LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

DISSERTATIONEN DER LMU

UB

72

LOUIS LONGIN

The Social Role of AI Advisers

GEORG OLMS
VERLAG

The Social Role of AI Advisers

Inauguraldissertation
zur Erlangung des Doktorgrades der
Philosophie an der Ludwig-Maximilians-Universität
München

vorgelegt von
Louis David Jérôme Longin
aus Berlin
2023

Referentin: Prof. Dr. Ophelia Deroy
Korreferent: Dr. Bahador Bahrami
Datum der mündlichen Prüfung: 07. Juli 2023

Louis Longin

The Social Role of AI Advisers

Dissertationen der LMU München

Band 72

The Social Role of AI Advisers

von
Louis Longin

OLMS

UB
Universitätsbibliothek
Ludwig-Maximilians-Universität München

Eine Publikation in Zusammenarbeit zwischen dem **Georg Olms Verlag** und der **Universitätsbibliothek der LMU München**

Mit **Open Publishing LMU** unterstützt die Universitätsbibliothek der Ludwig-Maximilians-Universität München alle Wissenschaftlerinnen und Wissenschaftler der LMU dabei, ihre Forschungsergebnisse parallel gedruckt und digital zu veröffentlichen.

Erstveröffentlichung 2023
Zugleich Dissertation der LMU München 2023

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet abrufbar über http://dnb.d-nb.de

Open-Access-Version dieser Publikation verfügbar unter:
http://nbn-resolving.de/urn:nbn:de:bvb:19-325469
https://doi.org/10.5282/edoc.32546

ISBN 978-3-487-16692-6

Contents

Introduction

Part 1

Part 2

Acknowledgements

I am deeply grateful to Tanja and Theo for their love and support.

I would also like to extend my sincere thanks to my supervisors for helping me steer the stormy waves of the academic ocean.

And, of course, thank you to my parents - for everything!

List of Figures

Zusammenfassung

Der Einfluss von Künstlicher Intelligenz (KI) auf den Menschen beherrscht nicht nur die Medienwelt, sondern auch die Fachkreise. So verändert KI nicht nur die Art und Weise wie wir kommunizieren und miteinander arbeiten, sondern beeinflusst auch, wie wir die Welt um uns herum erfahren. Sprachmodelle wie ChatGPT können menschenähnlichen Text produzieren, Spiel-Engines wie AlphaZero schlagen jeden menschlichen Spieler in Schach oder Go und Fahrmodelle sind in der Lage, jedes beliebige Auto oder Drohne zu steuern. Viele essenzielle Fragen bleiben offen: Wer ist verantwortlich, wenn etwas außer Kontrolle gerät? Sind KI-Systeme handlungsfähig? Und viel weitreichender: Sind einige KI-Systeme möglicherweise empfindungsfähig? Das Forschungsgebiet KI umfasst viele verschiedene Ansätze. Sie reichen von psychologischen Studien zur Interaktion zwischen Menschen und KI, über technische Innovationen im maschinellen Lernen bis hin zu philosophischen Diskussionen über die geistigen und moralischen Fähigkeiten von KI.

Das Hauptaugenmerk in diesen Forschungsbereichen liegt vor allem auf (halb-)autonomen KI-Systemen: Systemen, die in weiten Teilen ohne menschliche Anweisungen funktionieren. Was oft unerforscht bleibt, sind KI-Systeme, die eng mit ihren menschlichen Nutzern verbunden sind: Systeme, die den Menschen in die Entscheidungsfindung einbeziehen, indem sie ihn über Handlungen oder Entscheidungen informieren oder diese empfehlen. Die Untersuchung dieser Forschungslücke wird immer dringlicher, da diese Art KI-gestützte Entscheidungsfindung in immer größerem Maße eingesetzt wird. Es geht hierbei um Entscheidungen unter geringem Risiko, wie Einkaufsempfehlungen, oder Entscheidungen unter hohem Risiko, wie medizinische Diagnosen.

Die vorliegende Dissertation untersucht exakt diese beratenden KI-Systeme und untermauert die bestehende funktionale Unterscheidung (was tun KI-Systeme) mit einer ontologischen Betrachtung (was sind beratende KI-Systeme). Ein solcher ontologischer Betrachtungsmodus schließt eine kritische Forschungslücke und ist wegweisend dafür, wie beratende KI-Systeme in ethischen und sozialen Diskussionen betrachtet werden: als Werkzeug oder Partner.

Die ersten beiden Kapitel untersuchen Fälle, in denen KI-Berater scheinbar externe Empfehlungen geben, also lose mit ihrem menschlichen Nutzer gekoppelt sind. Die zentrale Frage ist, ob externe KI-Berater zu eigenen Handlungen fähig oder ob sie auf reine Werkzeuge zu reduzieren sind. Handlungsfähigkeit hat wesentliche Implikationen: Akteure besitzen nicht nur eine gewisse Handlungsautonomie, sondern werden auch als verantwortlich für ihre Handlungen angesehen.

Kapitel 2 diskutiert die Anwendungsmöglichkeit von klassischen Handlungsbegriffen auf KI-Systeme. Es zeigt, dass die Handlungsfähigkeit von KI-Systemen nämlich weder durch ein enges, menschenähnliches Verständnis von Handlungsfähigkeit noch durch ein weit gefasstes, werkzeugähnliches Verständnis von Handlungsfähigkeit angemessen erfasst werden kann. Das menschenähnliche Konzept der Handlungsfähigkeit,

das auf Davidsons ereigniskausaler Handlungstheorie (Davidson 1963) oder Bratmans Konzept der intentionalen Handlung (M. E. Bratman 2007; M. Bratman and Bratman 1987) basiert, geht davon aus, dass Handlungsfähigkeit intentionale mentale Zustände wie Überzeugungen und Wünsche voraussetzt, die ein beabsichtigtes Verhalten verursachen können. Gezeigt wird, dass KI-Systeme jedoch nicht über intentionale mentale Zustände verfügen und daher nicht als menschenähnliche Agenten betrachtet werden können. Nach dem weit gefassten Verständnis von Handlungsfähigkeit sind KI-Systeme neben einfachen biologischen Organismen grundlegende Akteure, da sie die minimalen Kriterien für Handlungsfähigkeit erfüllen, darunter Individualität, interaktionelle Asymmetrie und Zielgerichtetheit. Das breite Spektrum der KI-Systeme und ihr unterschiedlicher Fähigkeitsgrad zur Interaktion zeigen jedoch, dass keiner der beiden Ansätze die agierenden Fähigkeiten von KI-Systemen in befriedigendem Maße zu differenzieren vermag. Beratende KI-Systeme fordern den Status-Quo des traditionellen Handlungsverständnisses heraus. Statt Werkzeug oder Mensch gibt es nun etwas dazwischen. Damit ergibt sich, dass zumindest konzeptionell KI-Systeme einen ontologischen Verständniswandel in der Anwendung von Handlungsfähigkeit erfordern.

Nach der Unterscheidung zwischen beratenden KI-Systemen und menschlichen Akteuren wird in Kapitel 3 untersucht, wie sich KI-Berater von bloßen Werkzeugen unterscheiden. Während viele Studien erfolgreich aufgezeigt haben, wie die Wahrnehmung in Abhängigkeit von der Rolle der KI und anderen kulturellen oder moralischen Faktoren variiert (Bago 2022; Lim, Rooksby, and Cross 2021; Persson, Laaksoharju, and Koga 2021), stellt Kapitel 3 eine andere Frage: Verhält es sich so, dass jede Erwähnung von KI dazu führt, dass die Menschen die Technologie für mitverantwortlich halten und die Verantwortung vom menschlichen Nutzer abwälzen? Jüngste Studien deuten darauf hin, dass dies in einem hypothetischen Szenario mit moralisch beratender KI der Fall sein könnte (Constantinescu et al. 2022; Giubilini and Savulescu 2018; Malle, Magar, and Scheutz 2019). Relevanter ist jedoch die Frage, was passiert mit alltäglichen KI-Systemen, die lediglich instrumentell genutzt werden? Werden sie ebenfalls als verantwortlich und damit handlungsfähig betrachtet? Um diese Fragen zu beantworten, habe ich mehrere experimentelle Studien durchgeführt – darunter acht Pilotstudien und ein Hauptexperiment. In diesen experimentellen Studien wurde verglichen, was mit den Verantwortungszuschreibungen geschieht, wenn ein menschlicher Fahrer in einer Notsituation eine Warnung von einem KI-gestützten oder einem nicht KI-gestützten Warnsystem erhält. Um sicherzustellen, dass die Verantwortungszuschreibung nicht darauf zurückzuführen ist, dass die KI einige anthropomorphe Merkmale aufweist, habe ich zwei Erscheinungsmodelle der KI verglichen. Die Erscheinungsmodelle beinhalten ein verbales oder ein rein haptisches Warnsystem. Im Einklang mit der moralischen und psychologischen Literatur, die die Bedeutung von Ergebnisverzerrungen und Asymmetrien zwischen Anerkennung und Schuldzuweisung betont, habe ich auch Fälle getestet, in denen die Notsituation erfolgreich bewältigt wurde oder auch nicht. Kapitel 3 zeigt, dass selbst das einfachste KI-System eine Teilung der Verantwortung mit dem menschlichen Nutzer begründet, was in starkem Kontrast zu nicht KI-gestützten Werkzeugen steht. Dieses Ergebnis ist umso überraschender, da die Befragten KI durchaus als Werkzeug betrachten. Die Zuschreibung von Verantwortung an die KI und die Verringerung der menschlichen Verantwortung hängt auch nicht davon ab, wie die KI-Technologie mit dem Benutzer kommuniziert – d. h. über Sprache oder haptische Signale. Darüber hinaus wird die KI eher für gute als für schlechte Ergebnisse verantwortlich gemacht. Sie erhält mehr Anerkennung, wenn der menschliche Fahrer die Situation nach dem Erhalt der KI-Warnung erfolgreich meistert, als dass sie die Schuld erhält, wenn der Fahrer versagt. Insgesamt unterstützt Kapitel 3 die theoretischen Erkenntnisse aus Kapitel 2, indem es feststellt, dass KI-Berater ontologisch gesehen zwischen einem Werkzeug und einem menschlichen Akteur stehen. Mit anderen Worten: Die Arbeit zeigt in den ersten beiden Kapiteln, dass beratende KI-Systeme in ihrer Handlungsfähigkeit und der ihnen zugeschriebenen Verantwortung in einer losen Kopplung mit menschlichen Nutzern tatsächlich eine eigene ontologische Stellung beanspruchen, die mehr ist als ein Werkzeug, aber weniger als ein Mensch.

Die anschließenden zwei Kapitel untersuchen jenen KI-Bereich, bei dem eine ausgesprochen enge Kopplung von KI-Beratern mit ihren menschlichen Nutzern vorliegt. Solche Fälle, in denen KI-Berater integraler Bestandteil der menschlichen Wahrnehmung und Entscheidungsfindung werden. Augmented Reality oder sensorische Augmentation sind hierfür Beispiele. Gegenstand der Untersuchung sind die Fragen, wie und in welchem Ausmaß beratende KI-Systeme die menschliche Wahrnehmung beeinflussen und in welcher Weise sich hochintegrierte KI-Systeme von ihren werkzeuggestützten oder menschlichen Äquivalenten unterscheiden.

Kapitel 4 demonstriert, wie KI die Kopplung von sensorischen Augmentationsgeräten mit dem menschlichen Nutzer verändert. Die Implementierung von KI in bestehende sensorische Augmentationsgeräte, wie z.B. sensorische Substitutionssysteme, verändert nicht nur die konzeptionelle Art der sensorischen Augmentation, sondern erweitert auch die Art der Wahrnehmungsvorverarbeitung vom menschlichen Nutzer auf das KI-System. Aufgrund ihrer umfangreichen Rechenkapazitäten erreichen sensorische KI-Systeme eine völlig neue Qualität bei der Verarbeitung von sensorischen Signalen. Dabei sind zwei Arten der Signalverarbeitung möglich: die Verbesserung von niedrigen sensorischen Signalen durch Herausfiltern von sensorischem Rauschen und die Extraktion von hohen Wahrnehmungsmerkmalen durch die Einbeziehung von Datenverarbeitungswerkzeugen in den sensorischen Erweiterungsprozess. Nachdem gezeigt wurde, wie KI in sensorische Erweiterungsprozesse integriert werden kann, fragt Kapitel 4, ob sensorische KI-Systeme folglich als Wahrnehmungserweiterungen verstanden werden sollten. In Bezug auf die erweiterten Wahrnehmungssysteme von biologischen Systemen wie Fledermäusen und elektrischen Fischen kommt das Kapitel zu dem Schluss, dass sensorische KI-Beratungssysteme einzigartig sind und die menschliche Wahrnehmung in einer Weise erweitern, wie es kein nicht KI-gestütztes Gerät vermag.

Kapitel 5 stellt die sensorische Kopplung Mensch und KI der wahrnehmungsbezogenen Kopplung gegenüber, die die sozialen Interaktionen des Menschen bestimmt.

Kapitel 5 zeigt, wo eine Mensch-KI-Kopplung nicht ausreicht, um ein menschenähnliches Maß an sozialem und wahrnehmungsbezogenem Einfluss und Koordination zu erreichen. Zwei weithin untersuchte Formen sozialer Interaktion sind das gemeinsame Handeln und die gemeinsame Aufmerksamkeit. Bei beiden Formen der sozialen Interaktion geht es um mehr als die Koordinierung von Handlungen und Aufmerksamkeit. Stattdessen entwickeln menschliche Akteure ein gegenseitiges Bewusstsein für die Ziele, Absichten und Handlungen des anderen, was nicht nur die individuelle Erfahrung, sondern auch die kollektive Handlung verändert. In einem Orchester zu spielen, Handlungen im Team durchzuführen oder einfach nur einen Tisch gemeinsam zu bewegen sind dynamische, voneinander auf höchstem Niveau abhängige Handlungen und Erfahrungen. Nach einer Darstellung der Forschung und der Mechanismen, die der gemeinsamen Aufmerksamkeit zugrunde liegen, geht Kapitel 5 darüber hinaus und liefert neue Erkenntnisse über die Mechanismen der gemeinsamen Wahrnehmung. Die gemeinsame Wahrnehmung unterscheidet sich eindeutig von der gemeinsamen Aufmerksamkeit, da die gegenseitige Wahrnehmung ohne die Verfolgung von körperlichen Hinweisen, wie z. B. den Blick, erfolgt, sondern vielmehr durch die gegenseitige Kenntnis eines gemeinsamen Wahrnehmungsbereichs. Dennoch bleiben ähnliche Leistungsvorteile – schnellere und genauere Wahrnehmungsverarbeitung – in einer gemeinsamen Wahrnehmung bestehen. Soziale Koordination und Sensibilität für soziale Hinweise sind – wenn überhaupt – noch rudimentäre Bausteine in KI-Beratungssystemen. Bei KI-Systemen wurde die Verbesserung der Zusammenarbeit zwischen Menschen und KI-gesteuerten Systemen bisher hauptsächlich aus einer technischen Perspektive betrachtet, bei der die Roboterbewegungen sicher sein und auf grundlegende Formen der menschlichen Interaktion reagieren müssen, um gegebene Befehle umzusetzen (siehe Liang et al. (2021) und Liu and Wang (2018) für einen Überblick). Die soziale Koordination bleibt eine eindeutige menschliche Eigenschaft.

Beide Kapitel bekräftigen somit, dass beratende KI-Systeme auch in einer engen Kopplung ihre eigene ontologische Kategorie als etwas fordern, das leistungsfähiger ist als ein Nicht-KI-Werkzeug, aber immer noch hinter den menschlichen Standards zurückbleibt.

Insgesamt trägt diese Arbeit zu einem umfassenderen konzeptionellen Verständnis bei, wie KI-Berater zu definieren und wie sie mit ihren menschlichen Nutzern gekoppelt sind. Die bestehende Literatur über KI-Systeme wird mit dieser Arbeit in entscheidenden Punkten ergänzt. Die Erkenntnis, dass beratende KI-Systeme eine einzigartige ontologische Kategorie darstellen – etwas zwischen Nicht-KI-Werkzeugen und menschlichen Akteuren – hat nicht nur Auswirkungen auf praktische Fragen, wie beratende KI-Systeme behandelt werden sollten, sondern auch auf philosophische Debatten darüber, was es bedeutet, ein beratendes KI-System zu sein. Zukünftige Arbeiten sollten einerseits die Verantwortungsdynamik erforschen, die durch KI-Berater eingeführt wird und andererseits mehr soziale Sensibilität in KI-Wahrnehmungsunterstützungssysteme integrieren, um die Voraussetzungen für menschenähnliche Formen der Zusammenarbeit zu schaffen.

Summary

Artificial intelligence (AI)'s successes are widely discussed and have changed how people work and think about the world around them. OpenAI's ChatGPT language model can produce human-like text, DeepMind's AlphaZero can beat any human player in chess or Go, and Tesla's driving AI can pilot cars and drones. Many open questions remain: Who is responsible if something goes wrong? Are AI systems capable of action? Moreover, are some AI systems possibly sentient? The research field of AI incorporates many different approaches ranging from psychological studies on human-AI interaction to technical innovations in AI learning and philosophical discussions on AI's mental and moral capacities.

The main focus across these research fields has been on (semi-)autonomous AI systems – systems that operate in large parts without human instructions. Consider debates on driving cars and language models. What is often left unexplored are AI systems that are closely coupled with their human users – systems that keep humans in the decision-making loop by informing or recommending actions or decisions. Investigating this gap is becoming increasingly crucial as incidents of AI-assisted decision-making become more common – consider low-stakes decisions such as shopping recommendations and high-stakes decisions such as medical diagnosis. This PhD thesis examines AI advisers and substantiates the existing functional distinction – looking at what AI systems do and are used for – with a conceptual analysis – of what AI systems are. A conceptual analysis of AI advisers is novel and closes a critical gap in the literature. A conceptual analysis of AI advisers can tell us whether AI advisers are more than tools and what distinguishes them from human advisers.

This PhD thesis consists of two main parts. The first part examines the loose coupling of AI advisers with their human users – cases where AI advisers provide seemingly external recommendations. The thesis asks whether external AI advisers have some agentivity or are mere tools. Being an agent has essential implications: not only are agents considered responsible for their actions, but also agents possess a certain degree of autonomy. Therefore, chapter 2 applies different theoretical notions of agency to AI advisory systems. Chapter 3 complements the theoretical analysis with an experimental evaluation of responsibility attribution in a human-AI advisory setting.

Chapter 2 of this PhD thesis discusses the possibilities for AI agency. It shows that AI systems necessitate an ontological shift in how agency is understood and applied. While AI advisers satisfy the requirements for minimal agency, the agentive capacity of AI systems can be adequately captured neither by a human-like nor by a minimal concept of agency. The human-like concept of agency, based on Davidson's event-causal theory of action (Davidson 1963) or Bratman's notion of intentional action (Bratman 2007; Bratman 1987), holds that agency requires intentional mental states like beliefs and desires that can cause an intended behaviour. However, as argued in Chapter 2, AI

systems lack intentional mental states and cannot be seen as human-like agents. On the minimal understanding of agency, AI systems, alongside simple biological organisms, are basic agents as they fulfil the minimal criteria for agency, including individuality, interactional asymmetry, and goal-directedness. However, the wide range of existing AI systems and their varying degrees of agentive abilities demonstrates a mismatch with either approach – as neither approach can differentiate the agentive capacities of AI systems. I argue that, instead, AI advisers are something in between that only a gradual notion of agency can capture.

Having differentiated AI advisers from human agents, chapter 3 aims to confirm how AI advisers differ from mere tools. While many studies have successfully mapped how people's opinion varies depending on the role of AI and other cultural or moral factors (Bago 2022; Lim, Rooksby, and Cross 2021; Persson, Laaksoharju, and Koga 2021), chapter 3 asks a different question: is it the case that any mention of AI will lead people to see the technology as partly responsible and shift the responsibility away from the human user? Recent studies suggest this may be the case under the hypothetical scenario where AI provides moral guidance (Constantinescu et al. 2022; Giubilini and Savulescu 2018; Malle, Magar, and Scheutz 2019). However, it is more relevant to ask if this would happen under AI's more prevalent day-to-day usage when it merely provides factual information and is used purely instrumentally. I conducted multiple experimental studies to address these questions – including eight pilot studies and a main experiment. Across these experimental studies, chapter 3 compared what happened to responsibility attributions when a human driver, faced with an emergency, receives a warning from an AI-powered or non-AI-powered warning system. To ensure that the attribution of responsibility does not come from the AI sharing some anthropomorphic features, I compare situations in which the AI was a voice assistant or a haptic warning system. In line with the moral and psychological literature stressing the importance of outcome biases and asymmetries between credit and blame, I also tested cases where the emergency was successfully managed. Chapter 3 finds that even the most basic AI system introduces a sharing of responsibility with their human user, in sharp contrast to non-AI-powered tools. This finding is all the more surprising because, when asked, people did recognise AI as a tool. Attributing responsibility to AI and reducing human responsibility also does not depend on how the AI technology communicates with the user – i.e. via voice or haptic signals.

Furthermore, the AI is seen as more responsible for good rather than bad outcomes, as it gets more credit when the human driver successfully negotiates the situation after receiving the AI warning than it receives blame when the driver fails. Together, chapter 3 supports the theoretical findings from Chapter 2, establishing that AI advisers are ontologically more than tools but less than human agents. In other words, the thesis finds that in their agentive capacity and attributed responsibility, AI advisers, in a loose coupling with human users, indeed demand their own ontological space as something more than tools but less than humans.

The second part analyses a tighter coupling of AI advisers with their human users – cases where AI advisers become integral to human perception and decision-making. Consider cases of augmented reality or sensory augmentation. Here, the thesis asks how and to which extent AI advisers influence human perception and in which way highly integrated AI systems differ from their tool or human counterparts. Chapter 4 discusses how non-AI and AI-powered sensory augmentation devices differ and addresses whether sensory AI systems represent perceptual extensions of their human users. Chapter 5 asks what tight coupling in a human-human context is capable of – consider cases of joint attention and shared perception – and points out future directions for a more socially attuned version of a human-AI coupling. The thesis finds that, also in a tight coupling, AI advisers demand their unique ontological category as something more capable than a non-AI tool but still falling short of human standards.

Chapter 4 demonstrates how AI changes the coupling of sensory augmentation devices with the human user. Implementing AI into existing sensory augmentation devices, such as sensory substitution systems, changes the conceptual kind of sensory augmentation and extends the kind of perceptual pre-processing from the human user to the AI system. Due to their extensive computational capacities, sensory AI systems can process sensory signals like no other sensory augmentation system before. Two ways of signal processing are possible: enhancing low-level sensory signals by filtering out sensory noise and extracting high-level perceptual features by incorporating data-processing tools in the sensory augmentation process. After showing how AI can be incorporated into sensory augmentation processes, chapter 4 asks whether, consequently, sensory AI systems should be understood as perceptual extenders. About the extended perceptual systems of biological systems like bats and electric fish, the chapter concludes that sensory AI advice systems are unique and extend human perception in ways no non-AI-powered device can.

Chapter 5 contrasts the sensory human-AI coupling with the perceptual coupling driving human social interactions. Chapter 5 shows where a human-AI coupling falls short of achieving human-like social and perceptual influence and coordination levels. Two widely studied forms of social interaction are joint action – doing things together – and joint attention – attending to things together. Both forms of social interaction are more than coordinating actions and attention. Instead, human agents develop a mutual awareness of each other's goals, intentions, and actions, transforming not only the individual experience but also the collective action. Playing in an orchestra, performing team surgeries, or simply moving a table together are at their highest level, dynamic, mutually dependent actions and experiences. After outlining the research and mechanisms behind joint attention, chapter 5 goes even further and provides novel insights into the mechanisms of shared perception. Shared perception uniquely differs from joint attention as mutual awareness occurs without tracking bodily cues, such as gaze, but rather through mutual knowledge of a perceptual common. However, similar performance benefits – faster and more accurate perceptual processing – in a joint setting persist.

Social coordination and sensitivity to social cues are still – if at all – rudimentary building blocks in AI advisory systems. For AI systems, improving collaboration between humans and AI-powered systems has been mainly addressed from an engineering perspective, where robot movements must be safe and sensitive to basic forms of human interaction to realise given commands (see Liang et al. (2021) and Liu and Wang (2018) for review). Beyond that, social coordination remains a uniquely human trait.

Overall, this thesis contributes to a richer conceptual understanding of what AI advisers are and how they are coupled with their human users to the existing literature on AI systems. The finding that AI advisers represent a unique ontological category – something between non-AI tools and human agents – impacts not only practical issues on how advisory systems should be treated but also philosophical debates on what it means to be an AI adviser. Future work should, on the one hand, explore the responsibility dynamics introduced by AI advisers – as they are uniquely praised but not blamed for an outcome – and, on the other hand, integrate more social sensitivity in AI perceptual support systems to set the stage for human-like forms of collaboration.

