information Systems (AIS).

Seeger, A.-M., Pfeiffer, J., and Heinzl, A. (2021). Texting with humanlike conversational agents: Designing for anthropomorphism. *Journal of the Association for Information Systems*, 22(4):8.

Seiler, R. and Schär, A. (2021). Chatbots, conversational interfaces, and the stereotype content model. In *54th Hawaii International Conference on System Sciences (HICSS), Grand Wailea, HI, USA, 5-8 January 2021*, pages 1860–1867. University of Hawai'i at Manoa.

Serra-Garcia, M. and Szech, N. (2021). The (in)elasticity of moral ignorance. *Management Science*, 68(7):4815–4834.

Seymour, M., Yuan, L. I., Riemer, K., and Dennis, A. R. (2025). Less artificial, more intelligent: Understanding affinity, trustworthiness, and preference for digital humans. *Information Systems Research*, 36(2):1096–1128.

Shank, D. B., DeSanti, A., and Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, 22(5):648–663.

Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, 1(10):694–696.

Sharma, A., Jain, A., Gupta, P., and Chowdary, V. (2021). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9:4843–4873.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. (2023). Towards understanding sycophancy in language models. arXiv preprint. arXiv:2310.13548.

Sharot, T. and Sunstein, C. R. (2020). How people decide what they want to know. *Nature Human Behaviour*, 4(1):14–19.

Sheridan, T. B. (2016). Human–robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(4):525–532.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Short, J., Williams, E., and Christie, B. (1976). *The Social Psychology of Telecommunications*. Wiley.

Sicherman, N., Loewenstein, G., Seppi, D. J., and Utkus, S. P. (2015). Financial attention. *The Review of Financial Studies*, 29(4):863–897.

Sidoti, O. and McClain, C. (2025). 34% of U.S. adults have used ChatGPT, about double the share in 2023. Pew Research Center, Short Reads. https://www.pewresearch.org/short-reads/2025/06/25/34-of-us-adults-have-used-chatgpt-about-double-the-share-in-2023/. Accessed: 2026-01-30.

Silberling, A. (2025). Parents sue OpenAI over ChatGPT's role in son's suicide. TechCrunch. https://techcrunch.com/2025/08/26/parents-sue-openai-over-chatgpts-role-in-sons-suicide/. Accessed: 2026-01-30.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, 4(4):267–281.

Skitka, L. J., Bauman, C. W., and Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology*, 88(6):895–917.

Skitka, L. J., Hanson, B. E., Morgan, G. S., and Wisneski, D. C. (2021). The psychology of moral conviction. *Annual Review of Psychology*, 72:347–366.

Skitka, L. J., Mosier, K. L., and Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006.

Skjuve, M., Folstad, A., Fostervold, K. I., and Brandtzæg, P. B. (2021). My chatbot companion – a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149:102601.

Spielberger, C. D., Gonzalez-Reigosa, F., Martinez-Urrutia, A., Natalicio, L. F., and Natalicio, D. S. (1971). The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican journal of psychology*, 5(3 & 4):145–158.

Stafford, R. Q., MacDonald, B. A., Jayawardena, C., Wegner, D. M., and Broadbent, E. (2014). Does the robot have a mind? Mind perception and attitudes towards robots predict use of an eldercare robot. *International Journal of Social Robotics*, 6:17–32.

Stanford Institute for Human-Centered Artificial Intelligence (HAI) (2025). AI index report 2025: Science and medicine. https://hai.stanford.edu/ai-index/2025-ai-index-report/science-and-medicine. Accessed: 2026-01-30.

Statistisches Bundesamt (Destatis) (2022). Statistischer Bericht: Bevölkerungsfortschreibung Zensus 2022. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Publikationen/Downloads-Bevoelkerungsstand/statistischer-bericht-bevoelkerungsfortschreibung-zensus-2022-5124107.xlsx?__blob=publicationFile. Accessed: 2024-12-04.

Steffel, M. and Williams, E. F. (2018). Delegating decisions: Recruiting others to make choices we might regret. *Journal of Consumer Research*, 44(5):1015–1032.

Steffel, M., Williams, E. F., and Perrmann-Graham, J. (2016). Passing the buck: Delegating choices to others to avoid responsibility and blame. *Organizational Behavior and Human Decision Processes*, 135:32–44.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3):213–225.

Stolte, J. F. (1994). The context of satisficing in vignette research. *The Journal of Social Psychology*, 134(6):727–733.

Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication*, 60(3):556–576.

Susarla, A., Gopal, R., Thatcher, J. B., and Sarker, S. (2023). The Janus effect of generative AI: Charting the path for responsible conduct of scholarly activities in information systems. *Information Systems Research*, 34(2):399–408.

Susser, D., Roessler, B., and Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2):1–22.

Sweeny, K., Melnyk, D., Miller, W., and Shepperd, J. A. (2010). Information avoidance: Who, what, when, and why. *Review of General Psychology*, 14(4):340–353.

Swiderska, A. and Küster, D. (2020). Robots as malevolent moral agents: Harmful behavior results in dehumanization, not anthropomorphism. *Cognitive Science*, 44(7):e12872.