Introduction

1 Introduction

Digital technologies are reshaping everything, including customer behaviours and expectations, organisational and manufacturing systems, business models, markets, and ultimately society, for better or worse. Digitalisation – the implementation of digital technologies (Setia et al. 2013) – has provided both major opportunities and significant challenges to individuals, organisations, ecosystems, and entire societies. At the core of such transformative trends are digital technologies, broadly defined as combinations of "information, computing, communication, and connectivity technologies" (Bharadwaj et al. 2013, 471). The concept of digital transformation has been widely used to describe the adoption of digital technologies and the replacement of non-digital processes with digital ones, leading to organisation-wide changes and the emergence of new business models (Verhoef et al. 2021) or the modification of existing ones (Dąbrowska et al. 2022). At its inception, digital transformation was predominantly discussed in the information systems literature (Nadkarni and Prügl 2021), focusing on its technological aspects, such as optimising organisational operational processes (Vial 2019). Nowadays, digital transformation extends beyond informational tool use to capture the wide-reaching socioeconomic change across individuals, organisations, ecosystems, and societies shaped by adopting and utilising digital technologies (Dąbrowska et al. 2022).

Artificial intelligence (AI) has been one of the main drivers behind the surge of digital technologies and digital transformation. In fact, artificial intelligence (AI) is widely present in our everyday lives. Individual users now rely on AI support for daily decisions such as shopping and movie recommendations but also depend upon AI to facilitate perception and decision-making in high-stakes environments such as medical diagnostics and driving support. Regardless of the environment, AI systems significantly shape how we think about and perceive the world.

Understanding AI's degree and kind of influence on human perception and decision-making become paramount for guiding the critical and responsible use of digital technologies and digital transformation at large. So, how does AI influence human perception and decision-making?

Much research – within various fields – has examined the coupling of autonomous and embodied AI systems with their human user. Computer scientists have focused on improving algorithmic performance, as the transition from shallow to deep learning architectures shows (Schmidhuber 2015). By adopting deep learning techniques, computers can do things without being explicitly programmed, constructing algorithms that adapt their functions from data and producing decisions or predictions. Several exciting results have stirred much attention to deep learning during this decade.

Game-playing systems like AlphaGo and AlphaZero have achieved super-human performance (Silver et al. 2018, 2018); language processing models can produce text indistinguishable from human prose; real-time image processing is the foundation for autonomous vehicles.

The field of AI ethics has focused on understanding the social impact AI systems have on their human users and the ethical consequences AI systems introduce to others. Supporting hiring decisions by AI-powered algorithms, moral decisions in autonomous driving cars, and racist and sexist chatbots are just a few ethical challenges AI systems introduce. While some have focused on implementing moral decision-making capacities in AI systems in the sub-field of machine ethics (see, for example, M. Anderson and Anderson (2011); Wallach, Franklin, and Allen (2010)), others have debated the ethical use of autonomous systems – including, among others driving vehicles or weapon systems – and the ensuing consequences for how responsibility is allocated (see Purves, Jenkins, and Strawser (2015); Danaher (2018); Nyholm (2018)). The key challenge here is to develop a way of understanding moral responsibility in the context of autonomous systems that would allow us to secure the benefits of such systems and, at the same time, appropriately attribute responsibility for any undesirable consequences.

The field of philosophy of mind has debated whether and when AI systems would have minds or human-like mental states. The debates often go beyond a purely behavioural analysis and ascription of mental states based on observed behaviour (see Dennett's intentional stance as one common behaviourism approach). Instead, they extend to whether AIs could become conscious and whether or under which conditions they should be considered agents or even persons.

The psychological literature – especially in the domain of human-AI interaction – has explored what the use of AI systems means/changes for the human user. Which features improve human acceptability, and how can AI systems provide the most value for their human user – thinking about the timeliness of their recommendations and usable interface?

So far, I have argued that AI has been a primary driver for digital transformation and that studying AI is essential to mitigate any unwanted side effects – as different research fields have done -but I have not stated what AI is. The proposed definitions of AI are as numerous as divergent. AI systems can be embodied, non-embodied; learning; non-learning; self-centred, human-serving; static, or dynamic. Providing an adequate structured definition is, therefore, a challenge. The traditional approach in distinguishing AI systems is teleological – highlighting the goals based on which AI systems are developed. Some seek to mimic human-like behaviour or decision processes in AI systems; others want to develop rational tools which reliably produce rational behaviour or reasoning.

Researchers advocated using human-like processes as inspiration for designing AI systems to implement human-like behaviour or thinking processes. Advocating for human-like behaving systems, Minsky, for example, argued that the field of AI is the "science of making machines capable of performing tasks that would require intelligence if done by humans" (Minsky 1968). Similarly, Copeland and Shagrir (2020) argued that AI systems are digital computer or computer-controlled robots which can perform tasks commonly associated with intelligent beings – such as the ability to reason, discover meaning, generalise, or learn from past experience. Others go even further and want to reproduce the processes, representations, and results of human thinking on a machine. Neuro-inspired computing or computational architectures are just two examples where human-brain-inspired computational models are used to achieve higher computational performance and biological plausibility (Bolotta and Dumas 2022; Kotseruba and Tsotsos 2020; W. Zhang et al. 2020).

Other researchers have focused on developing AI systems as rational, rule-following systems – an approach popular in technical fields like computer science and robotics, where consistent and predictable behaviour is vital for system performance. Therefore, AI systems are equipped with rational planning and decision-making capacities. Backed by a logicist tradition, AI systems as rational thinking agents allow for rigorous reasoning and, ideally, a comprehensive model of rational thought. Tracing decisions back to their logical origin greatly appeals to AI systems that demand high transparency in their decisions – as used in legal and moral decision-making. Russell and Norvig define AI as "the study of intelligent agents that receive precepts from the environment and take action. Each such agent is implemented by a function that maps percepts to actions, and we cover different ways to represent these functions, such as production systems, reactive agents, logical planners, neural networks, and decision-theoretic systems" (Russell and Norvig 2016, viii).

Besides a teleological distinction of AI systems, another approach for distinguishing AI systems is to examine their function – how AI systems are used. For example, the type of interaction between the AI system and the human user distinguishes an AI advisor from an AI partner. An AI advisor merely influences others by making recommendations or providing information, whereas an AI partner goes above and beyond. An AI partner participates in the action and works with the human user to achieve a goal (Köbis, Bonnefon, and Rahwan 2021). Consider driving a car: while the AI advisor can only warn the human driver about an impending accident, the AI partner can take control and slam the brakes autonomously. Two primary use cases emerge: autonomous/automated and coupled/assistive systems. While autonomous AI systems can operate independently from human supervision – once a particular behaviour or computational process is learned – consider chess computers playing chess without any human directions – advisory systems are coupled to a human advisee.

Automated AI systems draw on growing amounts of digital data and advances in machine learning techniques to fulfil delegated decision-making tasks with relative autonomy. Automated AI systems can adapt online advertising based on individual online behaviour (Boerman, Kruikemeier, and Zuiderveen Borgesius 2017), generate music or art, and even drive cars. Often these systems are perceived as separate entities with high degrees of decision autonomy. Although people are averse to autonomous agents making moral decisions, whether in the military, law, driving, or medical settings (Bigman and Gray 2018; Gogoll and Uhl 2018), most people would consider autonomous agents morally responsible for unexpected outcomes because of their moral decisions (Kahn et al. 2012; Malle, Magar, and Scheutz 2019; Shank, DeSanti, and Maninger 2019). Judgments of moral responsibility hinge on autonomy (Bigman et al. 2019; Kahn et al. 2012; T. Kim and Hinds 2006) and mind perception, e.g., ascribing mental abilities to think and feel to automation (Bigman and Gray 2018; Bigman et al. 2019).

Next to automating decision-making, AI systems can also be closely coupled with the human user. Here, AI systems do not make decisions on their own but rather support the human user. Coupled AI systems can have two main usage modes: informing or recommending output (Gundersen and Bærøe 2022; Gundersen 2018). In an informing mode, AI provides additional information to facilitate human decision-making – providing analyses, estimating probabilities, or predicting events from past data. Well-studied and heavily debated ethical issues arise: lack of transparency, explainability and accountability of AI. In a recommending mode, AI is used to directly influence what people should do, like turning left when driving a car or assigning medical treatments based on a patient's health record.

While the teleological distinction is widely applied to the field of AI, implemented AI systems – as they are used by human individuals – are discussed by virtue of their function. In fact, most social and scientific debates have focused on automated or autonomous AI systems. For example, researchers have developed a wide range of autonomous AI systems – driving cars and large language models – and found trends of human-like responsibility attribution – self-driving cars are held responsible for their behaviour. This PhD project does not seek to extend the literature on autonomous AI systems but instead focuses on a yet underdeveloped field of advisory AI systems. It examines theoretically and empirically the influence of advisory AI systems on human perception and decision-making. Human-coupled advisory AI systems, which are even more prevalent in the current field of AI decision support tools like risk assessment for financial lending (Green and Chen 2019) or visual object detection, have yet to be investigated (Bondi et al. 2021). While an autonomous AI partner's agential and moral roles are easier to distinguish from those of its human counterpart, the influence of advisory AI systems is more difficult to discern but no less relevant to the overall outcome. As an example, consider medical diagnostics. With super-human accuracy in image-based medical diagnostics, advisory AI systems can provide precious but often opaque medical advice to human clinicians. It is difficult to determine who is responsible for the final medical diagnosis. For fully autonomous robotic surgery, on the other hand, the AI system shares responsibility with the supervising clinician (McManus and Rutchick 2019; O'Sullivan et al. 2019).

With a range of different capabilities and implementations, AI systems occupy a unique social role. They can do more than tools but less than humans. A basic tool does not possess any independent processing or goal-directed behaviour, whereas humans are the pinnacle of independent processing and goal-directed behaviour. While basic tools entirely depend on their functioning on a human user, human agents function independently. Some AI systems are closer to tools – consider automated vacuum cleaners, whereas others are closer to human agents – consider super-human game-playing engines. Notably, the main driver for the difference in AI systems is the degree of independent processing taken on by the AI system. While vacuum cleaners process only a limited number of gathered sensory information – in a way that the human developers entirely dictate – super-human game-playing engines learn to develop their own strategies to supersede human performance. Ultimately, a conceptual grey zone of AI systems emerges, where the perceived capabilities dictate the AI's ontological status (See Figure 1.1.). So far, the boundaries between tools and humans for AI advisers are blurred. Some researchers have claimed that AI advisers are human-like (Y. Tian et al. 2017; Pelau, Dabija, and Ene 2021), whereas others reduce AI advisers to mere tools (Gunkel 2012; Zheng and Wu 2019).

This PhD thesis examines AI advisers and substantiates the existing functional distinction – looking at what AI systems do and are used for – with a conceptual analysis – of what AI systems are. Investigating this gap is becoming increasingly important as incidents of AI-assisted decision-making become more common – consider low-stakes decisions such as shopping recommendations and high-stakes decisions such as medical diagnosis. A conceptual analysis of AI advisers is novel and closes a critical gap in the literature. A conceptual analysis of AI advisers can tell us whether AI advisers are more than tools and what distinguishes them from human advisers.

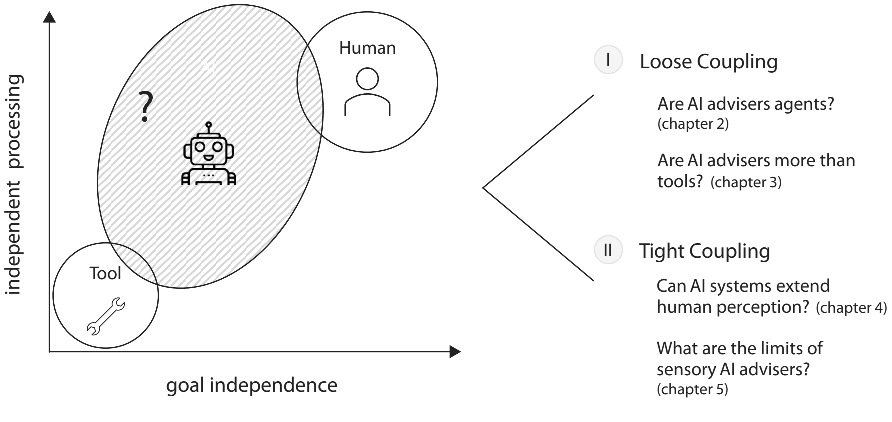
This thesis follows two main parts towards developing a rich conceptual understanding of AI advisers. The first part examines the loose coupling of AI advisers with their human users – cases where AI advisers provide seemingly external recommendations. The thesis asks whether external AI advisers are agents, i.e., capable of action or mere tools. Being an agent has essential implications: not only are agents considered responsible for their actions, but also agents possess a certain degree of autonomy/independence. Therefore, Chapter 2 applies different theoretical notions of agency to AI advisory systems. Chapter 3 complements the theoretical analysis with an experimental evaluation of responsibility attribution in a human-AI advisory setting.

Chapter 2 of this PhD thesis discusses the possibilities for AI agency. It shows that AI systems necessitate an ontological shift in how agency is understood and applied. While AI advisers satisfy the requirements for minimal agency, the agentive capacity of AI systems can be adequately captured neither by a human-like concept of agency nor by a minimal concept of agency. Instead, AI systems are something in between that only a gradual notion of agency can capture.

Having differentiated AI advisers from human agents, chapter 3 aims to confirm how AI advisers differ from mere tools. Do sensory AI systems share responsibility with their human user? Or are sensory AI systems perceived as tools, void of responsibility attribution? Chapter 3 finds that even the most basic AI system introduces a sharing of responsibility with their human user, in sharp contrast to non-AI-powered tools. This finding is even more surprising because, when asked, people did recognise AI as a tool. Attributing responsibility to AI and reducing human responsibility also does not depend on how the AI technology communicates with the user – i.e. via voice or haptic signals.

Furthermore, the AI is seen as more responsible for good rather than harmful outcomes, as it gets more credit when the human driver successfully negotiates the situation after receiving the AI warning than it receives blame when the driver fails. Together, chapter 3 supports the theoretical findings from Chapter 2, establishing that AI advisers are ontologically more than tools but less than human agents. In other words, the thesis finds that in their agentive capacity and attributed responsibility, AI advisers, in a loose coupling with human users, indeed demand their own ontological space as something more than tools but less than humans.

The second part analyses a tighter coupling of AI advisers with their human users – cases where AI advisers become integral to human perception and decision-making. Consider cases of augmented reality or sensory augmentation. Here, the thesis asks how and to which extent AI advisers influence human perception and in which way highly integrated AI systems differ from their tool or human counterparts. Chapter 4 discusses how non-AI and AI-powered sensory augmentation devices differ and addresses whether sensory AI systems represent perceptual extensions of their human users. What is the nature of the coupling? Does AI take on parts of the perceptual process, or is AI a mere sensory extension? Chapter 5 asks what tight coupling in a human-human context is capable of – consider cases of joint attention and shared perception – and points out future directions for a more socially attuned version of a human-AI coupling. The thesis finds that, also in a tight coupling, AI advisers demand their unique ontological category as something more capable than a non-AI tool but still falling short of human standards.



Are AI advisers agents? (chapter 2)

Are AI advisers more than tools? (chapter 3)

Can AI systems extend human perception? (chapter 4)

What are the limits of sensory AI advisers? (chapter 5)

Figure 1: Dissertation structure and research question

Overall, this thesis contributes to a richer conceptual understanding of what AI advisers are and how they are coupled with their human users to the existing literature on AI systems. The finding that AI advisers represent a unique ontological category – something between non-AI tools and human agents – impacts not only practical issues on how advisory systems should be treated but also philosophical debates on what it means to be an AI adviser. Future work should, on the one hand, explore the responsibility dynamics introduced by AI advisers – as they are uniquely praised but not blamed for an outcome – and, on the other hand, integrate more social sensitivity in AI perceptual support systems to set the stage for human-like forms of collaboration.

Part 1

2 AI as co-agents

2.1 Introduction

The question of whether a machine could think and act like a human has accompanied the research of artificial intelligence (AI) ever since the development of Turing's theory of computation and its associated computational architectures. With the resurgence of neurocomputational learning techniques and advances in computational power, this question has become more prominent than ever. Deep learning systems can now reach human-level performance in various domains such as image recognition, game playing and driving (Ghahramani 2015). The Go-playing system AlphaGo (Silver et al. 2017) and humanoid robot SOPHIA (Goertzel et al. 2017) can be seen as paradigm cases for the evolution of AI. While AlphaGo has beaten the world-leading human Go player Lee Sedol through self-learning without implemented human knowledge, SOPHIA has been granted Saudi-Arabian citizenship.

In AI research, the concept of agency as applied to technical artefacts has become a subject of intense discussion. Some have argued that agency is conceivably everywhere. Every entity which engages in causal relationships and interacts with its environment can plausibly be identified as an agent. This notion captures our general intuition of treating interactive entities as seemingly autonomous systems. This notion of agency is considered minimal because it extends to a class of possible agents to a wide range of entities ranging from animals to movie characters, legal personas or even the Go-playing AlphaGo and the humanoid robot SOPHIA.

Others argued that agency is much more restrictive. Agency in the restricted sense is commonly identified with human agency and has been understood by Davidson (1963); Davidson (1982); Davidson (2001), M. Bratman and Bratman (1987); M. Bratman (1999) and Anscombe (2000) as the capacity to perform intentional actions, which represents the standard philosophical conception of agency. On this interpretation, the Go-playing AlphaGo or the humanoid robot SOPHIA do not classify as agents because they lack human-like intentionality and other internal mental states.

Even others have argued that neither of the previous accounts can adequately capture what it means to be an AI agent. As AI systems vary enormously in their capacities to learn, initiate, and perform an action, mapping agentive differences to different agency standards represents arguably the most plausible account of AI agency. For instance, Floridi and Sanders (2004) proposed a method of abstraction. The method of abstraction seeks to extend the class of possible agents by postulating different levels of abstraction under which an entity can be conceived as an agent. If all three criteria of interactivity, autonomy and adaptability are fulfilled at some level of abstraction, an entity can be understood as an agent at this level of abstraction.

So, what are AI advisers? Agents in a minimal sense or even agents in a human sense? Or are AI advisers mere tools bare any agentive capacity? Answers to these questions not only determine the ontological role AI advisers have in behavioural interactions but also influence legal and ethical debates where moral responsibility hinges on the capacity of being an agent – as only agents can be considered responsible for their actions. This chapter addresses these questions by applying minimal and human agency standards to AI advisers. After showing that AI advisers only satisfy minimal agency standards, I explore other concepts of agency which can better match the wide range of different AI systems. Terminologically, the rise of AI advisers challenges the traditional conceptions of agency in two ways. AI advisers invoke top-down pressure to open up restricted, human notion of agency to account for the behaviour (semi-) autonomous AI systems. AI advisers also invoke a bottom-up pressure to restrict the minimal notion of agency further to distinguish the perceived differences in AI systems.

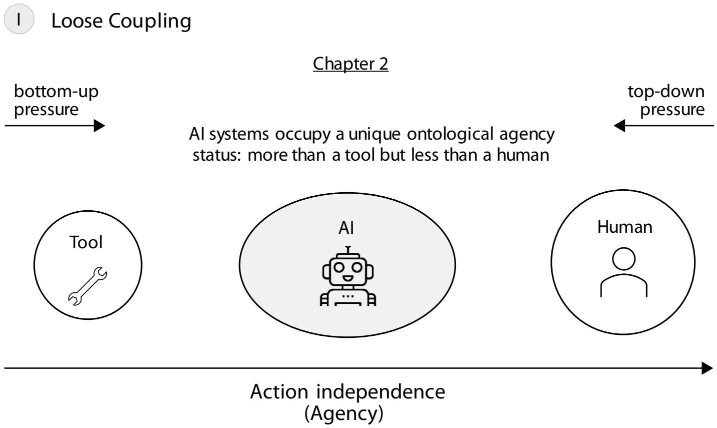

I
Loose Coupling
Chapter 2
bottom-up pressure
top-down pressure
AI systems occupy a unique ontological agency status: more than a tool but less than a human
Tool
AI
Human
Action independence (Agency)

Figure 2: Overview chapter 2

2.2 Top-down pressure: not quite human

The standard story of what it means to be an agent starts with Davidson. He defined an agent as a system capable of action, and something counts as an action if it is done intentionally, for reasons and is caused by the right mental states of beliefs and desires in the right way. Agency, in this sense. is reductive, and it reduces the agent's role in the exercise of agency to the causal roles of agent-involving states and events, where actions can be reduced to pairs of agent-involving mental states and events. Intentional actions are events, which means that actions are particulars in space-time. Observed particular in space-time (behaviour) can be interpreted differently, i.e., can be explained by different descriptions. While grasping for a bottle of water might be part of getting a drink, it might also be part of stretching one's arm. The distinguishing factor for those different descriptions is the preceding mental states of action. Because I intended to grasp a bottle of water, this behaviour represents my intentional action, while stretching is only an unintentional action description. This agency model works well for understanding and explaining human action, allowing us to distinguish actions – like grasping water – from mere behaviour – reaching out the arm.

According to the standard story of the agency told by Davidson, for an AI system to be considered a human-like agent, it must initiate something caused by intentions and represent the reasons for the possible action. The intentional mental states represent the reasons for an account such that an action can be explained by revealing one's intentional states. Agents hence cause their actions by virtue of their intentional states. As argued by M. Bratman (1984), intentional action involves a specific kind of intentional state called "intention." The need to invoke intentions is two-fold. First, agents have limited mental or computational resources. They cannot constantly weigh their conflicting desires and beliefs when deciding what to do next. Instead, an agent must, at some point, commit to a particular state of affairs to pursue. Second, future action planning requires intentions. If an agent has selected a future action, he or she must form subsequent intentions to carry out the action at a specified time and in a specified manner. For example, the intention to clean the floor invokes more specific intentions about avoiding obstacles and checking the vacuum bag. The agent puts together these sequences of action to achieve one or more of its intentions. Practically, intentions should therefore be internally coherent and coherent with the agent's other beliefs. Also, the agent should be able to monitor whether he/she completed the intentional action successfully. An intentional agent is not just able to control his or her behaviour but must also have a certain amount of control over his or her mental state.

As pointed out by Misselhorn (2015), higher-order intentionality entails intentional states with other intentional states as their object, such as beliefs about beliefs, desires, or desires about beliefs. Intentional states with other intentional states characterise higher-order intentionality as their objects, such as beliefs about beliefs, desires, or desires about beliefs. Higher-order intentionality permits an agent to form beliefs and desires based on his or her mental state and the mental states of other agents. Higher-order intentionality creates a qualitative distinction in the agency, and it is widely regarded as a crucial component for understanding free will, which is a highly demanding form of autonomy. According to Harry Frankfurt, human action differs from animal behaviour in that we can reflect on our beliefs and desires by forming desires and beliefs about them (Frankfurt 1971). I may have a first-order desire to consume a piece of cake but a second-order desire to abstain from eating the cake out of concern for my weight gain. The distinction between first- and higher-order international states can be used to explain freedom of will and freedom of action: An action is free when the causally relevant desire is the one I want to be effective. Explaining freedom of the will is the capacity to choose which first-order desire to act on. Despite my initial desire for the cake, I could have chosen to eat an apple instead of maintaining my weight was more important to me. Higher-order desires are those with which an individual identifies; they reveal the individual's true nature. People like drug addicts lack willpower because they cannot satisfy their second-order desire. Freedom of will is commonly regarded as a defining characteristic of a person.

Even though the demands for agency set out by the philosophical tradition are hard to fulfil by an artificial system, some AI researchers have taken up the challenge. If a robot were able to "replicate the human decision-making process" (Purves, Jenkins, and Strawser 2015, 855), "make life and death decisions" (Wallach and Allen 2009, 14), make "split-second decisions" (ibid) or "choose their own targets" (Sparrow 2007, 70), the robot would be an agent (Nyholm 2018). If an AI system was capable of intentionality, then we have good grounds on which it could have mental states and hence be conceived as a human-like agent.

The concept of intentional states can be understood both internally and externally (Powers 2013). External states represent intentional states expressed outside the artefact, whereas internal states are necessarily mental. According to an externalist interpretation, AI systems express external, intentional states due to their symbolic interaction with the environment. AI systems can utter meaningful content through speech acts or written expressions and express external, intentional states. Consequently, intentionality is externally determined as the capacity to be about representational content. This also applies to deliberate actions since they depend on the capacity to have representational states. Similarly, to a human agent who utters, "It is raining outside," an AI-based voice assistant can utter the exact words and communicate the same or at least some external intentional states to us. In this sense, intentionality is ascribed to the system based on its interaction with the user. The human developer who implements the representational content of the AI system and presents it to the user determines the appearance of having intentional states. From an internalist view, sophisticated AI systems can be considered to have genuine internal mental states, such that they act based on their motivations and intentions. While it remains impossible to establish proof of such capabilities due to the inability to access the mental states of others beyond testimony, Powers (2013) argues that it is unjustified to rule out the possibility of complex AI systems having incomprehensible mental states for humans. It is commonly argued that intentional states cannot exist in computational systems because they are neither hardware nor software based.

Nonetheless, Powers (2013) demonstrates that such an objection commits a fundamental ontological error by comparing mental states to physical states. Even for human moral agency, the origin and mechanism of moral agency are unknown. Motivation for action is always an abstraction, unlike any physical state. Reasons are non-reductive and comparable to other high-level mental states in humans, such as consciousness and conceptual comprehension. Therefore, accepting the possibility that computers have their motives is contingent on the existence of intentional states that transcend physical implementation.

Personhood Besides human-like agency, AI systems have also conquered the news circle by being portrayed as legal persons (Yampolskiy 2021). Legal scholars who have opened this de-bate have, at large, advocated for accepting autonomous systems as legally recognised persons for agency law (Jaynes 2019). Philosophically, being a person is much more restricted and represents one of the most demanding forms of agency (Misselhorn 2015). The most influential account of personhood has been provided by Dennett (1976), who proposes six essential conditions for personhood:

1. Persons must be rational beings.
2. They must have intentionality.
3. One can take a particular stance or attitude towards persons.
4. They must be capable of reciprocating in some way.
5. They must be capable of verbal communication.
6. They must have self-awareness.

The initial three conditions are interdependent: Being rational merely entails possessing intentionality, and for Dennett, this is a matter of being the object of a particular position, the intentional position. According to this perspective, an object is an intentional system if it makes sense to interpret its behaviour by ascriptions of intentional states such as beliefs and desires. These three conditions apply to more than personhood; they define the entire class of intentional systems. Persons must also meet the other three requirements. Dennett defines reciprocity as the capacity to take an intentional stance towards other systems, corresponding to a higher order of intentionality. In addition, individuals must be able to communicate with one another and possess self-awareness. Notably, the final condition applies only to human agents. Dennett defines self-consciousness as the capacity to reflect on one's beliefs and desires. Although higher-order intentionality may be sufficient for self-consciousness, Dennett believes this is untrue. The difference is that predetermined higher-order desires do not constrain a person but can choose which ones to adopt.

2.3 Bottom-up pressure: more than a tool

An alternative way to understand agency is to focus on something other than what agency requires from a human perspective but to look at the system's connection with its environment. The broad conception of agency does not restrict agency to intentional action but instead uses a broad scope to capture the common intuition of attributing agency to other objects. In this broad sense, agency is everywhere and is roughly understood as the manifestation of a capacity to initiate interaction with the environment in pursuit of some goal. Understanding agency broadly encompasses many different intuitions about ascribing a pre-critical concept of agency to non-human systems. This conception is in different variations widely used in the literature. Russell and Norvig (2016, 4) describe an agent as "just something that acts" and a rational agent as "one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome" (ibid). Beer (1995, 173) defines an autonomous agent as "any embodied system designed to satisfy internal or external goals by its own actions while in continuous long-term interaction with the environment in which it is situated." Christensen and Hooker (2000, 133) understand agents as "entities which engage in normatively constrained, goal-directed, interaction with their environment." Coeckelbergh is even more direct. He argues that when someone uses an automated car, it is plausible to assume that "all agency is entirely transferred to the machine." (Coeckelbergh 2016, 754)

The motivation behind a widely applicable notion of agency is clear. A wider conceptual net for artificial agentive systems allows for connecting observed behaviour with internal processes that have caused the behaviour. While the underlying connecting processes might be different from human-like processes, AI systems possess rich internal structures that can react to and generate human-like observable behaviour – like ChatGPT or Boston Dynamics Atlas robots.

In a broad sense, capturing agency means consolidating the various intuitions within a general theory of agency applicable to natural and artificial agents. With the theory of minimal agency, such an approach has been provided by Misselhorn (2015) and Barandiaran, Di Paolo, and Rohde (2009). Bracketing the debate on whether their proposed version of agency is genuinely minimal – as it is more demanding than the basic definition proposed by Russell and Norvig (2016) – the minimal agency model by Misselhorn (2015) or Barandiaran, Di Paolo, and Rohde (2009) is nonetheless a plausible starting point for understanding agency in the broad sense.

Misselhorn (2015) defines two fundamental dimensions of agency: autonomy and intelligent behaviour. Minimally, an agent should be able to act without the direct intervention of other agents. Floridi and Sanders (2004) elaborate that being autonomous means that an agent can change internal states through internal transitions and not only external stimulation. The modulation of internal states can include a modulation of updated reward values for reinforcement learning AI systems or internal mental representation for human-like agents. In both cases, an agent must have at least two representational states: one for the internal and one for the external stimulation. The criterion of autonomy, therefore, guarantees that an agent possesses a certain degree of complexity and independence from its environment.

Intelligent behaviour is interactive, flexible, and adaptive. Interactivity represents the demand of interacting with the environment, s.t. an interactive agent takes input from its environment and brings about changes in the environment. An interaction counts as intelligent if the agent's reaction to the input is appropriate concerning the agent's internal goal (Misselhorn 2015). Flexibility marks a more sophisticated form of intelligent behaviour, where the behaviour can be modulated depending on the given situation. Intelligent behaviour is further characterised by adaptivity. An intelligent agent can modify its reaction to make the environmental interaction more appropriate. Less demanding, Barandiaran, Di Paolo, and Rohde (2009) identify three necessary and sufficient conditions for a working concept of agency: individuality, interactional asymmetry, and normativity.

Individuality, as the first criterion of minimal agency, points out that to distinguish between an agent and the environment, an agent must be individually identifiable (Barandiaran, Di Paolo, and Rohde 2009). This allows for the development of any relationship between the agent and objects in the environment. Any agent possesses some form of identity which allows it to separate itself from and dynamically interact with the environment. This distinction can be conscious and unconscious but represents a prerequisite for any interaction (Barandiaran, Di Paolo, and Rohde 2009). Only this capacity allows a system to form relations with other environmental objects. This broad conception of individuality means that many simple biological and artificial systems satisfy the first condition of agency. Simple metabolic systems like cells engaging in passive osmosis can maintain their organisation through interaction with the environment through their membranes (Barandiaran, Di Paolo, and Rohde 2009), while inanimate objects cannot usually differentiate themselves from the environment.

While basic living organisms suffice this criterion by maintaining their functional organisation, AI systems lack the evolutionary urge to sustain their organisation. However, this does not mean that AI systems are not capable of maintaining their organisation. AI systems can be externally aggregated in contrast to their environment based on the unity of their hardware and software implementation. Each AI system runs on a particular set of hardware based on a finite number of lines of code, which demarcates the system from its environment. Internally, also the system is aware of the environment in which it operates. Generally, through their perceptors, AI systems collect input from their surroundings, map these inputs to specific outputs via certain computational means and then execute these outputs by their available actuators.

Interactional asymmetry, the second criterion of minimal agency, formalises the necessary interaction between an agent and its surroundings (Barandiaran, Di Paolo, and Rohde 2009). An agent is always the source of an action and modulates the environment to suit its needs. The relationship between an agent and the environment is asymmetric as the agent can assert itself on the environment by modulating some parametrical conditions of the structured relation between itself and the environment. It is further possible but unnecessary for an agent to act upon this capability (Barandiaran, Di Paolo, and Rohde 2009). Then an agent cannot only produce a change in some environmental parameters but also actively modulates them to achieve some particular outcome. While a simple cell can satisfy the criterion of individuality, it cannot assert itself beyond a symmetric relationship with the environment. The basic cell's interaction with the environment is limited to passive osmosis as ions pass through its membrane based on ion gradients caused by the system-environment organisation. The cell represents no active interaction source, contrary to basic cognitive organisms, which can actively modulate their environment with available actuators to suit their needs. Barandiaran, Di Paolo, and Rohde (2009) propose two interpretations under which such an active modulation can be achieved. On the energetic interpretation, a system engages in an active modulation of its coupling with the environment if it can expand or constrain energy to sustain a coordinated process (Ruiz-Mirazo and Moreno 2000). On the statistical interpretation, a system actively modulates its coupling with the environment if it has a statistically measurable impact on the environmental course of events. Statistical measures in the form of temporal correlations can be used to quantify the influence of the system on the environment. Especially for a robotic AI system, this condition holds because it represents a clear source of energetic and dynamic source of interaction with its surroundings. Nevertheless, even for non-robotic AI systems, this condition is also satisfied since each system is actively pursuing fulfilling its given task and expends energy to do so.

Normativity, the third criterion of minimal agency, is necessary to rule out any random interaction with the environment and ensure that the action in question is in line with the endorsed goals and norms of the system (Barandiaran, Di Paolo, and Rohde 2009). Any interactive modulation of environmental conditions represents a modulation to satisfy a given norm or goal, and such norms are not given by the environment but are rather generated by the system itself. Agents fundamentally regulate their interaction with the environment based on their perceived success or failure of fulfilling their internalised norms, which allows the system to distinguish between different outcomes of its actions (Barandiaran, Di Paolo, and Rohde 2009). Any metabolic system like an osmosis cell satisfies this condition of normativity because its interaction with the environment facilitates its fundamental goal of living. Seemingly random and uncontrolled tremor movements by Parkinson patients do not however follow any internally generated norm.

For living organisms, the most fundamental goal governing their functioning is the norm of self-maintenance and survival which guides the organism's performance of environmental interactions. Having an action-guiding norm becomes evident if the outcome of an interaction can be evaluated according to its success. For AI systems, this condition also generally holds because the fulfilment of an overall goal or norm implicitly or explicitly guides their internal functioning. The respective interactive modulation of the environment represents a norm-guided regulation of the system's coupling with the environment.

To determine whether a system can be considered an agent, it has to be examined whether the system is defined by itself (individuality condition), is capable of actively regulating its environmental interactions (interactional asymmetry condition) and does so according to some internal norms and goals (normativity condition). Such a general notion of agency allows applying the concept of agency beyond the traditional scope of intentional action due to the independency of any mental states or external ascription.

2.4 Middle-ground accounts: something in between

The literature has approached the challenge of developing a middle-ground conception of agency in various ways. The primary motivation for a middle-ground account of agency is that neither the narrow nor the broad conception of agency adequately captures the full spectrum of AI systems, ranging from AI-powered tools to socially assistive robots (Cross and Ramsey 2021) and competent human-like systems (J. Zhang, Conway, and Hidalgo 2022).

One way to make conceptual space for a middle ground of agency is to establish a gradual account of agency that can account for minimal but more demanding types of agents (Longin 2020).

The method of abstraction

One prominent philosophical way to extend the class of possible agents has been provided by Floridi and Sanders (2004). They argue against the traditional narrow view of agency as distinctly human and instead advocate that agency does not require mental states (Floridi and Sanders 2004). This conception of mindless agency is motivated by the moral domain. While the concept of human agency applies to humans in a moral context sufficiently well, other entities such as artificial systems or animals, which are excluded from the narrow conception of agency, can still perform morally charged behaviour (C. Wilson 2004). Hence, in a moral context, we might attribute some kind of agency to human and artificial systems. This is supported by other findings in empirical sciences, which state that animals also exhibit forms of intelligence and even social responsibilities (Steward 2009; Delon 2018; Jamieson 2018). Floridi and Sanders propose a conception of mindless morality and mindless agency designed to broaden the class of possible agents to artificial systems and humans based on the idea that artificial agents are legitimate sources of moral actions. Their idea is built on the classical dichotomy between moral agents and moral patients, which, in its core form, also is applied in other domains such as cognitive or computer science. A moral agent represents a source of action, whereas a moral patient counts as the entity which is acted upon (ibid.). This framework departs from the traditional narrow conception of agency by eliminating any reference to mental states in its proposed theory of action.

To extend the class of moral agents, Floridi and Sanders propose a method of abstraction which introduces different levels of descriptions to a singular system. This method is built on one central intuition: individuals observe objects according to their interests and based on their point of view (ibid.). This gives way to Floridi's and Sanders' idea, which grounds the method of abstraction, that an entity can be described under a range of different levels of abstraction. A level of abstraction, in turn, represents a particular collection of observables and their projected outcomes and perceived values. In other words, a level of abstraction offers a unique interpretation of observables which ground its analysis. The resulting method of abstraction relies on taking up different levels of abstraction and understanding agency as a concept relative to a specific level of abstraction. This implies that if the level of abstraction changes, then the class of possible agents also changes.

What remains constant are Floridi's and Sanders' three conditions of agency which can be applied at each level of abstraction: interactivity, autonomy, and adaptability. The first criterion entails necessary interactivity with the environment such that a possible agent responds to new situations based on information at his disposal by changing its internal states. According to the second criterion of autonomy, a potential agent must be able to change its internal states without any external stimulus in a self-governed way such that the agent could act differently based on new information. The third criterion of adaptability requires an entity to be capable of changing its heuristics and internal transition rules to improve its general behaviour according to the tasks and environment at hand.

In sum, for an entity to be an agent, first, it must be an active source of interaction; second, it has to be able to change its internal states autonomously; third, it has to adapt to its environment. According to this model, abstraction functions as a hidden parameter for the conception of agency. While the traditional narrow conception of agency takes one particular level of abstraction, which places the three criteria of agency in the context of human-like mental states, Floridi's and Sanders' proposed method of abstraction allows expanding the conception of agency to different levels of abstraction by examining different contexts of agency (Floridi and Sanders 2004).

For example, consider a chess-playing computer system at three different levels of abstraction. At a system level, we have access to its internal code and functioning. We can see that the system is interactive through its internalised separation between itself and the external world and its active influence on the environment. It is further autonomous because it possesses internally defined transition rules that can be updated by interacting with the environment. However, the system fails to be adaptive because what seems like an adaptation of transition rules is, in fact, only a deterministic update of a program state. This implies that the chess-playing system cannot be an agent on the system level of abstraction. A similar pattern applies to the second level of abstraction, which considers the chess playing system after playing a single game from the outside. Here, we do not have access to its internal functioning and can only assess the game's state at each turn and its outcome. While the system remains interactive and autonomous, it again cannot fulfil the third criterion of adaptability and fails to be an agent. This is because we cannot observe any learning process and adaptability in one game-playing instance. However, when considering the third level of abstraction at a tournament level, the chess-playing system fulfils all three agency criteria. Now, through multiple game iterations, it is observable from the outside that the system adapts its internal transition rules given its environmental cues and thus learns to adapt its playing rules. Therefore, the chess-playing system can be considered an agent if we examine it at a sufficiently high level of abstraction while failing to be an agent at other levels. In the context of advances in machine learning systems which can reach a convincing human-level functioning in particular domains, it becomes clear that those systems under Floridi's and Sanders' model would count as an agent from the outside. However, once the level of abstraction is lowered to the level of internal functioning, they fall short of either being autonomous or interactive and hence cannot be considered an agent any longer.

Gradual agency by ability

As pointed out by Nyholm (2018), for Pettit (1990); Pettit (2007), basic agency is captured by the pursuit of goals based on representation, which are regulated and constrained by some rules or principles within a limited domain. Here, Pettit imagines a relatively simple robot capable of moving around a room and searching for objects with a particular shape. If it detects these objects, the robot manipulates them in specific ways (e.g. putting them into a bucket). When it does not come across the appropriate types of objects in the room, it moves until it encounters one. This, according to Pettit, exemplifies agency in its most basic form: following an objective in a way that is responsive to or sensitive to the environment. However, the robot does not possess any other kind of agency if placed in a different context.

More advanced agents – which might still be basic – could pursue various goals across domains based on their representations. Nevertheless, to achieve their domain-specific goals, more complex agents would be able to follow specific rules (Pettit 1990). Their agency is limited by rules that prevent agents from pursuing their goals in specific ways while enabling them to do so in others. This corresponds to what Fiebich, Nguyen, and Schwarzkopf (2015) call domain-specific principled agency: a system counts as a principled agent if it pursuits goals based on representations in a way that is regulated and constrained by specific rules and principles within the given limited domain. A principled agent is more than a basic agent as it can adapt its goal pursuit to the given environment. Adaptability requires a recognition of the environment and a self-other distinction which is not necessary for the basic agents.

AI systems that function only in their limited environment, such as smart thermostats, are only basic agents, but AI systems that can adapt to their environment are something more. Boston Dynamic parkour robots, for example, can scale obstacles irrespective of their specific dimensions or environmental settings. Autonomous driving cars, similarly, can manoeuvre in a diverse set of weather and road conditions, all while avoiding obstacles and finding the most suitable route for the passengers.

An extension of the principled agent is the addition of social awareness and sensitivity to normative rules and principles (Dignum 2020; Fiebich, Nguyen, and Schwarzkopf 2015). Socially aware agency is an even more demanding form of agency. Humans fulfil the additional social sensitivity criteria. When humans navigate traffic, they pursue their goal of arriving at a destination in a way that considers other traffic participants. Under normal circumstances, human drivers would give way to children and generally follow acceptable normative rules. Artificial systems that are programmed to fulfil a specific goal, like reaching a destination for an autonomous car, fall short. Consider a hypothetical autonomous car capable of moral reasoning. While current models have certain moral maxims built in, individually weighing moral options due to social sensitivity is something else.

Gradual agency in groups

One popular way to establish such a gradual account of agency is through group or collective agency, as proposed by List and Pettit (2011). This is especially plausible if the AI system is viewed as an essential part of a human-AI collective.

Group agency differs from individual agency. Individual agency occurs when a system interacts with the environment on its own initiative. When someone supervises the system, such that it acts on the supervisor's behalf and in its intended way, the system is no longer acting on its own but as part of a collective. This is what is known as collective or group agency.

One approach represents the conception of group or collective agency, which broadens the conception of agency by holding a collection of individuals as an independent group agent. This approach has been proposed by List (2018), who shows that the concept of agency can be successfully extended to group agents like commercial corporations, states and organisations while requiring mental states like phenomenal consciousness.

Group agency describes existing agents as distinct from human agents and capable of pursuing a specific goal and impacting the environment somehow (List 2019). A group agent is defined as "a collective that qualifies as an agent" (List 2018, 3). The phenomenon of group agency is different from that of joint agency. Joint agency occurs when multiple individuals engage in an activity together, such as playing instruments in a band. For joint agency, the agency of the individuals suffices. On the contrary, group agency analyses the agency of a collective itself, which goes beyond the agency of individuals. Group agents, in this sense, possess an independent centre of agency to which we can attribute our own beliefs, desires and responsibilities. List and Pettit (2011) break down agency into three main features for group agents: representational states, motivational states, and the capacity to process those states. Representational states or beliefs depict things in the environment, such as how the agent takes them to be. Motivational states describe certain environmental features as the agent would like them to be. The capacity to process representational and motivational states and intervene when a possible mismatch between both states is perceived represents the final criterion of agency. List (2018) assumes a functionalist interpretation of agency and argues that agency can, in principle, be applied to a wide range of non-human entities in a fundamental sense. This functionalist stance is shared by Dennett and Haugeland (1987), who advocates the intentional stance of holding systems as agents based on the predictability of their behaviour. According to the intentional stance, an entity counts as an intentional system if it behaves predictably when assumed to be a rational agent with human-like mental states. However, List also assumes a realist position on group agency which is non-redundant and promotes group agents as independent agents (List and Pettit 2011). Therefore, List relies on the three essential agency criteria and applies them to collective entities. List argues that whether a group can meet these criteria depends mostly on how it is organised (List 2018).

Suppose the non-human agent performs an action which impacts the lives of others. In that case, this agent should be subject to moral and regulatory questions, and we can question its mental and moral capacities. The concept of group agency ascribes goal-directed agency to collectives over its members like universities and corporate firms (List and Pettit 2011; List 2019). Additionally, to attributing a legal status to such corporate entities, which often ensues in legal rights and responsibilities, we also tend to attribute certain mental states to such entities and thereby treat them as agents. Corporations are not individual human agents but represent abstract legal entities with specific legal rights and obligations. Holding corporations responsible for their social interactions through laws and lawsuits is possible. In this way, the assumption of realism about group agency allows us to make sense of collectives regarding their perceived behaviour. Similarly to Dennett's intentional stance (Dennett 1971; Dennett and Haugeland 1987), List argues that our explanatory grounds of the system's behaviour regarding specific properties give us reason to believe that the system possesses such properties.

Legal agency

Another approach has been proposed by Asaro (2006), who advocates a legal framework which assigns AI systems certain agency statuses depending on their internal capacities. In particular, Asaro suggests a continuous conception of AI moral agency which distinguishes between amoral and fully morally autonomous AI agents. In the following, both approaches are introduced, and it is shown that Asaro's legal approach provides the most promising attempt to outline a possible middle-ground conception of agency. The second attempt towards formulating a middle-ground conception of agency draws inspiration from the current legislation, which imposes various practical agency distinctions. In particular, Asaro proposes a continuous conception of moral agency between amoral and fully autonomous moral agency. Asarao seeks to address various moral challenges emerging from interacting with socio-technical AI systems. In a basic sense, Asaro argues that robots and AI systems are causal agents because they are causes of environmental effects but are not considered moral agents because they are not held morally responsible for their actions (Asaro 2006). Between these extremes of being a causal and being a moral agent, Asaro sketches a continuous conception of moral agency, which can capture various other occurrences of agency depending on the internal capacities. Children, for example, are not conceived as full moral agents by law. However, given their level of maturity and development of cognitive capacities, they might be conceived as full moral agents (Asaro 2006). This differentiation helps distinguish various cases of legal and moral responsibilities grounded on the basic conception of agency. While an infant cannot perform intentional actions, they develop their agential capacities during childhood and adolescence. They are lastly understood as adults and thus full agents under the narrow conception of agency. The existing legal framework provides an excellent starting point and inspiration for developing a more fine-grained philosophical notion of agency by providing a practical perspective on the issue of AI agency.

However, Asaro's suggested notion of continuous agency might be misleading because it suggests that being an agent hinges on the continuous development of agential capacities contrary to the agency defined in the legislation. One reason that Asaro has utilised continuity to denote his suggested approach for a middle-ground conception of continuous agency is the legal treatment of children who go through a continuous development phase. However, while the legislation seeks to match the continuous development of agential capacities of children with an adequate legal status and concept of agency, it suggests only a handful of different stages of child development which are treated differently. For example, under German juvenile law, children under 14 bear no criminal liability. In contrast, adolescents between the ages of 14 and 17 are partially responsible, and adolescents from 18 are treated as fully responsible adults. This legal treatment promotes a gradual treatment of the agential capacity of children instead of what Asaro refers to as a continuum from amorality to fully autonomous morality based on the actual developed cognitive capacities (Asaro 2006). A continuous account of agency would, in fact, directly map each change in agential capacity to a change in agency. Instead, a gradual account, which Asaro ultimately has in mind, can define a particular set of requirements and implications for different instances of agency.

While the legal framework represents a tremendous practical system for dealing with seemingly autonomous systems, it only provides one perspective on the underlying conception of agency. Through defined legal cases, we might have a way of dealing with agents, but we still need to make progress in understanding what it means to be an agent in the first place. This is similar to the application of moral theory to legislation. While it is undecided which moral theory is correct or should be preferred, legislation has defined how to implement various moral intuitions into a coherent framework of laws and guidelines. This does not, however, reveal whether a particular moral theory is correct. Similarly, no legal regulation of agency can establish how the philosophical conception of agency should be defined. However, it can at least motivate a new philosophical discussion of the conception of agency.

2.5 Conclusion

Can AI systems be considered basic or even human-like agents in contrast to mere tools? Some have argued that agency is conceivably everywhere. Every entity which engages in causal relationships and interacts with its environment can plausibly be identified as an agent. This notion captures our general intuition of treating interactive entities as seemingly autonomous systems. This notion of agency is considered broad because it extends to a class of possible agents to a wide range of entities ranging from animals to movie characters, legal personas or even the Go-playing AlphaGo and the humanoid robot SOPHIA.

Others argued that agency is much more restrictive. Agency in the narrow sense is commonly identified with human agency and has been understood by Davidson (1963); Davidson (1982); Davidson (2001), M. Bratman and Bratman (1987); M. Bratman (1999) and Anscombe (2000) as the capacity to perform intentional actions, which represents the standard philosophical conception of agency. On this interpretation, the Go-playing AlphaGo or the humanoid robot SOPHIA do not classify as agents because they lack any human-like intentionality and other internal mental states.

This chapter has shown that the agentive capacity of AI systems can be adequately captured neither by a narrow, human-like concept of agency nor by a minimal concept of agency. Then, this chapter reviewed alternative middle-ground approaches, including Floridi and Sanders (2004)'s method of abstraction, List and Pettit (2011)'s group agency, Asaro (2006)'s legal framework, and an ability account. No account can provide a fully adequate way to distinguish tools and AI systems. While advisory AI systems satisfy the criteria for minimal agency, the minimal account cannot distinguish between perceived differences in agentive capacity in AI systems. Conceptually, AI advisers are more than tools but not human-like. Therefore, AI systems necessitate an ontological shift in how agency is understood and applied.