Szczepański, M. (2019). Economic impacts of artificial intelligence (AI). Briefing PE 637.967, European Parliament, European Parliamentary Research Service (EPRS). Members' Research Service.

Taber, C. S. and Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in cognitive sciences*, 7(7):320–324.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25:44–56.

Touré-Tillery, M. and McGill, A. L. (2015). Who or what to believe: Trust and the differential persuasiveness of human and anthropomorphized messengers. *Journal of Marketing*, 79(4):94–110.

Trammell, P. and Korinek, A. (2023). Economic growth under transformative AI. Technical Report 31815, National Bureau of Economic Research, Cambridge, MA.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., and Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, 4(6):479–491.

UNESCO (2021). Recommendation on the ethics of artificial intelligence. Adopted by the General Conference at its 41st session, 23 November 2021. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence. Accessed: 2026-01-30.

U.S. Census Bureau (2023). National population totals and components of change: 2020-2023. https://www.census.gov/data/datasets/time-series/demo/popest/2020s-national-detail.html. Accessed: 2024-12-04.

U.S. Census Bureau (2024). Comparing race and hispanic origin. https://www.census.gov/topics/population/hispanic-origin/about/comparing-race-and-hispanic-origin.html. Accessed: 2024-12-04.

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*, 28(1):107–124.

Varian, H. R. (2019). Artificial intelligence, economics, and industrial organization. In Agrawal, A., Gans, J., and Goldfarb, A., editors, *The Economics of Artificial Intelligence: An Agenda*, pages 399–422. University of Chicago Press, Chicago, IL.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., and Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):233.

Wachter, S., Mittelstadt, B., and Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99.

Wallach, W. and Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Wang, Q., Saha, K., Gregori, E., Joyner, D., and Goel, A. (2021). Towards mutual theory of mind in human-AI interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14.

Waytz, A., Cacioppo, J., and Epley, N. (2010a). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8):383–388.

Waytz, A., Cacioppo, J., and Epley, N. (2010b). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232.

Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52:113–117.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., and Cacioppo, J. T. (2010c). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, 99(3):410–435.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., and Gabriel, I. (2021). Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Weil, E. (2023). You are not a parrot. Intelligencer, New York Magazine. `https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html`. Accessed: 2026-02-08.

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman, San Francisco.

Weller, A. (2019). Transparency: Motivations and challenges. `https://arxiv.org/abs/1708.01870`.

Wiese, E., Weis, P. P., Bigman, Y., Kapsaskis, K., and Gray, K. (2022). It's a match: Task assignment in human–robot collaboration depends on mind perception. *International Journal of Social Robotics*, 14(1):141–148.

Wilson, T. D., Centerbar, D. B., Kermer, D. A., and Gilbert, D. T. (2005). The pleasures of uncertainty: prolonging positive moods in ways people do not anticipate. *Journal of Personality and Social Psychology*, 88(1):5.

Winfield, A. (2019). Ethical standards in robotics and AI. *Nature Electronics*, 2(2):46–48.

Winner, L. (2017). Do artifacts have politics? In Weckert, J., editor, *Computer Ethics*, pages 177–192. Routledge, London. Reprint of: Winner (1980), Daedalus 109(1):121–136.

World Health Organization (2024). Artificial intelligence for health. `https://www.who.int/publications/m/item/artificial-intelligence-for-health`. Accessed: 2026-01-30.

Xu, X. and Sar, S. (2018). Do we see machines the same way as we see humans? A survey on mind perception of machines and human beings. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 472–475.

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., Cremer, D. D., Soh, H., and Gray, K. (2021). Robots at work: People prefer—and forgive—service robots with perceived feelings. *Journal of Applied Psychology*, 106(10):1557–1572.

Yang, Z. J. and Kahlor, L. (2013). What, me worry? The role of affect in information seeking and avoidance. *Science Communication*, 35(2):189–212.

Yeung, K. (2017). 'hypernudge': Big data as a mode of regulation by design. *Information, Communication & Society*, 20(1):118–136.

Young, A. D. and Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, 85:103870.

Yuan, L. and Dennis, A. R. (2019). Acting like humans? Anthropomorphism and consumer's willingness to pay in electronic commerce. *Journal of Management Information Systems*, 36(2):450–477.

Zatsu, V., Shine, A. E., Tharakan, J. M., Peter, D., Ranganathan, T. V., Alotaibi, S. S., Mugabi, R., Muhsinah, A. B., Waseem, M., and Nayik, G. A. (2024). Revolutionizing the food industry: The transformative power of artificial intelligence — a review. *Food Chemistry: X*, 24:101867.

Zhai, C., Wibowo, S., and Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11:28.

Zhang, B. and Dafoe, A. (2019). Artificial Intelligence: American attitudes and trends. Technical report, Center for the Governance of AI, Future of Humanity Institute, University of Oxford, Oxford, UK. Available at SSRN: https://ssrn.com/abstract=3312874.

Zhang, Z., Chen, Z., and Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, 101:104327.

Zhou, E. and Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, 3(3):pgae052.

Złotowski, J., Yogeeswaran, K., and Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies*, 100:48–54.

Zou, L. and Khern-am nuai, W. (2023). AI and housing discrimination: The case of mortgage applications. *AI and Ethics*, 3(4):1271–1281.