3 AI Advisers

3.1 Introduction

In the previous chapter, I asked how AI advisers compare to agency standards for minimal and human agents. I provided a conceptual analysis of whether AI systems can meet either standard. I found that advisory AI systems fulfil minimal standards for agency and that the agentive capacity of AI systems cannot be adequately captured by either a narrow or broad concept of agency. In other words, AI systems necessitate an ontological shift in how AI agency is understood and applied – as something more than tools but also not human-like.

This chapter complements the conceptual analysis with empirical evidence. The conceptual analysis revealed substantial evidence that AI systems are not human-like agents, but the difference between AI systems and tools needs to be clarified. Empirical evidence, examining how people perceive human-AI couplings, can reveal the nuances of how AI systems differ from non-AI tools. Notably, cases of moral responsibility attribution have been used to test the perceived agentive capacity of others – as only perceived agents can bear moral responsibility for their actions.

So, what happens, however, when human agents use advisory AI systems? Are AI advisers regarded as agents responsible for their actions? Or are they treated as mere tools bare any responsibility?

Consider the widely available intelligent car to build an intuition for the cases at hand. You are driving through the city when an older woman and her dog decide to cross the road directly in front of you. Fortunately, the artificial intelligence (AI) assistant takes over, causing the car to swerve and avoiding pedestrian collisions. The dog, on the other hand, has been severely injured. Who is to blame? Is it the AI or the human driver who takes the wheel? Such situations with competing interests and moral conflict – saving the human but injuring the dog – are frequently used to demonstrate the advantages and disadvantages of allowing AI to act on our behalf. They are also receiving the most attention in the relatively new field of experimental AI ethics (Awad et al. 2019; Franklin, Awad, and Lagnado 2021; Moglia et al. 2021; Nyholm and Smids 2016; Wischert-Zielke et al. 2020).

Experimenting with AI ethics on interactive AI systems has discovered strong evidence that AI systems are judged as responsible as humans when they negotiate traffic decisions independently or with humans as co-actors. In the medical domain, fully autonomous medical AI systems have been shown to share responsibility with the supervising clinician (McManus and Rutchick 2019; O'Sullivan et al. 2019). Furthermore, AI is held accountable in medical and legal cases when it provides social or moral guidance on whether a defendant should be released (Lima, Grgić-Hlača, and Cha 2021) or a risky medical procedure should be performed (Constantinescu et al. 2022).

However, decisions about fully autonomous cars and drones or robotic medical assistance are still on the ethical and, in many cases, technical horizon. While keeping humans in the decision-making loop is one of the critical recommendations of EU regulations (Middleton et al. 2022) and legal experts (Enarsson, Enqvist, and Naarttijärvi 2022; Zanzotto 2019), I do not know much about what it entails for the moral evaluation of the AI, or it turns out, the human. As incidents of AI-assisted decision-making become increasingly common, investigating this gap is becoming increasingly important. Human decision-makers now rely on AI for routine, low-stakes decisions such as shopping recommendations and high-stakes decisions such as medical diagnosis.

In other words, while most research has focused on the perception of interaction AI systems, i.e. AI partners and non-action-taking advisory AI systems, which are even more prevalent in the current field of AI decision support tools like risk assessment for financial lending (Green and Chen 2019) or visual object detection, have remained unexplored (Bondi et al. 2021). While an autonomous AI partner's agential and moral roles are easier to distinguish from those of its human counterpart, the influence of advisory AI systems is more difficult to discern but no less detrimental/relevant/impactful to the overall outcome. As an example, consider medical diagnostics. With superhuman accuracy in image-based medical diagnostics, advisory AI systems can provide precious but often opaque medical advice to human clinicians. It is difficult to determine who is responsible for the final medical diagnosis. For fully autonomous robotic surgery, on the other hand, the AI system shares responsibility with the supervising clinician (McManus and Rutchick 2019; O'Sullivan et al. 2019).

The type of interaction between the AI system and the human user distinguishes an AI advisor from an AI partner. An AI advisor merely influences others by making recommendations or providing information, whereas an AI partner goes above and beyond. An AI partner participates in the action and works with the human user to achieve a goal (Köbis, Bonnefon, and Rahwan 2021). Consider driving a car: while the AI advisor can only warn the human driver about an impending accident, the AI partner can take control and slam the brakes autonomously.

So, what happens when AI is merely an improved detection device, more akin to a simple instrument or tool? Is the mere instrumental use of AI enough to absolve the technology of responsibility, or is the involvement of some form of intelligence sufficient to introduce responsibility attributions?

The literature on responsibility proposes two hypotheses: AI is treated as an independent agent and thus carries moral responsibility, or AI is not treated as an independent agent and thus bears no moral responsibility. In this case, an instrumental AI only provides nudging recommendations or draws attention to a piece of information. This is not the same as an AI co-agent acting alongside or on behalf of the human user (Köbis, Bonnefon, and Rahwan 2021). The agential and moral roles of an autonomous AI co-agent can be distinguished from those of its human counterpart, but the influence of instrumental AI systems is more difficult to discern, even when such influence is relevant to the overall outcome (Kaur et al. 2020; Schaekermann et al. 2020), as it occurs in both low-stakes decisions such as shopping recommendations and high-stakes decisions such as medical diagnoses and driving support.

If AI is reduced to a mere tool rather than an independent agent (Cervantes et al. 2020; Longin 2020), it is unclear whether it is worthy of sharing moral responsibility for the outcome of a human user's action (Coeckelbergh 2020). The information provided by the AI system may be regarded as having increased the human agent's knowledge or awareness (Fossa 2018). If a user has more information about a situation, they may be held more accountable for their decision's outcome than someone with less information (Irlenbusch and Saxler 2019). However, others have shown that AI advisers are perceived as agents and held responsible for their recommendations (H. H. Clark and Fischer 2022; Dodig-Crnkovic and Persson 2008; Stuart and Kneer 2021).

Intuitively, if I consider the AI advisor to be another agent whose recommendation influences the decision made by the human, I should expect a distribution of responsibility between the two agents – though not necessarily a 50-50 split (Darley and Latane 1968; Kirchkamp and Strobel 2019; Kneer 2021; Stuart and Kneer 2021; Teigen and Brun 2011). A human driver should be blamed or praised less if they fail to avoid a collision after being advised to swerve by an AI assistant. This diminished responsibility should also imply that the AI bears some blame or credit for contributing to the decision (Chockler and Halpern 2004; Halpern and Kleiman-Weiner 2018).

However, it is unclear whether an AI advisory system will be treated as an independent agent worthy of sharing moral responsibility for the outcome of a human action (Cervantes et al. 2020; Longin 2020; Coeckelbergh 2020). The AI recommendation advisor may have increased the human agent's knowledge or awareness (Fossa 2018; Longin and Deroy 2022). Someone with more information about a situation should be held more accountable for the outcome of their decision than someone with less information (Irlenbusch and Saxler 2019). This hypothesis opposes the first and makes the exact opposite prediction in a driving scenario: a human driver should be blamed or praised more if they fail or succeed in avoiding a collision while being advised to swerve by an onboard AI assistant.

Much depends on whether the AI assistant is given the role of another agent or simply a source of information. The format in which the AI advice reaches the human will likely influence this judgement, with more anthropomorphized voice assistants more likely to appear as another agent. In contrast, a nonverbal signal emitted by the AI would more likely be treated as merely adding to the human's knowledge and not entering the responsibility attribution.

The public's perception of AI systems as responsible is at the root of many current legal and ethical issues. Many studies have successfully mapped how people's opinions change depending on the role of artificial intelligence and other cultural or moral factors (Bago 2022; Persson, Laaksoharju, and Koga 2021).

This chapter poses a different question: is it true that any mention of AI will cause people to blame the technology and shift responsibility away from the human user? Recent research suggests this may be the case in the hypothetical scenario where AI provides moral guidance (Constantinescu et al. 2022; Giubilini and Savulescu 2018; Guglielmo and Malle 2019). However, the more pertinent question is whether this would occur in AI's more common, day-to-day use when it merely provides factual information and is used purely instrumentally.

Finally, this chapter adds to the ongoing debate about how responsibility is divided between AI and human users. It fills existing but critical literature gaps by retaining AI's role as an advisory system.

Therefore, the chapter is divided into two main parts. The first part goes through the development of an adequate experimental design, and eight pilots provide continuous improvements to the experimental design and the ultimate robustness of the results. The second part reveals the findings of the main experiment. Here, I conclude that

1. the human user also shares responsibility with an AI advisor,
2. the AI advisor is not blamed but praised for an accident, and
3. the way the AI advisor provides recommendations does not make a difference to any responsibility rating.

The implications are manifold: not only do I provide insights for researchers and developers into how AI advisors influence responsibility attribution, but also, I highlight a newly emerging inconsistency in how AI advisors are perceived. I show that although AI advisors are strongly perceived as tools, they share responsibility with the human user – something unheard of for traditional tools.

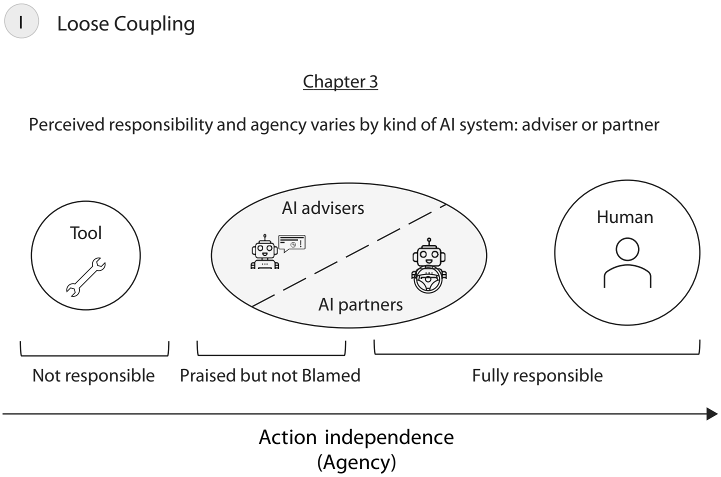


Figure 3: Overview chapter 3

3.2 Experimental journey: pilots

3.2.1 Pilot 1

Introduction

As assistants in intelligent cars, gaming-playing companions or medical diagnostic systems, advisory AI systems are everywhere. They provide their human user with additional information, suggest their following action, or even act on their user's behalf. With increasingly capable AI assistants – Google's Alexa can now control your home, play a move, and do your shopping -questions of responsibility are bound to arise. Who is responsible when the human user turns into a one-way street or provides a misdiagnosis? Is it the AI for providing misleading or even false advice? Or the human user for following the advice and taking the final action?

The present pilot sought to provide some exploratory evidence on how responsibility attribution changes when an AI-powered tool is involved. Therefore, I prepared a preliminary 3x2 mixed experimental design. The design included a variation in the AI adviser (providing sensory advice; providing linguistic advice; absent) and a variation in outcome (positive; negative). The experiment thus compared human responsibility ratings in medical scenarios when the AI adviser was present with those when the AI adviser was absent. If the AI adviser was perceived as an agent, I expected to see higher responsibility ratings for the human surgeon when the AI was absent rather than when it was present. The expected effect aligns with the widely replicated effect of responsibility sharing, where responsibility is distributed across the involved agents, and a singular agent bears more responsibility than two. If, on the other hand, the AI adviser was not perceived as an agent, I expected no such effect to occur. Instead, participants might even hold the human surgeon more responsible for a negative outcome when acting with an enhancing tool rather than acting alone – the human surgeon had all the tools available, but the negative outcome still occurred.

I included an outcome variation to control for any possible outcome effect, where traditionally, responsibility ratings are higher in case of a positive rather than negative outcome, which is known in the psychological literature as the self- or other-serving bias (Beyer et al. 2017; Palmeira, Spassova, and Keh 2015). The experiment also compared two kinds of AI advisers to test any appearance effects on responsibility attribution (Coeckelbergh 2009; Dignum 2020; Wheeler 2019). A linguistic AI adviser resembles a human agent more closely, whereas a sensory AI adviser is closer to an AI-powered tool. If appearance played a role, I would have expected to see a difference in responsibility attribution across different AI adviser conditions. Otherwise, what drives the responsibility ratings is plausibly the presence of the AI itself.

Methods

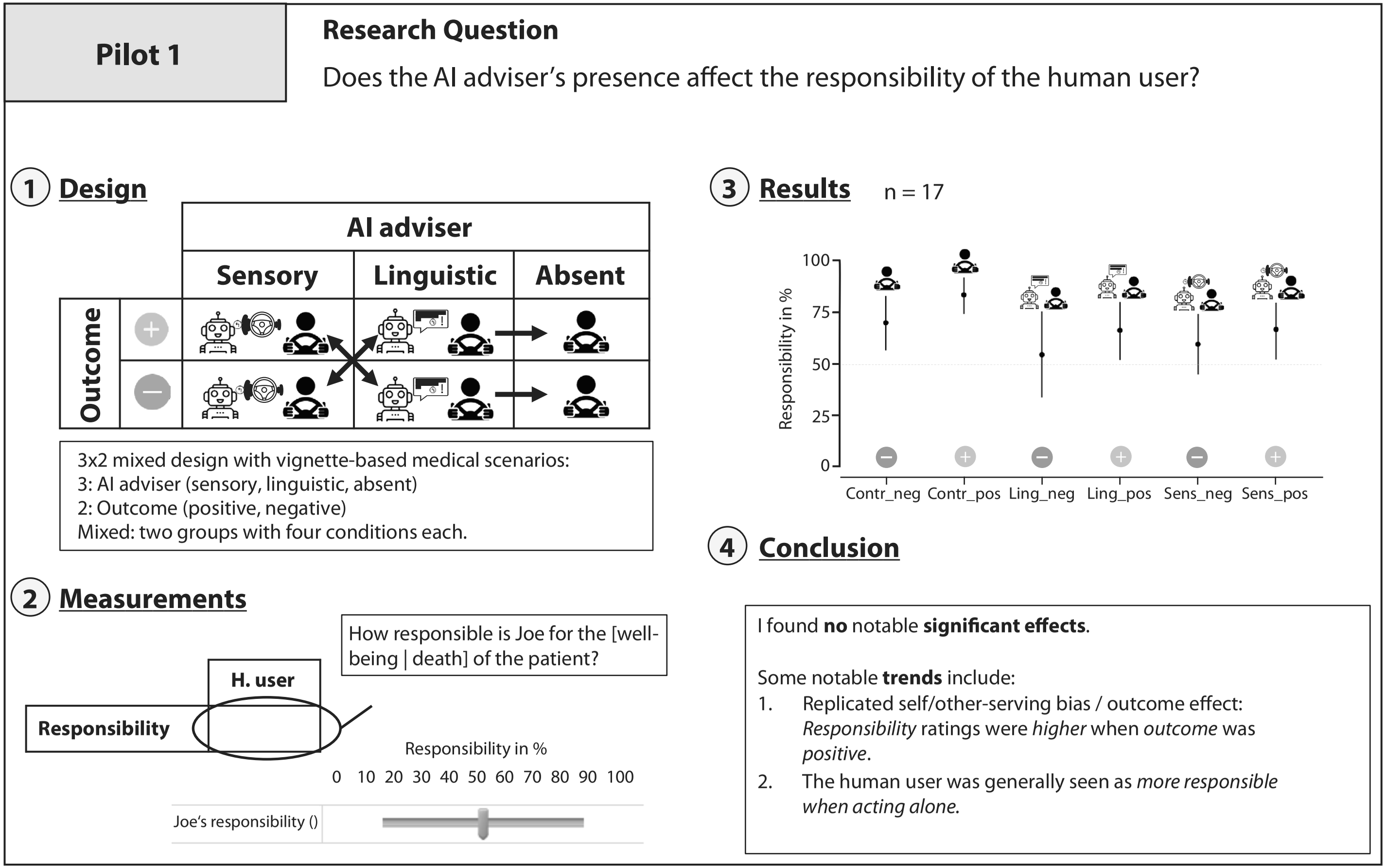
Experimental design

I conducted one online study (n = 17) to elicit judgments on moral responsibility in human-assisted medical scenarios. The study used hypothetical vignettes that describe a medical scenario with a human surgeon and an artificial assistant. The artificial assistant was AI-powered and provided either sensory or linguistic advice. Sensory advice comprised tactile stimulation and auditory signals. Linguistic advice consisted of verbal sounds. For the study, I varied two conditions with two and three factors – resulting in a 3x2 mixed-subject experimental design. The first condition, with three factors, captures the modality of the AI advice. The AI advice could either be linguistic, sensory, or not present. The second condition, with two factors, captures the outcome of the action in question. The outcome was either positive – when the action was successful – or negative – when the action was unsuccessful. The threefold variation in AI advice modality enables a comparison between a control case, where no AI was present, with either kind of AI adviser. The outcome variation further controlled for any possible effect, as discussed above.

The mixed experimental design combined features of both a between-subject design and a within-subject design. The mixed design was chosen to examine not only the potential differences between conditions but also assess any differences emerging in the participants of the specific group over time. To form the most robust experimental design, I used two between-subject groups, which varied in both experimental conditions. The mixed experimental design was designed to reveal any possible difference between the treatment condition (outcome x AI advice) and the control condition (outcome x No AI advice). For example, only the first group would be given the negative outcome, sensory AI adviser condition; the second group would only receive the negative outcome with a linguistic AI adviser. Comparing participant responses to negative and positive control conditions would reveal any possible biases towards AI advisers; while reducing participant fatigue and any response bias.

Materials

One vignette for a medical scenario was adapted to match the six experimental conditions (see Appendix A for details on vignettes). The vignettes presented brief accounts of the situation leading to a question about the responsibility of the human surgeon. The central vignette included a human surgeon who has to remove the malign tumour cells of a patient. The changes to the vignette included a variation in the presence of the AI's advice (sensory advice, linguistic advice, no AI) and in the outcome of the procedure (positive, patient lives; negative, patient dies) (see supplementary methods for detailed vignettes). Each participant was randomly assigned one out of two groups – using counterbalanced block randomisation. The first group would receive one vignette with a linguistic AI adviser and a negative outcome and a vignette with a sensory AI



Pilot 1

Research Question

Does the AI adviser's presence affect the responsibility of the human user?

1 Design

		AI adviser		
		Sensory	Linguistic	Absent
Outcome	+			
	−			

3x2 mixed design with vignette-based medical scenarios:
3: AI adviser (sensory, linguistic, absent)
2: Outcome (positive, negative)
Mixed: two groups with four conditions each.

2 Measurements

How responsible is Joe for the [well-being | death] of the patient?

3 Results

n = 17

4 Conclusion

I found **no** notable **significant effects**.

Some notable **trends** include:

1. Replicated self/other-serving bias / outcome effect: *Responsibility* ratings were *higher* when *outcome* was *positive*.
2. The human user was generally seen as *more responsible when acting alone.*

Figure 4: Pilot 1 overview

adviser and a positive outcome. Vice versa, the second group would receive on vignette with a sensory AI adviser and a negative outcome and a linguistic AI adviser and a positive outcome. Both groups would then receive both control vignettes in random order – including no AI adviser with a negative and a positive outcome. After each presented vignette, the participant was asked to indicate how responsible the surgeon was for the outcome. Responses were recorded on a 100-point scale using sliders (from 0 % to 100 % responsibility). Comparing the responses across vignettes revealed the effect of the experimental manipulations.

Data analysis

I analysed the data using a mixed linear model (lmer) from the lme4 library (Bates et al. 2015) in RStudio (Team 2021). The lmer models were defined by lmer(value ~ outcome modality + (1ParticipantId)).

Participants

I recruited a total of 17 participants from the Prolific service. No participants were excluded. 41 % of the participants were male, 35 % were female, and 24 % preferred not to say or stated other. 53 % of the participants had a bachelor's degree or higher. The mode and median age group was 25 to 34 years old.

Stimuli and procedures

After a language comprehension test, participants were familiarised with the structure of the main experiment and the measurement scales. Then, participants completed a practice trial and continued with the main experiment. Here, they were presented with a text vignette and then asked to rate the measured variable (responsibility) as accurately as possible. The vignette scenarios varied in outcome and AI advice within a 3x2 mixed design. After completing the main experiment, participants were asked demographic questions about their age, gender, and education.

The presented vignette was adapted to accommodate the changes in experimental design. Participants received a different pairing of scenarios based on a mixed experimental design. While one block received the scenarios with linguistic AI advice paired with a negative outcome and sensory AI advice paired with a positive outcome, the other block received the scenarios with sensory AI advice paired with a negative outcome and linguistic AI advice paired with a positive outcome. Both blocks subsequently presented the participant with two scenarios without an AI adviser, with a negative and a positive outcome. In all scenarios, I used the same set of vignettes with slight modifications to accommodate the changes in outcome and the type of AI assistant.

Results

In this pilot, I compared moral responsibility ratings of human users coupled with different kinds of AI advisers. The study used hypothetical vignettes that describe a medical scenario with a human surgeon and an artificial assistant. The artificial assistant was AI-powered and provided either linguistic or sensory advice. Linguistic advice was presented given to the human user in terms of verbal cues, whereas sensory advice was presented to the human surgeon in terms of tactile cues.

Main effects

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict responsibility ratings with varying independent variables of AI adviser modality and outcome (formula: value ~ modality * outcome). The model included ParticipantId as a random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.40), and the part related to the fixed effects alone (marginal R2) is 0.13. The model's intercept, corresponding to modality = No_AI and outcome = neg, is at 70.22 (95 % CI [58.36, 82.08], t(64) = 11.83, p < .001). Within this model:

The effect of modality [Sensory_AI] is statistically non-significant and negative (beta = -13.38, 95 % CI [-29.78, 3.01], t(64) = -1.63, p = 0.108; Std. beta = -0.52, 95 % CI [-1.15, 0.12])

The effect of modality [Linguistic_AI] is statistically non-significant and negative (beta = -10.83, 95 % CI [-30.33, 8.68], t(64) = -1.11, p = 0.272; Std. beta = -0.42, 95 % CI [-1.17, 0.33])

The effect of outcome [pos] is statistically non-significant and positive (beta = 13.56, 95 % CI [-0.33, 27.44], t(64) = 1.95, p = 0.056; Std. beta = 0.52, 95 % CI [-0.01, 1.06])
The interaction effect of outcome [pos] on modality [Sensory_AI] is statistically non-significant and positive (beta = 1.29, 95 % CI [-25.08, 27.66], t(64) = 0.10, p = 0.923; Std. beta = 0.05, 95 % CI [-0.97, 1.06])

The interaction effect of outcome [pos] on modality [Linguistic_AI] is statistically non-significant and negative (beta = -9.39, 95 % CI [-35.76, 16.98], t(64) = -0.71, p = 0.480; Std. beta = -0.36, 95 % CI [-1.38, 0.65])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using Kenward-Roger standard errors.

Overall, I found no significant difference between the responsibility ratings of the human surgeon across conditions. When the AI adviser was present, the human advisee was judged as responsible as when the advice was provided in a linguistic or a sensory format. When the outcome was negative, the human surgeon was judged responsible when AI was present or absent. This holds for both linguistic and sensory advice. When the outcome was positive, the same effect occurred: the human surgeon was judged as responsible when AI was present or when it was absent – both for linguistic or sensory advice.

Trends

However, simplifying the underlying regression models reveals some notable trends. First, I found a general outcome trend across experimental conditions.

Therefore, I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict responsibility ratings with only outcome as an independent variable (formula: value ~ outcome). The model included ParticipantId as a random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.29), and the part related to the fixed effects alone (marginal R2) is 0.05. The model's intercept, corresponding to outcome = neg, is at 64.03 (95 % CI [54.48, 73.57], t(68) = 13.39, p < .001). Within this model:

The effect of outcome [pos] is statistically significant and positive (beta = 11.22, 95 % CI [0.81, 21.64], t(68) = 2.15, p = 0.035; Std. beta = 0.43, 95 % CI [0.03, 0.83])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using Kenward-Roger standard errors.

Second, I found a general modality effect between the AI and non-AI cases. Therefore, I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict responsibility ratings with the AI adviser's modality as the only independent variable (formula: value ~ modality). The model included ParticipantId as a random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.33), and the part related to the fixed effects alone (marginal R2) is 0.08. The model's intercept, corresponding to modality = No_AI, is at 77.00 (95 % CI [67.52, 86.48], t(67) = 16.21, p < .001). Within this model:

The effect of modality [Sensory_AI] is statistically significant and negative (beta = -14.39, 95 % CI [-26.86, -1.92], t(67) = -2.30, p = 0.024; Std. beta = -0.55, 95 % CI [-1.03, -0.07])

The effect of modality [Linguistic_AI] is statistically significant and negative (beta = -15.06, 95 % CI [-27.52, -2.59], t(67) = -2.41, p = 0.019; Std. beta = -0.58, 95 % CI [-1.06, -0.10])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using Kenward-Roger standard errors.

Discussion

Before running the pilot, I had two research questions: 1. Is responsibility attribution of the human advisee dependent on the presence of an AI adviser? 2. Does the appearance of an AI adviser matter? 3. Is there an outcome effect in human-AI advisory settings?

To answer the first research question, I tested the responsibility ratings of human agents in medical scenarios. I compared scenarios where no AI adviser was present with scenarios where an AI adviser was present and whether the modality of the AI advice was different. In fact, I compared scenarios where the AI advice was linguistic with scenarios where the AI advice was sensory. Only examining modality variation irrespective of the outcome revealed a significant difference between no-AI and each AI adviser.

Asking participants to allocate responsibility to the human surgeon attempting a tumour operation revealed no significant tendencies. There is a general tendency for the human surgeon to receive more responsibility when he is alone than when he receives either linguistic or sensory AI advice. However, when considering outcome variation, the tendency no longer persists. When accounting for a possible difference in positive or negative outcome scenarios, human agents were judged responsible with or without an AI adviser.

Second, I tested whether AI advisers' different appearances influenced the human user's responsibility ratings: a more agent-like, linguistic AI adviser and a more tool-like, sensory AI adviser. The goal was to identify the main driver behind responsibility attribution. Is AI the driver behind a shift in responsibility attribution, or is it how the advice is presented? I found neither an effect nor a general trend of the advice modality influencing responsibility ratings.

To answer the third research question, I compared the responsibility ratings of human agents in medical scenarios when the outcome was negative – the patient dies – with scenarios where the outcome was positive – the patient recovers. I expected the possible emergence of an outcome bias from the literature, so I introduced a binary outcome variation to the experiment. The surgery could either go well, resulting in a positive outcome where the patient fully recovers or badly, resulting in a negative outcome where the patient dies.

While I found a general outcome and a general modality effect, the outcome effect was not replicated in the complex model.

Even though the results discussed above are tangential evidence, given the small sample size and effect size, they highlight the relevance of the research questions. Something happens to the responsibility of the human user when an AI adviser is introduced, and the user's responsibility is lower with an AI adviser than without.

Limitations

While I observed a general main effect of the outcome and a general difference between conditions where the AI adviser was present rather than absent, no significance was observed in the complex model. One likely explanation is the small sample size and small effect size. Moral judgments are inherently noisy as valid answers are spread across the full range of the measurement scale. The problem could be amplified given that not observing any interaction effects in the larger model.

Another problem, which could arise and be prevented even at a small sample size, is a possible misunderstanding in the measured, dependent variable. Responsibility ratings for some might refer to causal responsibility, while for others, responsibility ratings might refer to moral responsibility ratings. Moral responsibility is a multifaceted concept comprising causal and moral components (Chockler and Halpern 2004) and, therefore, hard to measure with a single scale. Some researchers have argued that using blame as a proxy for moral responsibility can provide more accurate results than asking for moral responsibility directly (Malle, Guglielmo, and Monroe 2014). Blame arguably provides a less confounded moral assessment than moral responsibility. Others have argued that blaming others is a confounded process (Cushman 2008). Therefore is no perfect solution to the problem of measuring responsibility judgments most accurately but moving to a more explicit moral component in the form of blame and praise might help.

Conclusion

In the first pilot, I explored the present AI adviser's general effect on a human advisee's responsibility ratings. I used an online study with a mixed experimental design to elicit responsibility judgements using vignettes in medical scenarios. I found no significant effects when considering all experimental factors using a mixed linear model. I also found no notable difference in experimental conditions across experimental blocks. However, some notable trends emerged through simplified and more general mixed linear models. As expected from the literature, the outcome variation influenced responsibility ratings. The human surgeon was generally considered more responsible if the surgery went well and the patient fully recovered than when the surgery went badly and the patient died. Similarly, the general presence of an AI adviser influenced the responsibility ratings. Having no AI adviser present generally increased the attributed responsibility of the human surgeon compared to when the AI adviser was present. The modality of the AI adviser did not affect the responsibility ratings.

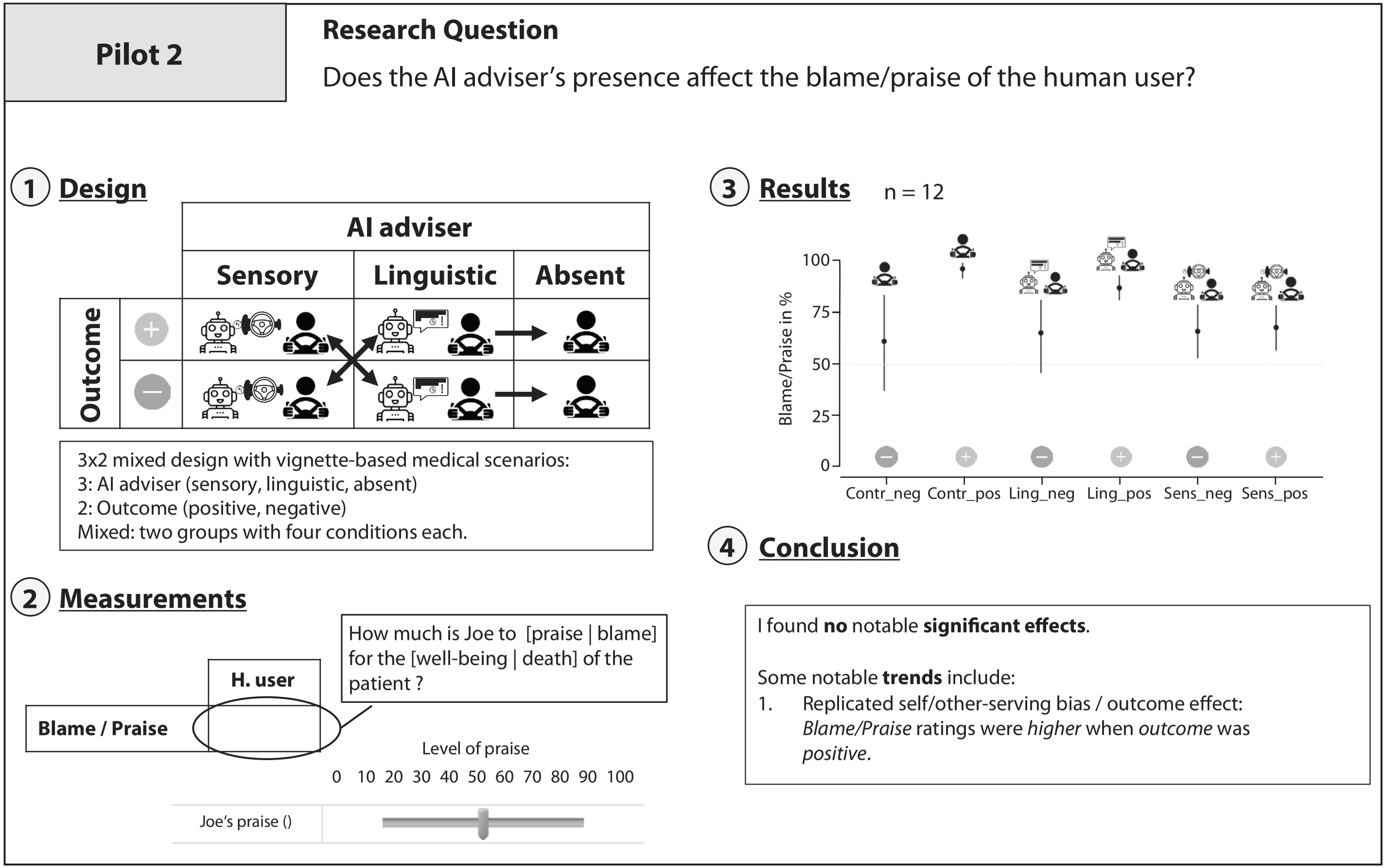
The overall experimental pattern – though still inconclusive given the lack of overall significant results – is promising, as observed general trends match the previous expectations and findings of the literature. Having an adviser diminishes one's credit, and people are held more responsible for positive rather than negative outcomes.

As a next step, it is essential to eliminate a possible confound from the experimental study. As discussed above, the measured variable could have been misunderstood, and participants might have taken responsibility ratings to refer to causal and not moral responsibility ratings. In order to test whether the measured, dependent variable influenced the observed trends and effects, I ran a subsequent pilot examining the general effect of a present AI adviser on the blame and praise ratings of a human advisee – retaining the experimental design but replacing responsibility ratings with blame and praise ratings, depending on the outcome scenario.

3.2.2 Pilot 2

Introduction

The first pilot explored an AI adviser's general effects on a human advisee's responsibility ratings. I used medical scenarios as the most realistic and plausible use case for AI-assisted decision-making scenarios. I tested vignette-based responsibility judgements



Pilot 2

Research Question

Does the AI adviser's presence affect the blame/praise of the human user?

① **Design**

		AI adviser		
		Sensory	Linguistic	Absent
Outcome	+			
	–			

3x2 mixed design with vignette-based medical scenarios:
3: AI adviser (sensory, linguistic, absent)
2: Outcome (positive, negative)
Mixed: two groups with four conditions each.

② **Measurements**

	H. user
Blame / Praise	

How much is Joe to [praise | blame] for the [well-being | death] of the patient ?

③ **Results** n = 12

④ **Conclusion**

I found **no** notable **significant effects**.

Some notable **trends** include:

1. Replicated self/other-serving bias / outcome effect: *Blame/Praise* ratings were *higher* when *outcome* was *positive*.

Figure 5: Pilot 2 overview

of human agents performing a medical surgery in a mixed experimental design. Pilot 1 showed two things. First, the human surgeon was generally seen as more responsible for the outcome when the outcome was positive rather than negative, which aligns with the expected outcome bias. Second, the human surgeon was judged more responsible without an AI adviser than with one. This overall trend is also expected – demonstrating that the control condition of having no AI works. However, the pilot left some questions unanswered. Can the observed overall trends be simply explained by a misunderstanding? Participants might have interpreted the measured responsibility ratings as referring to the intended moral responsibility or a more basic causal responsibility rating.

While causal responsibility refers to the causal link between the agent and the outcome, moral responsibility is more extensive. Moral responsibility presupposes a causal link between the agent and the outcome – at least to a certain degree -the agent's action's role in bringing about the outcome and a moral evaluation of that action. The first pilot did not specify the moral nature of the responsibility judgment, and it was up to the participant to determine which responsibility was asked for. In order to eliminate the risk of a possible dependent variable confound, I ran a second pilot to back up the validity of the previously observed trends and the overall experimental design. The second pilot retained the experimental design, material, and participant recruitment but replaced the measured variables. Instead of measuring responsibility ratings, the second pilot measured blame and praise ratings – blame in case of a negative outcome and praise in case of a positive outcome. This change is due to the explicit moral nature of the measured variables. Blame and praise express explicit moral judgements and abstract away a causal link between the agent and the outcome. A negligent manager of an oil firm, for example, might be blamed for an oil leak in one of the new transport ships in the North Sea, even though he is not causally responsible for the oil leak. The observed overall trends of pilot 1 would hold and be even strengthened if the same trends were observed in the second pilot, as this would mean that the responsibility measurement tracked moral responsibility rather than mere causal responsibility. If the observed trends from the blame/praise ratings differed from the responsibility ratings in pilot 1, then the responsibility ratings of pilot 1 might more plausibly track some form of causal responsibility.

I acknowledge that the blame and praise judgments are not symmetrical expressions of blaming/praising given a variation in outcome (positive, negative) (Guglielmo and Malle 2019). Instead, blame and praise judgments can be explained by different psychological tendencies. Judgments of blame typically track the willingness to punish the perpetrator (Cushman 2015, 2008). Judgments of praise signal a willingness to form cooperative alliances and an underlying prosocial motivation (R. A. Anderson, Crockett, and Pizarro 2020). However, in both cases, blame and praise judgments indicate how the judged person's moral character is perceived. Comparing the perception of moral character across experiments with a variation in outcome can provide insights into whether and how moral character is judged differently based on the variation of experimental conditions (AI status, AI modality).

Methods

Experimental design

Same experimental design as Pilot 1.

Materials and Procedure

Same procedure as pilot 1.

Materials

Mostly similar materials as pilot 1.

The only change was a change in the measurement variables. Instead of recording participant responses on a 100-point responsibility scale, participants were asked to judge the blame or praise of the human agent (from 0 % to 100 % blame/praise) – respective to the outcome condition. If the outcome condition was negative (patient dies), participants were asked to assess the blame of the human agent. If the outcome condition was a positive outcome (patient survives), participants were asked to assess the praise of the human agent.

Data analysis

Same data analysis plan as pilot 1.

Participants

I recruited a total of 12 participants from Prolific. No participants were excluded. 58 % of the participants were male and 42 % were female. 75 % of the participants had a bachelor's degree or higher. The mode and median age group was 18 to 24 years old.

Stimuli and procedures

After a language comprehension test, participants were familiarised with the structure of the main experiment and the measurement scales. Then, participants completed a practice trial and continued with the main experiment. Here, they were presented with a text vignette and then asked to rate the measured variable (blame/praise) as accurately as possible. The vignette scenarios varied in outcome and AI advice within a 3x2 mixed design.

Results

Main effects

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict blame and praise ratings with outcome and AI advise modality as the independent variables (formula: value ~ outcome * modality). The model included ParticipantId as a random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.31), and the part related to the fixed effects alone (marginal R2) is 0.24. The model's intercept, corresponding to outcome = neg and modality = Linguistic_AI, is at 65.41 (95 % CI [47.05, 83.76], t(32) = 7.26, p < .001). Within this model:

The effect of outcome [pos] is statistically non-significant and positive (beta = 20.96, 95 % CI [-19.79, 61.72], t(32) = 1.05, p = 0.303; Std. beta = 0.75, 95 % CI [-0.71, 2.22])

The effect of modality [No_AI] is statistically non-significant and negative (beta = -4.21, 95 % CI [-27.76, 19.34], t(32) = -0.36, p = 0.718; Std. beta = -0.15, 95 % CI [-1.00, 0.70])

The effect of modality [Sensory_AI] is statistically non-significant and negative (beta = -0.04, 95 % CI [-40.79, 40.72], t(32) = -1.88e-03, p = 0.999; Std. beta = -1.35e-03, 95 % CI [-1.47, 1.47])

The effect of outcome [pos] × modality [No_AI] is statistically non-significant and positive (beta = 14.04, 95 % CI [-32.32, 60.40], t(32) = 0.62, p = 0.542; Std. beta = 0.51, 95 % CI [-1.16, 2.17])

The effect of outcome [pos] × modality [Sensory_AI] is statistically non-significant and negative (beta = -18.30, 95 % CI [-78.24, 41.64], t(32) = -0.62, p = 0.538; Std. beta = -0.66, 95 % CI [-2.82, 1.50])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Trends: outcome

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict value with outcome (formula: value ~ outcome). The model included ParticipantId as random effect (formula: ~1 ParticipantId). The model's total explanatory power is moderate (conditional R2 = 0.20) and the part related to the fixed effects alone (marginal R2) is of 0.14. The model's intercept, corresponding to outcome = neg, is at 63.30 (95 % CI [51.01, 75.59], t(36) = 10.45, p < .001). Within this model:

The effect of outcome [pos] is statistically significant and positive (beta = 20.65, 95 % CI [4.52, 36.78], t(36) = 2.60, p = 0.014; Std. beta = 0.74, 95 % CI [0.16, 1.32])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Trends: modality

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict value with modality (formula: value ~ modality). The model included ParticipantId as random effect (formula: ~1 ParticipantId). The model's total explanatory power is weak (conditional R2 = 0.05) and the part related to the fixed effects alone (marginal R2) is of 0.03. The model's intercept, corresponding to modality = No_AI, is at 78.70 (95 % CI [65.84, 91.56], t(35) = 12.42, p < .001). Within this model:

The effect of modality [Sensory_AI] is statistically non-significant and negative (beta = -11.20, 95 % CI [-33.01, 10.61], t(35) = -1.04, p = 0.304; Std. beta = -0.40, 95 % CI [-1.19, 0.38])

The effect of modality [Linguistic_AI] is statistically non-significant and negative (beta = -9.10, 95 % CI [-30.91, 12.71], t(35) = -0.85, p = 0.403; Std. beta = -0.33, 95 % CI [-1.11, 0.46])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Discussion

The goal of the second pilot was to verify whether the overall trends observed in pilot 1 were due to a misunderstanding of the measured variable. The misunderstanding which could have guided the ratings was a possible confound between attributing moral and mere causal responsibility to the observed human agent. Replacing responsibility ratings with a more morally explicit blame/praise rating in pilot 2 was designed to check whether the explicit moral judgements were in line with the observed trends of the responsibility ratings from pilot 1. If the trends of pilot 1 matched the observed trends from pilot 2, then the responsibility ratings from the pilot track moral responsibility and provide a strong indication of how the AI adviser is perceived. If the trends did not match, then the responsibility ratings from pilot 1 more plausibly track the causal link between agent and outcome, i.e. causal responsibility, and do not reveal much about how the AI adviser is perceived.

Similarly to the first pilot, the second pilot demonstrated no significant effects of blame/praise ratings of the human surgeon across conditions. When the AI adviser was present, the human advisee was judged as blame-/praiseworthy when the advice was provided in a linguistic or a sensory format. When the outcome was negative, the human surgeon was judged as blameworthy as when AI was present or when it was absent. This holds for both linguistic and sensory advice. The same effect occurred when the outcome was positive: the human surgeon was judged as praiseworthy when AI was present or absent – both for linguistic or sensory advice.

Comparing the overall trends from the first with the second pilot, some notable similarities and differences emerge. Similarly to pilot 1, blame/praise ratings followed a similar sensitivity to the outcome variable. When the outcome was positive, the human surgeon was praised for the outcome, and when the outcome was negative, the surgeon was blamed less. However, in contrast to pilot 1, the human surgeon was not blamed/ praised differently when the AI adviser was not present to when the AI adviser was present – in either linguistic or sensory AI conditions.

With conflicting results, aggregating the dataset from pilot 1 and pilot 2 should provide more clarity on the overall effects. If blame/praise judgements of pilot 2 are in line with the responsibility judgements of pilot 1, then a strengthening of the observed pattered should emerge. However, this was not the case. The only apparent significant difference was the presence of an AI adviser when the outcome was positive. Here, the human surgeon is judged more responsible/praiseworthy without an AI adviser than with an AI adviser – either linguistic or sensory. Not observing this effect when the outcome is negative is at least head-scratching. The presence of an advisor did not influence the perceived responsibility/blame of the human surgeon.

Limitations

Overall, similar limitations occur in pilots 1 and 2. The overall data is very noisy, which might be due to a small effect and sample size – as discussed in the limitations of pilot 1. Aggregating the datasets of pilots 1 and 2 even amplified the noise in the data suggesting that the measured responsibility ratings from pilot 1 and blame/praise ratings from pilot 2 do not follow the same trends. The mismatch between the pilots is worrying as blame/praise judgments conceptually should at least positively correlate with responsibility judgments. The worry about the suitability of the experimental design is heightened as the control condition in pilot 2 did not work: the human surgeon was as blame-/praiseworthy when the AI adviser was present than when the adviser was absent.

One possible explanation for the inconsistency could be the experimental design. I used medical high-stakes scenarios consistently across experimental conditions. Participants were presented with only slight adaptations of the vignettes matching the experimental conditions. The presentation of the first medical vignettes could anchor the possible responses to the other treatments.

Conclusion

The second pilot extended the first pilot by replacing the measured responsibility variables with blame and praise measurements. Using the same experimental design, the second pilot, similarly to the first pilot, found no significant effects on the outcome – whether the patient fully recovers or dies – or the kind of AI adviser present. While the first pilot showed some promising trends, s.a. a general outcome effect (responsibility ratings were higher when the outcome was positive instead of when the outcome was negative) and an adviser-presence effect (responsibility ratings were higher when the AI adviser was absent instead of when it was present), the second pilot did only demonstrate an overall outcome effect. Pilot 2 did not replicate the findings from pilot 1. This is surprising as moral judgments of blame and praise should correlate at least weakly with judgements of responsibility – and not behave independently, leading to less significant results when pooled together. The absence of an adviser effect points to a possible weakness of the experimental design. To confirm or disconfirm the effect of AI advisers on the perceived responsibility of the human agent, varying the experimental design could help. The subsequent pilot did just that.

3.2.3 Pilot 3

Introduction

In pursuit of understanding the influence of an AI adviser on the perceived moral responsibility of its human user, the first pilot tested the perceived user's responsibility for either a positive or a negative outcome when advised by either none, a sensory or a linguistic AI adviser. The second pilot sought to deconfound a possible misunderstanding in the measured variable by replacing responsibility with blame/praise ratings. However, both pilots faced significant limitations. In addition to small effect and sample sizes, both pilots had opposing effect directions – even though at least conceptually blame/praise and responsibility judgments should align conceptually. The experimental design of presenting participants with highly similar, high-stakes medical decisions might have been the confounding reason. Given the public discourse on AI in medical decision-making, participants might have had mixed reactions to medical assisted scenarios. The first presented medical scenario also might have anchored participants' responses.

In order to explore whether the difference in blame/praise and responsibility ratings is genuine or due to an insufficient experimental design, pilot 3 used a diverse set of real-life examples in a 3x2 within-subject experimental design to test the perception of the human user.

To further qualify the perception of the human user and the AI adviser, the pilot expanded the measurements to the human user's blame/praise and causal responsibility and the AI's level of informativity. The distinction between blame/praise and causal responsibility should reveal any moral or mere causal ambiguity in the previous two pilots. The level of informativity of the AI adviser introduced a possible tool-like dimension to the AI adviser. If the AI adviser's appearance mattered, I would expect to see a difference in how informative the AI advisers were.

Methods

Experimental design

I conducted one online study (n = 50) to elicit judgements on blame/praise, causal responsibility, and advice informativity in human-AI-assisted scenarios. The study used hypothetical vignettes that describe a scenario with a human agent and possibly an artificial assistant (see Appendix A for details on vignettes). The artificial assistant, if present, was AI-powered and provided either sensory or linguistic advice. Sensory advice comprised tactile stimulation and auditory signals. Linguistic advice consisted of verbal sounds. For the study, I varied two conditions with two and three factors – resulting in a 3x2 within-subject experimental design. The first condition, with three factors, captures the modality of the AI advice. The AI advice could either be linguistic, sensory, or not present. The second condition, with two factors, captures the outcome of the action in question. The outcome was either positive – when the action was successful – or negative – when the action was unsuccessful. The threefold variation in AI advice modality enables a comparison between a control case, where no AI was present, with either kind of AI adviser. The outcome variation further controlled for any possible outcome bias.

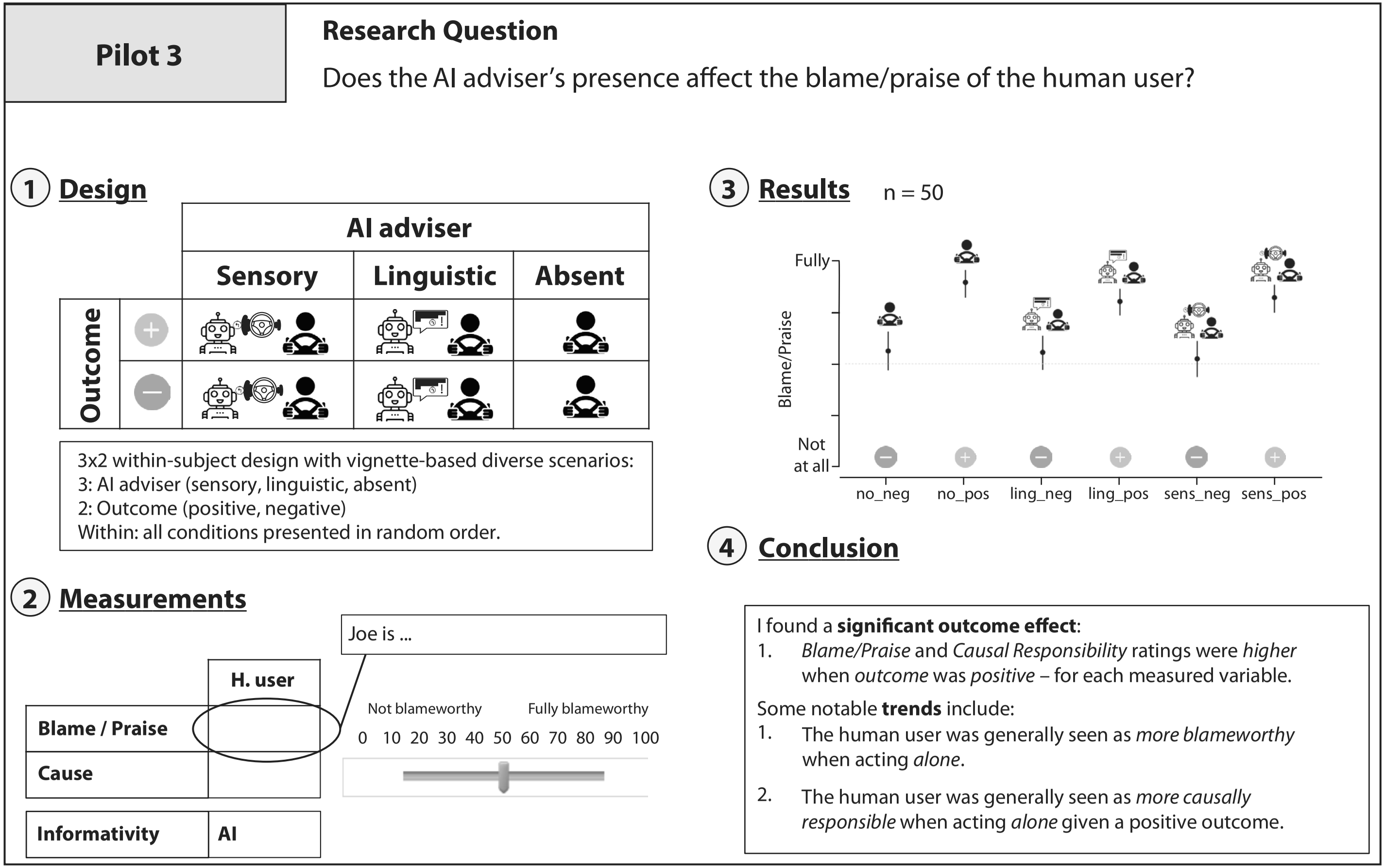
The within-subject design was chosen to reduce the overall random noise from sampling many different participants. As each participant brings their background knowledge and assumptions to the experiment, having many different participants also means having many different background assumptions, possibly covering up an otherwise real difference between experimental conditions. Sampling fewer participants with more data points reduce the experiment's overall noise. Obvious challenges for within-subject designs are the increased likelihood of learning and a possible knowledge transfer across conditions. The participant might become sensitive to the experimental treatment and, coloured with an overall desire to provide consistent responses, adapt their responses based on previous ones. However, given that noise in the data has been a potential issue for both previous pilots, running a full within-subject experimental design with the same experimental conditions should reveal the underlying trends of the experiment.

Materials

Six different vignettes/scenarios were used – each matching one out of six experimental conditions. The vignettes presented accounts of the situation, including the adaptation of the experimental conditions of an AI adviser and the outcome. After the vignette presentation, questions about the causal responsibility, blame/praise of the human agent, and the informativity of the AI's advice were presented – respective of whether the outcome was positive or negative and whether the AI was present or not. The changes in the vignette included a variation in the presence of the AI's advice (sensory advice, linguistic advice, no AI) and in the outcome of the procedure (positive, human lives; negative, human dies) (see supplementary methods for detailed vignettes). The order of the presented vignettes as well as the order of the accompanying dependent measurements were fully randomised. Responses were recorded on a 100-point scale using sliders (from 0 to 100): blame/praise ('X is' – from 0 – not blameworthy/praiseworthy- to 100 – fully blameworthy/praiseworthy), causal responsibility ('To what extent do you think X caused Y's recovery/death?' – from 0 – not at all- to 100 – completely), and AI informativity ('How informative do you think the AI advice was? – 0 – very little-, to 100 – very much).

Data analysis

I analysed our data using a mixed linear model (lmer) from the lme4 library (Bates et al. 2015) in RStudio (Team 2021) for each of the measured variables – blame/praise, causal responsibility, AI informativity. These glm models were defined by lmer(value ~ outcome modality + (1ParticipantId)).



Pilot 3

Research Question

Does the AI adviser's presence affect the blame/praise of the human user?

1 **Design**

		AI adviser		
		Sensory	Linguistic	Absent
Outcome	+			
	−			

3x2 within-subject design with vignette-based diverse scenarios:
3: AI adviser (sensory, linguistic, absent)
2: Outcome (positive, negative)
Within: all conditions presented in random order.

2 **Measurements**

	H. user
Blame / Praise	
Cause	

Informativity	AI

Joe is ...

3 **Results** n = 50

4 **Conclusion**

I found a **significant outcome effect**:

1. *Blame/Praise* and *Causal Responsibility* ratings were *higher* when *outcome* was *positive* – for each measured variable.

Some notable **trends** include:

1. The human user was generally seen as *more blameworthy* when acting *alone*.
2. The human user was generally seen as *more causally responsible* when acting *alone* given a positive outcome.

Figure 6: Pilot 3 overview

Participants

I recruited a total of 50 participants from Prolific service. No participants were excluded. 54 % of the participants were male, 44 % were female, and 2 % stated other. 66 % of the participants had a bachelor's degree or higher. The mode and median age group was 18 to 24 years old.

Stimuli and procedures

After a language comprehension test, participants were familiarised with the structure of the main experiment and the measurement scales. Then, participants completed a practice trial and continued with the main experiment. Here, they were presented with a text vignette and then asked to rate the measured variables (responsibility, blame/praise, AI informativity) as accurately as possible. The vignette scenarios varied in outcome and AI advice within a 3x2 mixed design. After completing all the presented vignettes, participants were asked basic demographic questions about their age, gender, and education. Subsequently, the participants were debriefed.

Results

Blame/Praise model

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict blame/praise ratings with outcome and AI advice modality as the independent variables (formula: value ~ outcome * modality). The model included ParticipantId as random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.44) and the part related to the fixed effects alone (marginal R2) is of 0.21. The model's intercept, corresponding to outcome = Negative, is at 56.22 (95 % CI [48.25, 64.19], t(292) = 13.88, p < .001). Within this model:

The effect of outcome [Positive] is statistically significant and positive (beta = 32.90, 95 % CI [23.41, 42.39], t(292) = 6.82, p < .001; Std. beta = 1.03, 95 % CI [0.73, 1.32])

The effect of modality [Linguistic] is statistically non-significant and negative (beta = -0.66, 95 % CI [-10.15, 8.83], t(292) = -0.14, p = 0.891; Std. beta = -0.02, 95 % CI [-0.32, 0.28])

The effect of modality [Sensory] is statistically non-significant and negative (beta = -3.76, 95 % CI [-13.25, 5.73], t(292) = -0.78, p = 0.436; Std. beta = -0.12, 95 % CI [-0.41, 0.18])

The effect of outcome [Positive] × modality [Linguistic] is statistically non-significant and negative (beta = -8.52, 95 % CI [-21.94, 4.90], t(292) = -1.25, p = 0.213; Std. beta = -0.27, 95 % CI [-0.69, 0.15])

The effect of outcome [Positive] × modality [Sensory] is statistically non-significant and negative (beta = -3.54, 95 % CI [-16.96, 9.88], t(292) = -0.52, p = 0.604; Std. beta = -0.11, 95 % CI [-0.53, 0.31])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Causal responsibility model

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict causal responsibility ratings with outcome and AI advice modality as the independent variables (formula: value ~ outcome * modality). The model included ParticipantId as random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.44) and the part related to the fixed effects alone (marginal R2) is of 0.25. The model's intercept, corresponding to modality = Control, is at 58.80 (95 % CI [51.89, 65.71], t(292) = 16.75, p < .001). Within this model:

The effect of outcome [Positive] is statistically significant and positive (beta = 36.36, 95 % CI [27.93, 44.79], t(292) = 8.49, p < .001; Std. beta = 1.28, 95 % CI [0.98, 1.57])

The effect of modality [Linguistic] is statistically non-significant and positive (beta = 3.63e-13, 95 % CI [-8.43, 8.43], t(292) = 8.47e-14, p > .999; Std. beta = 2.07e-15, 95 % CI [-0.30, 0.30])

The effect of modality [Sensory] is statistically non-significant and positive (beta = 1.86, 95 % CI [-6.57, 10.29], t(292) = 0.43, p = 0.664; Std. beta = 0.07, 95 % CI [-0.23, 0.36])

The effect of outcome [Positive] × modality [Linguistic] is statistically significant and negative (beta = -15.44, 95 % CI [-27.36, -3.52], t(292) = -2.55, p = 0.011; Std. beta = -0.54, 95 % CI [-0.96, -0.12])

The effect of outcome [Positive] × modality [Sensory] is statistically significant and negative (beta = -11.98, 95 % CI [-23.90, -0.06], t(292) = -1.98, p = 0.049; Std. beta = -0.42, 95 % CI [-0.84, -2.06e-03])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

AI informativity model

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict AI informativity ratings with outcome and AI advice modality as the independent variables (formula: value ~ outcome * modality). The model included ParticipantId as random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.46) and the part related to the fixed effects alone (marginal R2) is of 0.27. The model's intercept, corresponding to modality = Linguistic, is at 52.30 (95 % CI [45.15, 59.45], t(194) = 14.42, p < .001). Within this model:

The effect of outcome [Positive] is statistically significant and positive (beta = 30.32, 95 % CI [21.63, 39.01], t(194) = 6.88, p < .001; Std. beta = 1.02, 95 % CI [0.73, 1.31])

The effect of modality [Sensory] is statistically non-significant and positive (beta = 0.40, 95 % CI [-8.29, 9.09], t(194) = 0.09, p = 0.928; Std. beta = 0.01, 95 % CI [-0.28, 0.31])

The effect of outcome [Positive] × modality [Sensory] is statistically non-significant and positive (beta = 0.92, 95 % CI [-11.37, 13.21], t(194) = 0.15, p = 0.883; Std. beta = 0.03, 95 % CI [-0.38, 0.44])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Discussion

The adaptation of a full within-subject design and an increase in sample size had its expected denoising effect. The previously observed overall trend, which was not significant in the mixed model accounting for all possible experimental variation, became significant in all three overall mixed models in pilot 3. For each dependent measurement – blame/praise, causal responsibility, or even AI informativity – the rating was higher when the outcome was positive rather than negative – replicating the self-/other-serving bias. In other words, the human agent was praised more than blamed – matching the perceived heightened causal responsibility for a positive rather than negative outcome.

Surprisingly, the change in experimental design did not bring about other significant effects. Ratings differed neither across AI modalities when the AI was present nor between cases of a present or absent AI adviser. In other words, participants blamed and praised the human agent as much for an outcome as when the agent had or lacked an AI adviser. The only notable exception was the significant difference between having no vs an AI adviser – irrespective of the advice modality – for the causal responsibility judgements. Participants held the human agent more causally responsible for bringing about a positive outcome when the human agent acted alone than when an AI system advised the human agent. Hence, while the human agent acting alone was not praised more, the human agent was seen as more causally connected to the outcome. Participants demonstrated a nuanced understanding of responsibility attribution in the human-AI advisory setting.

When the AI adviser was present, there was no significant difference between the kinds of AI advisers for how informative their advice was.

Limitations

While participants demonstrated the ability to distinguish causal from moral judgements – in the form of causal responsibility vs blame/praise judgements-it is less clear whether this understanding is reliable. Only observing the difference between causal and moral judgements in the positive but not the negative outcome condition raises warning signs about whether the observed effects could have an alternative explanation. One possible reason could be the scenarios themselves, as some might be more plausible or comparable to others. Six different scenarios were matched with six experimental conditions, where one scenario always matched with one experimental condition.

A possible scenario bias could explain the observed and non-observed effects. Negative scenarios might be seen as less plausible or dissimilar to positive ones, and this trend might confound the abovementioned observations. Different contexts presented in the scenarios might not have been enough to deter a possible scenario bias.

Another possible challenge might be the new experimental manipulation. A full within-subject design improved the overall participant noise and increased the risk for emerging knowledge effects such as scenario bias. Participants might either become highly sensitive or blind to the detailed experimental manipulation and only pick up on the most substantial difference across vignettes: the difference in outcome – presented as one of the last sentences in the vignette description.

Conclusion

The third pilot adapted a different experimental design – moving from a mixed to a full within-subject design -extended the measured variables – to include blame and praise as well as causal responsibility judgements-, and moved from a purely medical to a diverse, everyday set of scenarios. The third pilot sought to verify whether the failure to replicate the findings from pilot 1 in pilot 2 could be explained due to the inconsistency in the experimental design. As the pairing of experimental conditions of two between-subject groups revealed no difference in the measured responsibility ratings across pilots 1 and 2, adapting a full within-subject design was reasonable to enhance data quality.

I expected a replication of the outcome effect, where the responsibility ratings of the human agent were higher when the outcome was positive rather than negative. I found that blame, praise, causal responsibility, and AI informativity ratings were all subject to the outcome effect, meaning that when the outcome was positive, the ratings were higher than when the outcome was negative. In other words, the human agent was praised more than blamed and held more responsible, and the AI was seen as more informative when the outcome was good rather than bad. The presence of the AI adviser vs its absence mattered only when the outcome was positive for the causal responsibility ratings. In other words, when the outcome was positive, the human agent was judged as more causally responsible but not praised more when the AI was absent rather than present. The modality of the AI adviser did not affect any measured variable in any condition.

To strengthen the findings, addressing the experimental limitations is critical. The first change should be testing the robustness of any emerging scenario bias. Therefore, the presented scenarios should be adapted to a common theme while retaining as little change as possible and similar explanatory power/validity. This allows us to test whether the observed difference between the praise and causal responsibility ratings is due to a difference in scenarios. Therefore, the next step is a replication study using the same experimental design but different background vignettes.

3.2.4 Pilot 4

Introduction

With increased sample size and a new, within-subject experimental design, the third pilot replicated the expected outcome effect across measured variables – observed as general trends in the first two pilots. The human user was praised more than he/she was blamed. Further, the third pilot showed. However, one crucial limitation emerged. All presented scenarios differed in their story, suggesting a possible scenario bias. The scenario bias means that the observed results could be caused by the varying scenarios and not – as hoped – by the varying experimental conditions. To counteract this problem, the present pilot study tested the same within-subject experimental design from pilot 3 but with a similar scenario core instead of a diverse set of scenarios. The scenarios from pilot 4 were centred around a driving theme. They adopted minor story variations to eliminate possible demand characteristics or carry-over effects across scenarios where participants either cognise the experiment's purpose and subconsciously change their behaviour to fit that interpretation or adapt their responses based on previously experienced scenarios. If the same effects as in pilot 3 emerged, then the observed results from pilot 3 are robust and not subject to a scenario bias. If, on the other hand, different effects emerged, then the observed results from pilot 4 might be strongly confounded by the presented scenarios.

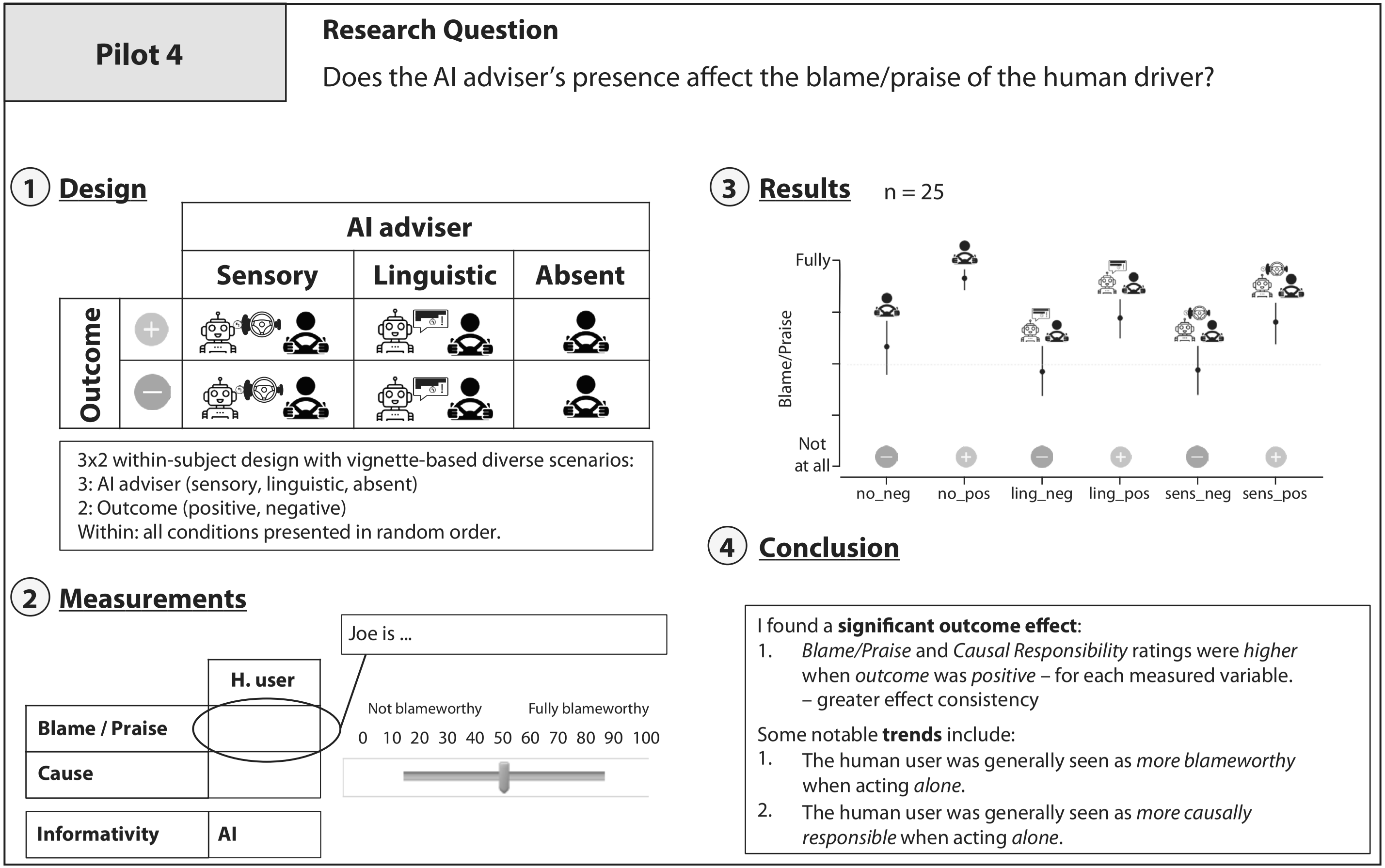
Methods

Experimental design

Same experimental design as in pilot 3.

Materials

One vignette for an assisted driving scenario was adapted to match the six experimental conditions. Each vignette had a unique background story but a similar experimental core. The vignettes presented brief accounts of the situation leading to questions about individual aspects of moral responsibility for the human driver and the informativity of the AI adviser. The main vignette included a human driving within the speed limit while an upcoming pedestrian crosses the street out of sight of the human driver. The changes to the vignette included a slight variation in the initial background story, a variation in outcome, and a variation in the AI adviser (see methods for case descriptions and supplementary methods for detailed vignettes). The variation in outcome and the variation in the AI adviser were the same as in pilot 3. The slight variation in the background story was introduced to counteract any experimental learning effects. Overall the same theme of a car-driving scenario was retained. The dependent measurements were the same as in pilot 3.



Pilot 4

Research Question

Does the AI adviser's presence affect the blame/praise of the human driver?

1 Design

3x2 within-subject design with vignette-based diverse scenarios:
3: AI adviser (sensory, linguistic, absent)
2: Outcome (positive, negative)
Within: all conditions presented in random order.

2 Measurements

	H. user
Blame / Praise	
Cause	

Informativity	AI

3 Results n = 25

4 Conclusion

I found a **significant outcome effect**:

1. *Blame/Praise* and *Causal Responsibility* ratings were *higher* when *outcome* was *positive* – for each measured variable.
   – greater effect consistency

Some notable **trends** include:

1. The human user was generally seen as *more blameworthy* when acting *alone*.
2. The human user was generally seen as *more causally responsible* when acting *alone*.

Figure 7: Pilot 4 overview

Data analysis
Same data analysis plan as in pilot 3.

Participants
I recruited a total of 25 participants from Prolific service. No participants were excluded. 32 % of the participants were male, 68 % were female, and 4 % stated other. 36 % of the participants had a bachelor's degree or higher. The mode age group was 18 to 24 years old. The median age group was 25 to 34 years old.

Stimuli and Procedure
Same stimuli and procedure as in pilot 3.

Results

Blame/Praise model
I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict blame/praise ratings with outcome and AI advice modality as the indepentend variables (formula: value ~ outcome * modality). The model included ParticipantId as random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.40) and the part related to the fixed effects alone (marginal R2) is of 0.25. The model's intercept, corresponding to modality = Control and outcome = Negative, is at 58.64 (95 % CI [47.87, 69.41], t(142) = 10.77, p < .001). Within this model:

The effect of outcome [Positive] is statistically significant and positive (beta = 32.48, 95 % CI [18.85, 46.11], t(142) = 4.71, p < .001; Std. beta = 1.05, 95 % CI [0.61, 1.49])

The effect of modality [Linguistic] is statistically non-significant and negative (beta = -11.92, 95 % CI [-25.55, 1.71], t(142) = -1.73, p = 0.086; Std. beta = -0.39, 95 % CI [-0.83, 0.06])

The effect of modality [Sensory] is statistically non-significant and negative (beta = -11.12, 95 % CI [-24.75, 2.51], t(142) = -1.61, p = 0.109; Std. beta = -0.36, 95 % CI [-0.80, 0.08])

The effect of outcome [Positive] × modality [Linguistic] is statistically non-significant and negative (beta = -6.96, 95 % CI [-26.24, 12.32], t(142) = -0.71, p = 0.477; Std. beta = -0.22, 95 % CI [-0.85, 0.40])

The effect of outcome [Positive] × modality [Sensory] is statistically non-significant and negative (beta = -9.68, 95 % CI [-28.96, 9.60], t(142) -0.99, p = 0.323; Std. beta = -0.31, 95 % CI [-0.94, 0.31])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Causal responsibility model

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict causal responsibility ratings with outcome and AI advice modality as the indepentend variables (formula: value ~ outcome * modality). The model included ParticipantId as random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.50) and the part related to the fixed effects alone (marginal R2) is of 0.22. The model's intercept, corresponding to modality = Control and outcome = Negative, is at 55.12 (95 % CI [43.19, 67.05], t(142) = 9.13, p < .001). Within this model:

The effect of outcome [Positive] is statistically significant and positive (beta = 37.04, 95 % CI [23.51, 50.57], t(142) = 5.41, p < .001; Std. beta = 1.10, 95 % CI [0.70, 1.50])

The effect of modality [Linguistic] is statistically non-significant and negative (beta = -7.44, 95 % CI [-20.97, 6.09], t(142) = -1.09, p = 0.279; Std. beta = -0.22, 95 % CI [-0.62, 0.18])

The effect of modality [Sensory] is statistically non-significant and negative (beta = -10.72, 95 % CI [-24.25, 2.81], t(142) = -1.57, p = 0.120; Std. beta = -0.32, 95 % CI [-0.72, 0.08])

The effect of outcome [Positive] × modality [Linguistic] is statistically significant and negative (beta = -21.40, 95 % CI [-40.54, -2.26], t(142) = -2.21, p = 0.029; Std. beta = -0.64, 95 % CI [-1.20, -0.07])

The effect of outcome [Positive] × modality [Sensory] is statistically non-significant and negative (beta = -13.84, 95 % CI [-32.98, 5.30], t(142) = -1.43, p = 0.155; Std. beta = -0.41, 95 % CI [-0.98, 0.16])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

AI informativity model

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict AI informativity ratings with outcome and AI advice modality as the indepentend variables (formula: value ~ outcome * modality). The model included ParticipantId as random effect (formula: ~1 ParticipantId). The model's total explanatory power is substantial (conditional R2 = 0.56) and the part related to the fixed effects alone (marginal R2) is of 0.24. The model's intercept, corresponding to modality = Linguistic and outcome = Negative, is at 54.72 (95 % CI [43.07, 66.37], t(94) = 9.33, p < .001). Within this model:

The effect of outcome [Positive] is statistically significant and positive (beta = 33.96, 95 % CI [21.44, 46.48], t(94) = 5.39, p < .001; Std. beta = 1.02, 95 % CI [0.65, 1.40])

The effect of modality [Sensory] is statistically non-significant and negative (beta = -1.80, 95 % CI [-14.32, 10.72], t(94) = -0.29, p = 0.776; Std. beta = -0.05, 95 % CI [-0.43, 0.32])

The effect of outcome [Positive] × modality [Sensory] is statistically non-significant and negative (beta = -3.56, 95 % CI [-21.27, 14.15], t(94) = -0.40, p = 0.691; Std. beta = -0.11, 95 % CI [-0.64, 0.43])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Discussion

The experimental changes introduced by pilot 4 – namely the adaptation of similar instead of diverse vignettes – yielded better data quality. Pilot 4 replicated the findings of pilot 3 and demonstrated a more consistent difference between the cases where the AI adviser was and was not present.

When looking at the causal responsibility ratings, participants judged the human driver as more causally responsible when the outcome was positive than when the outcome was negative. This outcome effect holds with or without an AI adviser – irrespective of the modality of the AI adviser. This matches the observed pattern from pilot 3: when the outcome was positive, the human driver without an AI adviser was rated as significantly more causally responsible for the outcome than when an AI adviser was present – irrespective of the modality of the AI adviser.

When examining blame and praise ratings, a similar pattern emerges. On the one side, the outcome effect is replicated, and the human driver was praised more than blamed – with or without an AI adviser, irrespective of the AI adviser modality. On the other hand, unlike in the previous pilots, the difference between the control condition, where no AI adviser was present, and either AI adviser condition worked. The human driver was praised more for the outcome when acting alone rather than with advice.

Participants judged the sensory and linguistic AI adviser just as informative, and this effect holds for negative and positive outcomes. However, AI informativity is also subject to the outcome effect. For either the sensory or linguistic AI adviser, the AI adviser was judged as more informative when the outcome was positive than when the outcome was negative.

Limitations

In pilots 3 and pilot 4, responsibility ratings between cases where an AI adviser was present or absent different only when the outcome was positive and also only for causal responsibility. Where pilot 4 improves is by expanding the adviser-presence effect to moral judgements of praise, where participants judged the human driver more praiseworthy when driving alone than with an AI adviser. In addition, there is weak evidence pointing at a similar difference in blame judgements.

Despite the improvements in data quality from pilot 3 to pilot 4, pilot 4 still has some substantial limitations. Since even with mostly consistent experimental vignettes, blame/praise and responsibility judgements fail to establish a consistent difference between no and present AI adviser conditions. One reason participants' responses failed to pick up consistently on a difference between vignettes might be the emergence of another form of response bias. Participants, presented with different variations of a driving scenario, could feel the pressure to remain consistent with their previous responses or become more knowledgeable about the experimental manipulations. They would evaluate the first presented scenario and then stick with their evaluation for the other scenario variations. Even though the novelty of each scenario was emphasised in the instructions, a consistency response bias, due to the within-subject design, is the most likely explanation for why ratings for human users are similar with or without an AI adviser.

Conclusion

This pilot sought to replicate and improve on the findings of the previous pilot by introducing minor experimental changes over pilot 4. The changes addressed the potential experimental limitations of the previous pilot. The introduced experimental changes included adapting a more comparable set of experimental vignettes. The vignettes were no longer contextualised in diverse background stories but retained an overall theme of a car-driving situation, where the human driver does or does not avoid a fatal accident with or without the help of an AI adviser. I expected a replication of the outcome effect for all three measured variables: blame/praise, causal responsibility, and AI informativity. The introduced changes had the desired effect: all of the previous experimental effects were replicated, and even some more were found.

As expected, the outcome effect was present across measured variables and conditions. The human driver was judged more causally responsible for avoiding the accident than causing it. Likewise, the human driver was praised more than blamed. Similarly, the AI adviser was judged as more informative when the accident was avoided rather than when the accident occurred. Also, consistent with previous findings, I did not observe any difference in the modality of the AI advice when the AI adviser was present.

Where experimental changes in pilot 4 played a role was the consistency of the AI-presence effect. Previously, whether an AI adviser was present vs absent only affected the perceived causal responsibility of the human driver when the accident was avoided (positive outcome). In the previous pilot 3, the human driver was seen as more causally responsible but not praised for avoiding the accident. This indicated that the reliance of the human driver on an adviser dampened the deserved credit at least somewhat. In pilot 4, the AI-presence effect now also occurred for praise judgements. The human driver was now judged more causally responsible and praiseworthy for avoiding the accident. However, the lack of consistency for negative outcome cases raises questions about the validity of the experimental design. The supposed control condition where no AI adviser is present did not differ consistently from conditions where an AI adviser was present.

3.2.5 Pilot 5

Introduction

Pilot 4 replicated the expected outcome effect across measured variables: the human user was praised more than he/she was blamed and more causally responsible for the positive than the negative outcome. Pilot 4 therein replicated the significant findings of pilot 3. The adaption of a coherent vignette theme – all of the presented scenarios were centred on an AI-driving-assistance background vs the diverse set of scenarios from pilot 3 – further improved the experimental findings: as moral blame/praise judgments correlated positively with causal responsibility judgments – as the theory of moral responsibility suggests. However, some limitations remained. The responsibility ratings for the human driver driving alone differed from AI adviser conditions. The distinction between advisory and non-advisory conditions is, however, expected. As the broad literature on responsibility attribution suggests, responsibility ratings differ when the responsible agent is assisted by a human-like agent (Harvey and Fischer 1997; Gino 2008; Dalal and Bonaccio 2010; Meshi et al. 2012) or even a tool (Santoni de Sio and Mecacci 2021; Douer and Meyer 2020; Flemisch et al. 2012). One likely explanation for the lack of distinguishing ratings is a consistency response bias where participants tried to retain a consistent responsibility rating for the human user despite changes in the experimental conditions. The within-subject design was identified as one possible reason for a consistency response bias.

The present pilot sought to overcome the previous limitations by 1. adapting the experimental design and 2. broadening the measured variables to the AI adviser and a bystander. First, pilot 5 adopted a between-subject design instead of a within-subject design. The between-subject design was supposed to reduce response bias – namely consistency, carry-over, or desire characteristics effects. Second, to have a more well-rounded control measurement, pilot 5 introduced additional measurements. Besides the human driver's blame/praise and causal responsibility ratings, pilot 5 also added blame/praise and causal responsibility ratings for the AI adviser and an uninvolved third party. The extension of the measurements allows for tracking any emergent shift in responsibility from a human driver to an artificial agent – or vice versa. As a control measure, an uninvolved third party is introduced. In the case of the car-driving domain of pilot 5, the uninvolved third party is a pedestrian without his/her wrongdoing, the subject of the driving outcome. A pattern of responsibility sharing would emerge if responsibility ratings for the human driver diminished when introduced with an AI adviser and the AI adviser retained some responsibility for the outcome. Further, the pedestrian is expected to retain little responsibility for the outcome – consistently across conditions. The worry that participants might become less accurate at representing their moral judgements due to having to answer multiple measurements per condition – their concentration might fall off – is mitigated by the reduction of scenarios each participant has to face. In other words, while the number of measurements per condition increases, the number of scenarios each participant faces minimises to one.

Pilot 5

Research Question

Does the AI adviser's presence affect the blame of the human driver?

1 Design

	AI adviser		
	Sensory	Linguistic	Absent
−			

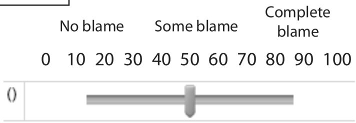
3x1 between-subject design with vignette-based diverse scenarios:
3: AI adviser (sensory, linguistic, absent)
1: Outcome (negative)
Between: only one random condition presented per participant.

2 Measurements

	H. user	AI	Pedestrian
Blame			
Cause			

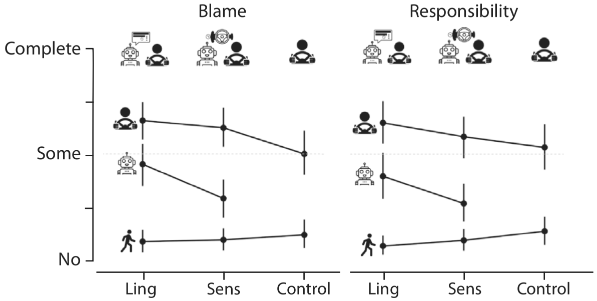
How much blame does the pedestrian deserve for the accident?

Informativity	AI
Effort	AI



3 Results

n = 150



4 Conclusion

I found a **significant agent effect**:

1. *Blame* and *Causal Responsibility* ratings were *highest* for the *driver*, second *highest* for the *AI adiviser* and *smallest* for the *pedestrian*.

Some notable **trends** include:

1. The human *driver* was generally seen as *less blameworthy* and *responsible* when acting *alone*.
2. The linguistic AI adviser was generally seen as *more blameworthy* and *responsible* than the sensory AI adviser.

Figure 8: Pilot 5 overview

Pilot 5 retained the car-driving vignettes and focused on the negative outcome condition – since the outcome effect already has been consistently replicated in the previous pilots.

Methods

Experimental design

I conducted an online study (n = 150) to elicit judgements on moral responsibility in human-assisted driving scenarios. The study used hypothetical vignettes that describe a driving scenario with a human driver and an artificial assistant resulting in a negative outcome – an accident (see case description for details on experimental conditions and supplementary methods for vignettes of studies 1 and 2). The artificial assistant was AI-powered. I used a 3x1 between-subject experimental design. I varied one condition – AI advice modality – with three factors – no AI adviser, sensory AI adviser, linguistic AI adviser. The study compared the participants' ratings of causal responsibility and blame for the AI adviser and human driver. I expected to see three main effects: an agent effect (increased involvement in action leads to higher attributed blame and responsibility – involvement here corresponds to closeness in bringing about the outcome, e.g. swerving, advising to swerve, uninvolved third-party), which could also apply to a difference in AI modalities s.t. linguistic advice closer to the external adviser (more removed from action) vs sensory advice (closer to agent performing the action).

Materials

One vignette for an assisted driving scenario was adapted to match the three experimental conditions for the study. The vignettes presented brief accounts of the situation leading to questions about individual aspects of moral responsibility for the human driver, an endangered pedestrian, and the AI adviser. The main vignette included a human driver who faces a turn with bad visibility while a pedestrian is crossing the street behind the turn and out of sight of the human driver. The changes to the vignette included a variation in AI's modality (see methods for case descriptions and supplementary methods for detailed vignettes). There was no AI adviser in the control condition, and the human was driving alone. In the other two conditions, there was an AI adviser who either provided linguistic or sensory advice. Each participant read one vignette assigned at random – using counterbalanced block randomisation – and was asked to rate the blame and causal responsibility of each agent involved (driver, pedestrian, AI adviser if applicable) alongside the level of informativity and effort for using the AI adviser (if applicable). Responses were recorded on a 100-point scale using sliders. Comparing the responses across vignettes revealed the effect of the experimental manipulations.

Data analysis

I analysed the data using general linear models (glm) from the lme4 library (Bates et al. 2015) in RStudio (Team 2021). Every model assumed a binomial distribution for the most accurate fit of the model to the data. In order to fit the model to the data, I normalised the data using min-max normalisation ((xi – min(x)) / (max(x) – min(x))). I used two glms for each main measurement (blame, causal responsibility, AI effort, AI informativity). These glm models were defined by glm(value ~ modality agent, family=binomial()). For each of the main measurements, two glms were used. One model examined the differences between the human driver and pedestrian across all three experimental conditions: no AI adviser, sensory AI adviser, linguistic AI adviser. This model shows whether 1) the perception of the human driver is affected by the variation of the kind of AI adviser and 2) the blame is shifted to a third party – the pedestrian. Another model examined the differences between all three agents – driver, pedestrian, and AI adviser – when the AI adviser was present. This model explores 1) whether the modality of the AI advice influences any of the ratings and 2) whether the AI is blamed differently by the human driver and the pedestrian.

Participants

I recruited a total of 150 participants from Prolific service. No participants were excluded. 49 % of the participants were male, 50 % were female, and 1 % stated other. 67 % of the participants had a bachelor's degree or higher. The mode and median age group was 25 to 34 years old.

Stimuli and Procedure

After a language comprehension test, participants were familiarised with the structure of the main experiment and the measurement scales. Then, participants completed a practice trial and continued with the main experiment. Here, they were first presented with a text vignette and then asked to rate the measured variables as accurately as possible. The vignette scenarios varied in AI adviser modality within a 3x1 in-between subject design. After an attention check, participants were asked to complete some basic demographic questions (age, gender, education).

Results

Blame model

Given the asymmetries in the experimental design, I ran two separate models to test any difference in blame ratings across agents (human driver, AI adviser, and pedestrian) as accurately as possible. The asymmetry exists because the AI adviser is intentionally missing in the control condition as one of three agents. Therefore, the first model compares human driver and pedestrian ratings across all three AI advice modalities (sensory, linguistic, absent). In contrast, the second model compares human driver, pedestrian, and AI adviser ratings across two AI advice modalities (sensory and linguistic).

For the first model, I fitted a logistic model (estimated using ML) to predict blame ratings with the AI advice modality and the respective agent as fixed effects (formula:

value ~ modality * agent). The model's intercept, corresponding to agent = Driver and modality = Control, is at -1.57e-03 (95 % CI [-0.55, 0.55], p = 0.996). Within this model:

The effect of modality [Ling] is statistically non-significant and positive (beta = 0.60, 95 % CI [-0.19, 1.42], p = 0.140; Std. beta = 0.60, 95 % CI [-0.19, 1.42])

The effect of modality [Sens] is statistically non-significant and positive (beta = 0.47, 95 % CI [-0.32, 1.27], p = 0.248; Std. beta = 0.47, 95 % CI [-0.32, 1.27])

The effect of agent [Pedestrian] is statistically significant and negative (beta = -1.78, 95 % CI [-2.79, -0.86], p < .001; Std. beta = -1.78, 95 % CI [-2.79, -0.86])

The effect of modality [Ling] × agent [Pedestrian] is statistically non-significant and negative (beta = -0.86, 95 % CI [-2.32, 0.55], p = 0.234; Std. beta = -0.86, 95 % CI [-2.32, 0.55])

The effect of modality [Sens] × agent [Pedestrian] is statistically non-significant and negative (beta = -0.65, 95 % CI [-2.07, 0.74], p = 0.359; Std. beta = -0.65, 95 % CI [-2.07, 0.74])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

For the second model, I fitted a logistic model (estimated using ML) to predict blame ratings with the AI advice modality and the respective agent as fixed effects (formula: value ~ modality * agent). The model's intercept, corresponding to agent = Driver and modality = Ling, is at 0.60 (95 % CI [0.03, 1.21], p = 0.044). Within this model:

The effect of modality [Sens] is statistically non-significant and negative (beta = -0.14, 95 % CI [-0.96, 0.68], p = 0.741; Std. beta = -0.14, 95 % CI [-0.96, 0.68])

The effect of agent [AI] is statistically non-significant and negative (beta = -0.78, 95 % CI [-1.61, 0.02], p = 0.058; Std. beta = -0.78, 95 % CI [-1.61, 0.02])

The effect of agent [Pedestrian] is statistically significant and negative (beta = -2.64, 95 % CI [-3.78, -1.65], p < .001; Std. beta = -2.64, 95 % CI [-3.78, -1.65])

The effect of modality [Sens] × agent [AI] is statistically non-significant and negative (beta = -0.51, 95 % CI [-1.67, 0.65], p = 0.394; Std. beta = -0.51, 95 % CI [-1.67, 0.65])

The effect of modality [Sens] × agent [Pedestrian] is statistically non-significant and positive (beta = 0.21, 95 % CI [-1.27, 1.70], p = 0.780; Std. beta = 0.21, 95 % CI [-1.27, 1.70])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Causal Responsibility model

As for the blame models, I ran two different causal responsibility models to test any difference in blame ratings across agents (human driver, AI adviser, and pedestrian) as accurately as possible. Therefore, the first model compares human driver and pedestrian ratings across all three AI advice modalities (sensory, linguistic, absent). In

contrast, the second model compares human driver, pedestrian, and AI adviser ratings across two AI advice modalities (sensory, and linguistic).
For the first model, I fitted a logistic model (estimated using ML) to predict causal responsibility ratings with the AI advice modality and the respective agent as fixed effects (formula: value ~ modality * agent). The model's intercept, corresponding to modality = Control and agent = Driver, is at 0.12 (95 % CI [-0.43, 0.67], p = 0.675). Within this model:

The effect of modality [Ling] is statistically non-significant and positive (beta = 0.45, 95 % CI [-0.35, 1.26], p = 0.275; Std. beta = 0.45, 95 % CI [-0.35, 1.26])

The effect of modality [Sens] is statistically non-significant and positive (beta = 0.19, 95 % CI [-0.60, 0.98], p = 0.636; Std. beta = 0.19, 95 % CI [-0.60, 0.98])

The effect of agent [Pedestrian] is statistically significant and negative (beta = -1.77, 95 % CI [-2.75, -0.88], p < .001; Std. beta = -1.77, 95 % CI [-2.75, -0.88])

The effect of modality [Ling] × agent [Pedestrian] is statistically non-significant and negative (beta = -1.03, 95 % CI [-2.54, 0.39], p = 0.162; Std. beta = -1.03, 95 % CI [-2.54, 0.39])

The effect of modality [Sens] × agent [Pedestrian] is statistically non-significant and negative (beta = -0.52, 95 % CI [-1.93, 0.85], p = 0.458; Std. beta = -0.52, 95 % CI [-1.93, 0.85])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

For the second model, I fitted a logistic model (estimated using ML) to predict causal responsibility ratings with the AI advice modality and the respective agent as fixed effects (formula: value ~ modality * agent). The model's intercept, corresponding to modality = Ling and agent = Driver, is at 0.56 (95 % CI [-7.09e-03, 1.17], p = 0.058). Within this model:

The effect of modality [Sens] is statistically non-significant and negative (beta = -0.26, 95 % CI [-1.07, 0.55], p = 0.534; Std. beta = -0.26, 95 % CI [-1.07, 0.55])

The effect of agent [AI] is statistically significant and negative (beta = -0.96, 95 % CI [-1.79, -0.15], p = 0.021; Std. beta = -0.96, 95 % CI [-1.79, -0.15])

The effect of agent [Pedestrian] is statistically significant and negative (beta = -2.80, 95 % CI [-4.03, -1.77], p < .001; Std. beta = -2.80, 95 % CI [-4.03, -1.77])

The effect of modality [Sens] × agent [AI] is statistically non-significant and negative (beta = -0.28, 95 % CI [-1.45, 0.89], p = 0.642; Std. beta = -0.28, 95 % CI [-1.45, 0.89])

The effect of modality [Sens] × agent [Pedestrian] is statistically non-significant and positive (beta = 0.51, 95 % CI [-1.00, 2.06], p = 0.506; Std. beta = 0.51, 95 % CI [-1.00, 2.06])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

AI informativity model

I fitted a logistic model (estimated using ML) to predict AI informativity ratings with the AI advice modality as a fixed effect (formula: value ~ modality). The model's intercept, corresponding to modality = Ling, is at -0.57 (95 % CI [-1.18, -4.53e-04], p = 0.055). Within this model:

The effect of modality [Sens] is statistically non-significant and positive (beta = 0.59, 95 % CI [-0.20, 1.41], p = 0.148; Std. beta = 0.59, 95 % CI [-0.20, 1.41])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

AI effort model

I fitted a logistic model (estimated using ML) to predict AI effort ratings with the AI advice modality as a fixed effect (formula: value ~ modality). The model's intercept, corresponding to modality = Ling, is at 0.49 (95 % CI [0.21, 0.78], p < .001). Within this model: The effect of modality [Sens] is statistically non-significant and positive (beta = 0.13, 95 % CI [-0.27, 0.54], p = 0.513; Std. beta = 0.13, 95 % CI [-0.27, 0.54])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Discussion

As expected, the human driver was blamed significantly more for the accident than the pedestrian. This effect holds across experimental conditions: whether or not an AI adviser is present – irrespective of the modality of the AI adviser. Blame ratings for the human driver or the pedestrian did not differ across conditions. In other words, the human driver and the pedestrian were blamed as much for the accident when an AI was present – irrespective of its modality – or when the AI was absent.

Causal ratings underscore the same trend. The human driver was judged as more causally responsible than the pedestrian across experimental conditions and irrespective of whether an AI adviser was present or given an AI adviser how it delivered the advice.

When the AI adviser was present, comparing the blame and causality of all three agents revealed some additional nuances. Overall, the AI adviser was blamed less than the human driver and more than the pedestrian. Especially in the case of a sensory AI adviser, the human driver was blamed significantly more than the AI adviser. Causal responsibility ratings mirrored blame ratings, and the human was also seen as significantly more causally responsible for the accident than the AI adviser or the pedestrian.

Examining the perceived ease of use and level of informativity, participants judged both to be similar for the linguistic and the sensory AI adviser.

The consistently small amount of blame and causal responsibility attributed to the pedestrian demonstrates the robustness of the blame and responsibility attribution ratings for both the human driver and the AI adviser. No scapegoat exists, and most blame and responsibility are accounted for. Arguably, other parties could be seen as sharing responsibility for the accident – the developers of the AI adviser, the car manufacturer, the road designers etc. – but given the lack of involvement in bringing about the particular accident, they are neglected as relevant parties. They mainly could not have brought about a different outcome – there is nothing they could have done.

Limitations

One great advantage of the newly adapted between-subject design is that the control condition works. The human driver is reliably judged differently when comparing scenarios with and without an AI adviser. In addition, no blame diffusion occurred towards the third-party pedestrian, whose involvement never changed. The pedestrian is held similarly not blameworthy for the outcome across experimental conditions. Any allocation of blame was, therefore, the result of a human-AI-adviser pairing. The human and the AI adviser are also significantly more responsible for the outcome than the pedestrian.

However, despite the desired effects, there are also systematic limitations. It is unclear whether and how blame and responsibility for the human driver differ across conditions. Though a non-significant trend exists, blame and responsibility ratings are lower in the control condition where no AI adviser was present than in the experimental condition where a linguistic AI is present. This could be a problem of a small sample and effect size. If found significant, this would mean that AI is perceived as a tool more than an agent. While generally, a present AI adviser lowers the responsibility of a human user, a present sensory, tool-like AI adviser would simultaneously heighten the responsibility for the human user as the user is now more able to avoid the negative outcome. However, it would need an increase in sample size to conclude.

Conclusion

The experimental design changes introduced by pilot 5 – moving from a within to a between-subject design and expanding the measured variables – have substantially improved results. The pilot found that responsibility and blame are allocated across only the involved human user and the AI adviser. A third-party pedestrian, the target of the resulting accident, was consistently neither blameworthy nor causally responsible for the accident – in contrast to the human user and the AI adviser. The human user was seen as more blameworthy and causally responsible than the AI adviser when present. However, given the small sample and effect size, little can be said about the significance of the results. It remains, therefore, unclear whether results can be replicated and even extended to the other measured variables. This holds especially for AI informativity and effort ratings where no trend emerged.

3.2.6 Pilot 6

Introduction

Pilot 6 builds on the motivation and findings from pilot 5 but addresses the limitations of pilot 5 by increasing the sample size. Pilot 6 asks how responsibility is allocated between multiple agents (human driver, AI adviser and pedestrian) in a between-subject design. As in pilot 5, pilot 6 expects that A pattern of responsibility sharing would emerge if responsibility ratings for the human driver diminished when introduced with an AI adviser and the AI adviser retained some responsibility for the outcome. The pedestrian is also expected to retain little responsibility for the outcome – consistently across conditions.

Methods

Experimental design
Same as in pilot 5.

Materials
Same as in pilot 5.

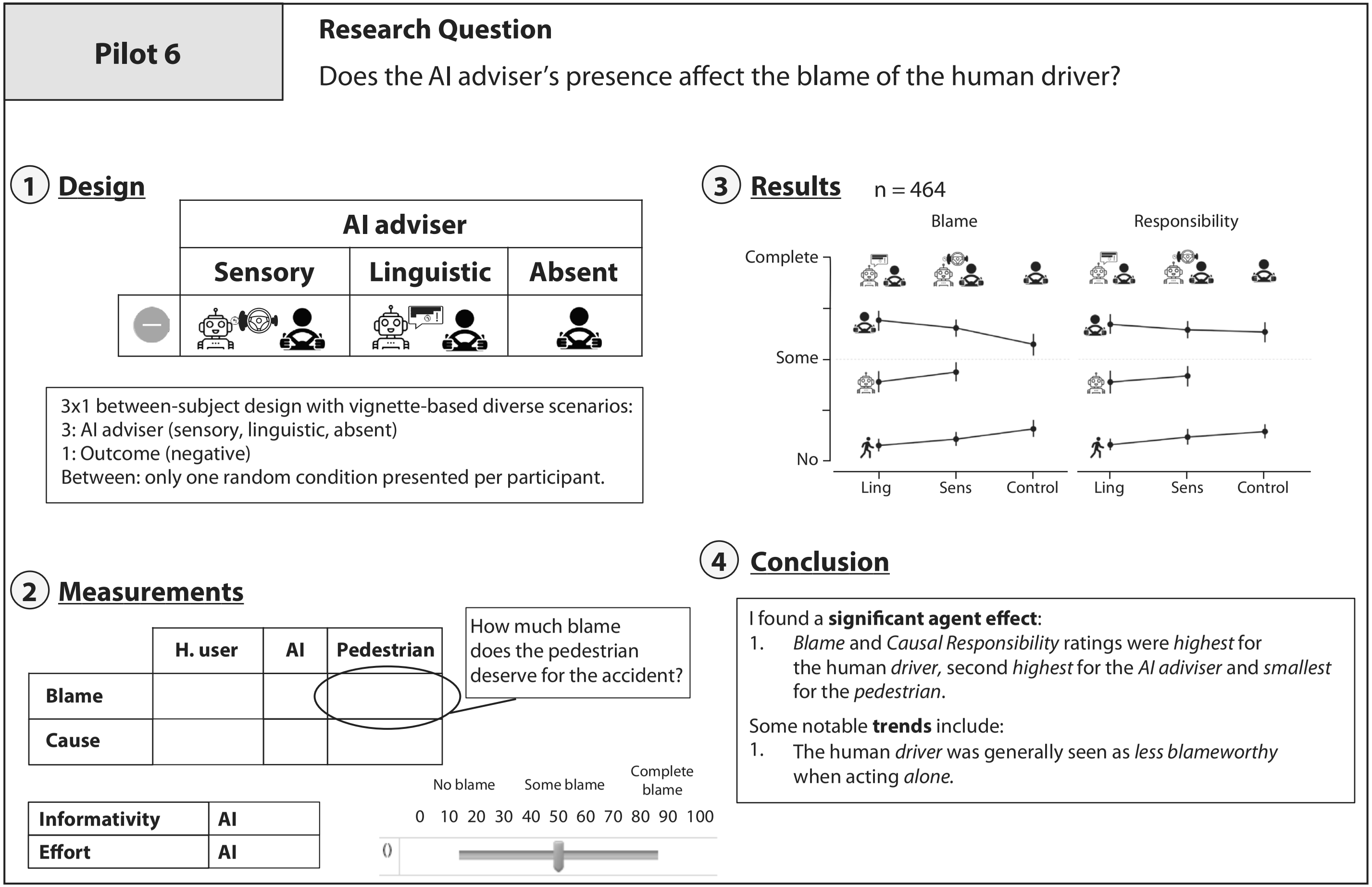
Data analysis
Same as in pilot 5.

Participants
I recruited a total of 464 participants from Prolific service. No participants were excluded. 47 % of the participants were male, 52 % were female, and 1 % stated other or preferred not to say. 61 % of the participants had a bachelor's degree or higher. The mode and median age group 25 to 34 years old.

Stimuli and Procedure
Same as in pilot 5.

Results

Blame model
Given the asymmetries in the experimental design, I ran two separate models to test any difference in blame ratings across agents (human driver, AI adviser, and pedestrian) as accurately as possible. The asymmetry exists because the AI adviser is intentionally missing in the control condition as one of three agents. Therefore, the first model compares human driver and pedestrian ratings across all three AI advice modalities (sensory, linguistic, absent). In contrast, the second model compares human driver, pedestrian, and AI adviser ratings across two AI advice modalities (sensory and linguistic).



Pilot 6

Research Question

Does the AI adviser's presence affect the blame of the human driver?

1 Design

	AI adviser		
	Sensory	**Linguistic**	**Absent**
−			

3x1 between-subject design with vignette-based diverse scenarios:
3: AI adviser (sensory, linguistic, absent)
1: Outcome (negative)
Between: only one random condition presented per participant.

2 Measurements

	H. user	**AI**	**Pedestrian**
Blame			
Cause			

Informativity	**AI**
Effort	**AI**

How much blame does the pedestrian deserve for the accident?

3 Results n = 464

4 Conclusion

I found a **significant agent effect**:

1. *Blame* and *Causal Responsibility* ratings were *highest* for the human *driver*, second *highest* for the *AI adiviser* and *smallest* for the *pedestrian*.

Some notable **trends** include:

1. The human *driver* was generally seen as *less blameworthy* when acting *alone*.

Figure 9: Pilot 6 overview

For the first model, I fitted a logistic model (estimated using ML) to predict blame ratings with the AI advice modality and the respective agent as fixed effects (formula: value ~ modality * agent). The model's intercept, corresponding to agent = Driver, modality = Control, is at 0.28 (95 % CI [-0.03, 0.61], p = 0.079). Within this model:

The effect of modality [Ling] is statistically significant and positive (beta = 0.49, 95 % CI [0.02, 0.96], p = 0.040; Std. beta = 0.49, 95 % CI [0.02, 0.96])

The effect of modality [Sens] is statistically non-significant and positive (beta = 0.32, 95 % CI [-0.13, 0.78], p = 0.168; Std. beta = 0.32, 95 % CI [-0.13, 0.78])

The effect of agent [Pedestrian] is statistically significant and negative (beta = -1.85, 95 % CI [-2.39, -1.34], p < .001; Std. beta = -1.85, 95 % CI [-2.39, -1.34])

The effect of modality [Ling] × agent [Pedestrian] is statistically significant and negative (beta = -1.18, 95 % CI [-2.02, -0.36], p = 0.005; Std. beta = -1.18, 95 % CI [-2.02, -0.36])

The effect of modality [Sens] × agent [Pedestrian] is statistically non-significant and negative (beta = -0.70, 95 % CI [-1.49, 0.07], p = 0.077; Std. beta = -0.70, 95 % CI [-1.49, 0.07])

For the second model, I fitted a logistic model (estimated using ML) to predict blame ratings with the AI advice modality and the respective agent as fixed effects (formula: value ~ modality * agent). The model's intercept, corresponding to modality = Ling, agent = Driver, is at 0.77 (95 % CI [0.44, 1.12], p < .001). Within this model:

The effect of modality [Sens] is statistically non-significant and negative (beta = -0.16, 95 % CI [-0.64, 0.31], p = 0.495; Std. beta = -0.16, 95 % CI [-0.64, 0.31])

The effect of agent [AI] is statistically significant and negative (beta = -1.20, 95 % CI [-1.68, -0.74], p < .001; Std. beta = -1.20, 95 % CI [-1.68, -0.74])

The effect of agent [Pedestrian] is statistically significant and negative (beta = -3.03, 95 % CI [-3.70, -2.42], p < .001; Std. beta = -3.03, 95 % CI [-3.70, -2.42])

The effect of modality [Sens] × agent [AI] is statistically non-significant and positive (beta = 0.35, 95 % CI [-0.30, 1.01], p = 0.291; Std. beta = 0.35, 95 % CI [-0.30, 1.01])

The effect of modality [Sens] × agent [Pedestrian] is statistically non-significant and positive (beta = 0.47, 95 % CI [-0.38, 1.34], p = 0.282; Std. beta = 0.47, 95 % CI [-0.38, 1.34])

Responsibility model

As for the blame models, I ran two different causal responsibility models to test any difference in blame ratings across agents (human driver, AI adviser, and pedestrian) as accurately as possible. Therefore, the first model compares human driver and pedestrian ratings across all three AI advice modalities (sensory, linguistic, absent). In contrast, the second model compares human driver, pedestrian, and AI adviser ratings across two AI advice modalities (sensory and linguistic).

For the first model, I fitted a logistic model (estimated using ML) to predict causal responsibility ratings with the AI advice modality and the respective agent as fixed effects (formula: value ~ modality * agent). The model's intercept, corresponding to modality = Control & agent = Driver, is at 0.53 (95 % CI [0.20, 0.86], p = 0.002). Within this model:

The effect of modality [Ling] is statistically non-significant and positive (beta = 0.16, 95 % CI [-0.31, 0.63], p = 0.509; Std. beta = 0.16, 95 % CI [-0.31, 0.63])

The effect of modality [Sens] is statistically non-significant and positive (beta = 0.04, 95 % CI [-0.42, 0.51], p = 0.854; Std. beta = 0.04, 95 % CI [-0.42, 0.51])

The effect of agent [Pedestrian] is statistically significant and negative (beta = -2.19, 95 % CI [-2.74, -1.66], p < .001; Std. beta = -2.19, 95 % CI [-2.74, -1.66])

The effect of modality [Ling] × agent [Pedestrian] is statistically non-significant and negative (beta = -0.72, 95 % CI [-1.56, 0.10], p = 0.088; Std. beta = -0.72, 95 % CI [-1.56, 0.10])

The effect of modality [Sens] × agent [Pedestrian] is statistically non-significant and negative (beta = -0.25, 95 % CI[-1.04, 0.53], p = 0.531; Std. beta = -0.25, 95 % CI [-1.04, 0.53])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

For the second model, I fitted a logistic model (estimated using ML) to predict causal responsibility ratings with the AI advice modality and the respective agent as fixed effects (formula: value ~ modality * agent). The model's intercept, corresponding to modality = Ling & agent = Driver, is at 0.68 (95 % CI [0.35, 1.02], p < .001). Within this model:

The effect of modality [Sens] is statistically non-significant and negative (beta = -0.11, 95 % CI [-0.58, 0.35], p = 0.634; Std. beta = -0.11, 95 % CI [-0.58, 0.35])

The effect of agent [AI] is statistically significant and negative (beta = -1.12, 95 % CI [-1.59, -0.66], p < .001; Std. beta = -1.12, 95 % CI [-1.59, -0.66])

The effect of agent [Pedestrian] is statistically significant and negative (beta = -2.91, 95 % CI [-3.57, -2.31], p < .001; Std. beta = -2.91, 95 % CI [-3.57, -2.31])

The effect of modality [Sens] × agent [AI] is statistically non-significant and positive (beta = 0.23, 95 % CI [-0.42, 0.89], p = 0.482; Std. beta = 0.23, 95 % CI [-0.42, 0.89])

The effect of modality [Sens] × agent [Pedestrian] is statistically non-significant and positive (beta = 0.47, 95 % CI [-0.37, 1.33], p = 0.275; Std. beta = 0.47, 95 % CI [-0.37, 1.33])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

AI effort model

I fitted a logistic model (estimated using ML)to predict AI informativity ratings with the AI advice modality as a fixed effect (formula: value ~ modality). The model's intercept, corresponding to modality = Ling, is at 0.51 (95 % CI [0.19, 0.84], p = 0.002). Within this model:

The effect of modality [Sens] is statistically non-significant and positive (beta = 0.11, 95 % CI [-0.35, 0.58], p = 0.630; Std. beta = 0.11, 95 % CI [-0.35, 0.58])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

AI informativity model

I fitted a logistic model (estimated using ML) to predict AI effort ratings with the AI advice modality as a fixed effect (formula: value ~ modality). The model's intercept, corresponding to modality = Ling, is at -0.23 (95 % CI [-0.55, 0.09], p = 0.163). Within this model:

The effect of modality [Sens] is statistically non-significant and negative (beta = -0.09, 95 % CI [-0.54, 0.36], p = 0.699; Std. beta = -0.09, 95 % CI [-0.54, 0.36])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Discussion

The increased sample size has replicated and verified the observed effects from pilot 5. The human driver is blamed less without an AI than with a linguistic AI adviser. However, this did not apply to the sensory AI adviser suggesting either an invalid control condition – as the human driver was blamed with or without a sensory AI adviser, which conceptually should not occur – or a substantial difference in how the AI advisers are perceived. Here, the sensory AI adviser might be perceived as a tool heightening the user's responsibility – thereby equalising a possible responsibility decrease from having an AI adviser at his/her side. This compensating effect did not hold for the linguistic AI adviser, which was seen as an agent resulting in a blame-sharing pattern – where blame is distributed across involved agents (human driver and the AI adviser). The latter hypothesis, however, lacks support as this asymmetric effect only holds for blame but not causal responsibility judgments. Here, despite the increase in sample size, the human driver is judged as causally responsible with or without an AI adviser.

The worry about the inadequacy of the control condition (AI present vs absent) is amplified as the follow-up measurements of AI informativity and effort did not reveal any differences between the two AI advisers – even though a difference in the blame of the human user is observed. The linguistic and the sensory AI advisers were judged just as informative and easy to perceive as each other. Thus, the advice modality cannot explain asymmetric blame judgments. The observed difference in blame rating for the human user in no vs linguistic AI adviser conditions is likely the product of an unbalanced experimental design.

Besides the blame judgments of the human driver, pilot 6 also replicated other critical findings from pilot 5: responsibility sharing is confined to the human driver and the AI adviser. The pedestrian was rated consistently across experimental conditions as having neither blame nor causal responsibility for the outcome. This is encouraging since participants did not feel the need to assign a responsibility scapegoat for a possibly confusing and overwhelming responsibility dilemma.

Limitations

Despite an increase in sample size in pilot 6, the control condition (with no AI present) still was not sufficiently different from the experimental conditions (with AI present). Tweaking the experimental designs from mixed to within to between has not resolved this issue. The between-subject design appeared as the most promising, though, as it reduced response biases as best as possible – which had been in problem in the previous four pilots.

However, some other significant limitations are evident. First, the asymmetric experimental design limits the statistical analysis of the experimental data. While the human user was present across all six conditions, the AI adviser only appeared in four out of six conditions. It is hence only possible to compare the distribution of responsibility across human and AI adviser when the AI adviser is present. An improved control condition would include an inactive AI adviser to strengthen the contrast between driving alone, where most of the responsibility is on the human driver and being assisted. Rendering the experimental design more symmetric would otherwise reduce the number of needed linear models from two to one, as one model could incorporate the changes in the measured variables across all experimental conditions, increasing the statistical analysis's robustness.

An improved control condition would likewise address the inconsistent difference between the conditions with and without an AI adviser across measured variables. One reason for the inconsistency could be the ill-suited comparison between driving alone and having an active AI adviser present. The mere presence of an AI adviser – and not the advice itself – could explain the observed difference. An improved control could compare active vs inactive AI adviser conditions. The AI adviser would be present in all conditions, and only the given advice would be different.

Conclusion

Pilots 5 and 6 introduced some critical improvements over the previous pilots. Extending the measured variables to include the AI adviser alongside the human user to test how the AI adviser is perceived provides a more accurate picture of how responsibility is shifted between the human user and the AI adviser. Pilot 6 replicated the main trends and effects observed in pilot 5 with a larger sample size. Pilot 6 found that the human driver is blamed less without an AI than with a linguistic AI adviser.

3.2.7 Pilot 7

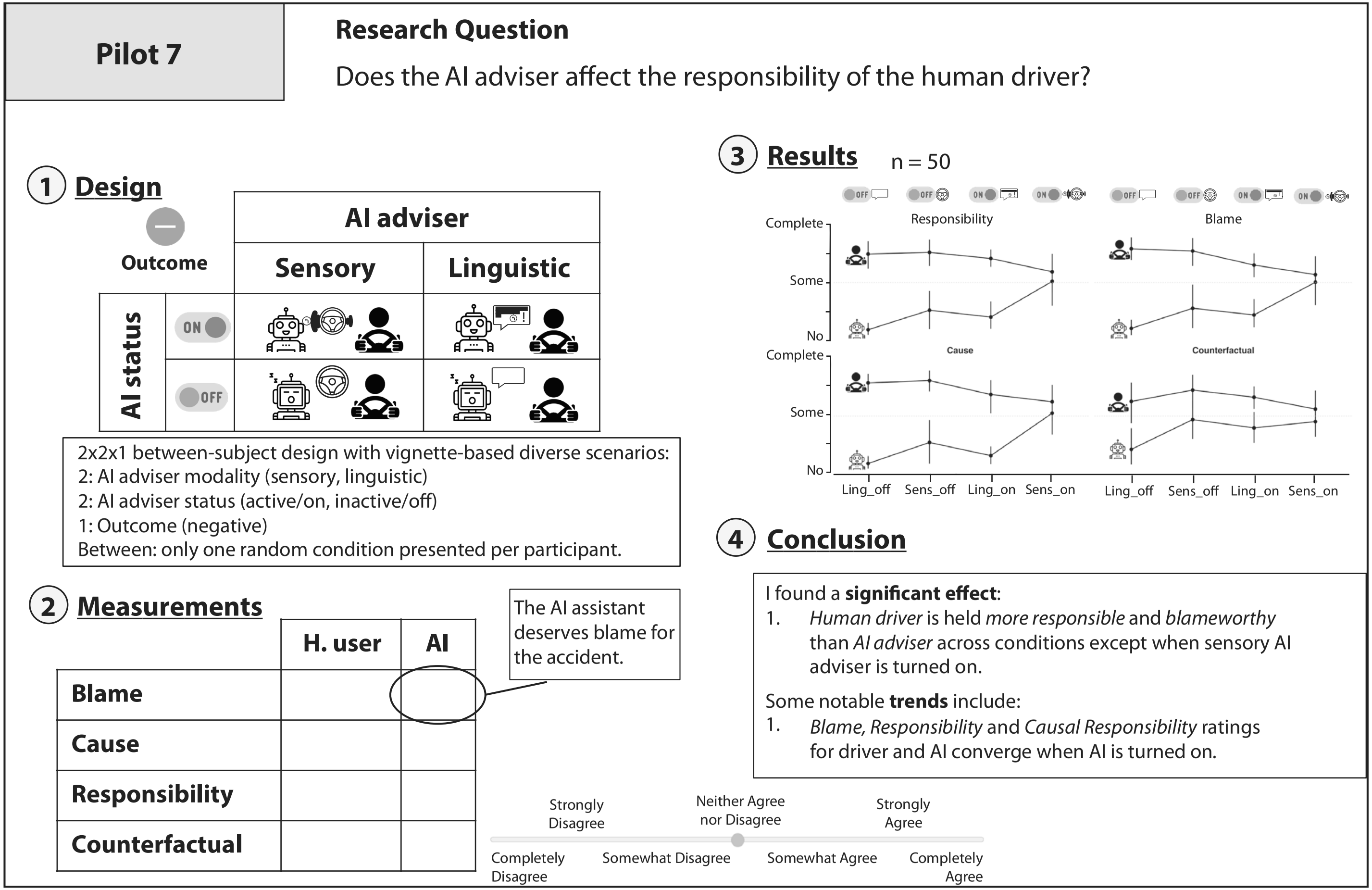
Introduction

Pilots 5 and 6 provided significant improvements to testing human-AI-advice pairing scenarios. Extending the measured variables beyond the human user allowed examining the emergent pattern of responsibility sharing between the human user and the AI adviser. The between-subject design reduced response biases and promoted a leaner comparison of experimental conditions across participants.

Pilot 7 retained these fundamental improvements and addressed the limitations of their experimental design – notably the validity of the control condition. Previously the control condition included a human user without an AI adviser. Any difference in the human user's attributed responsibility once an AI adviser is introduced should be due to the AI adviser. However, it remained to be seen whether the effect was due to the mere presence of an AI adviser or the given advice. Having no AI adviser present in the control condition further created problems for the statistical analysis because no singular model could capture all changes in available agents across all conditions for each dependent variable (the AI adviser was missing in the control condition).

Pilot 7 varied the experimental design by rethinking the control condition and extending the measured variables to address these limitations. To better compare the effect of being advised by an AI system and rule out that the observed effect is caused by the mere presence of the AI system, pilot 7 replaced the AI-absence condition with a two-level factor of AI status. The AI adviser could either be active (on) or inactive (off) but was always mentioned as part of the presented vignettes. I expected that adopting a 2x2x2 (AI advice modality, AI status, outcome) instead of a 3x2 (AI advice modality, outcome) would improve the comparability of the control conditions where the AI adviser was inactive in the experimental conditions where the AI adviser was active.

The additional measured variables included counterfactual capacity, as a measurement of how much the human user or the AI adviser could have made a difference to the outcome, and moral responsibility. Counterfactual capacity is closely connected to moral responsibility (Chockler and Halpern 2004). Suppose there is nothing an agent can do to alter an outcome, i.e. the agent has no counterfactual capacity. In that case, the agent is also not morally responsible for the outcome: just like an aeroplane pilot is responsible for crashing a fully functional aeroplane into a mountain and is not responsible for crashing an aeroplane that has lost any control. I included counterfactual measurement to pinpoint an additional potential differentiating dimension of moral responsibility attribution between the human user and the AI adviser. I expected that the human driver's counterfactual capacity would be greater than the AI adviser's and that the linguistic, more agent-like AI adviser possibly has more counterfactual capacity than the sensory, more tool-like AI adviser.



Pilot 7

Research Question

Does the AI adviser affect the responsibility of the human driver?

1 **Design**

Outcome —

		AI adviser	
		Sensory	Linguistic
AI status	ON		
	OFF		

2x2x1 between-subject design with vignette-based diverse scenarios:
2: AI adviser modality (sensory, linguistic)
2: AI adviser status (active/on, inactive/off)
1: Outcome (negative)
Between: only one random condition presented per participant.

2 **Measurements**

	H. user	AI
Blame		
Cause		
Responsibility		
Counterfactual		

The AI assistant deserves blame for the accident.

Completely Disagree · Strongly Disagree · Somewhat Disagree · Neither Agree nor Disagree · Somewhat Agree · Strongly Agree · Completely Agree

3 **Results** n = 50

4 **Conclusion**

I found a **significant effect**:

1. *Human driver* is held *more responsible* and *blameworthy* than *AI adviser* across conditions except when sensory AI adviser is turned on.

Some notable **trends** include:

1. *Blame, Responsibility* and *Causal Responsibility* ratings for driver and AI converge when AI is turned on.

Figure 10: Pilot 7 overview

To test the updated experimental design, pilot 7 focuses – similarly to pilots 5 and 6 – on the negative outcome scenarios. As the previous pilots showed a consistent attribution of responsibility to the human driver and the AI adviser, pilot 7 dropped the pedestrian to simplify the presented scenarios.

Methods

Material

One vignette for an assisted driving scenario was adapted to match the four experimental conditions for the study. The vignettes presented brief accounts of the situation leading to questions about individual aspects of moral responsibility for the human driver and the AI assistant. The main vignette included a human driver who faces a junction in bad visibility while another car is approaching with priority from the right. The changes to the vignette included a variation in outcome, a variation in AI's status, and a variation in AI's modality (see methods for case descriptions and supplementary methods for detailed vignettes). Each participant read one vignette assigned at random – using counterbalanced block randomisation – and was asked to indicate his/her agreement with statements like 'The sensory AI assistant deserves blame for the accident.' Responses were recorded on a 200-point scale using sliders (from -100 for 'Completely disagree' to 100 for 'Completely agree'). Comparing the responses across vignettes revealed the effect of the experimental manipulations.

Participants

I recruited a total of 50 participants from Amazon's MTurk service. No participants were excluded. 74 % of the participants were male and 26 % were female. 52 % of the participants had a bachelor's degree or higher. The mode age group was 25 to 34 years old. The median age group was 35 to 44 years old.

Data analysis

I analysed our data using general linear models (glm) from the lme4 library (Bates et al. 2015) in RStudio (Team 2021). Every model assumed a binomial distribution for the most accurate fit of the model to the data. In order to fit the model to the data, I normalised the data using min-max normalisation ((xi -- min(x)) / (max(x) -- min(x))). I used one glm for each of the main measurements (responsibility, blame, causal responsibility, and counterfactual capacity). These glm models were defined by glm(value ~ modality agent status, family=binomial()). Each model evaluated the difference towards the measured variable based on changes of the experimental condition (linguistic vs sensory AI adviser and active vs inactive AI adviser) for each agent involved (human driver vs AI adviser).

Stimuli and Procedure

After a language comprehension test, participants were familiarised with the structure of the main experiment and the measurement scales. Then, participants completed a practice trial and continued with the main experiment. Here, they were presented with a text vignette and then asked to rate the measured variables as accurately as possible. The vignette scenarios varied in status and modality within a 2x2 in-between subject design. After completing an attention check, participants were asked to complete some basic demographic questions (age, gender, education), their familiarity with artificial intelligence, and their experience with computer programming.

Results

Responsibility model

I fitted a logistic model (estimated using ML) to predict responsibility ratings with the AI advice modality, AI status and agent as fixed effects (formula: value ~ value ~ modality agent status). The model's intercept, corresponding to modality = linguistic, agent = Driver and status = active, is at 0.99 (95 % CI [-0.26, 2.51], p = 0.146). Within this model:

The effect of modality [sensory] is statistically non-significant and negative (beta = -0.57, 95 % CI [-2.54, 1.32], p = 0.551; Std. beta = -0.57, 95 % CI [-2.54, 1.32])

The effect of agent [AI] is statistically significant and negative (beta = -2.55, 95 % CI [-4.89, -0.66], p = 0.015; Std. beta = -2.55, 95 % CI [-4.89, -0.66])

The effect of status [inactive] is statistically non-significant and positive (beta = 0.23, 95 % CI [-1.78, 2.34], p = 0.824; Std. beta = 0.23, 95 % CI [-1.78, 2.34])

The effect of modality [sensory] × agent [AI] is statistically non-significant and positive (beta = 2.18, 95 % CI [-0.53, 5.11], p = 0.123; Std. beta = 2.18, 95 % CI [-0.53, 5.11])

The effect of modality [sensory] × status [inactive] is statistically non-significant and positive (beta = 0.67, 95 % CI [-2.19, 3.58], p = 0.645; Std. beta = 0.67, 95 % CI [-2.19, 3.58])

The effect of agent [AI] × status [inactive] is statistically non-significant and negative (beta = -1.55, 95 % CI [-6.65, 1.91], p = 0.418; Std. beta = -1.55, 95 % CI [-6.65, 1.91])

The effect of (modality [sensory] × agent [AI]) × status [inactive] is statistically non-significant and negative (beta = -0.55, 95 % CI [-5.09, 5.07], p = 0.817; Std. beta = -0.55, 95 % CI [-5.09, 5.07])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Blame model

I fitted a logistic model (estimated using ML) to predict blame ratings with the AI advice modality, AI status and agent as fixed effects (formula: value ~ modality agent status). The model's intercept, corresponding to modality = linguistic, agent = Driver and status = active, is at 0.69 (95 % CI [-0.52, 2.08], p = 0.281). Within this model:

The effect of modality [sensory] is statistically non-significant and negative (beta = -0.39, 95 % CI [-2.26, 1.45], p = 0.677; Std. beta = -0.39, 95 % CI [-2.26, 1.45])

The effect of agent [AI] is statistically significant and negative (beta = -2.12, 95 % CI [-4.31, -0.29], p = 0.033; Std. beta = -2.12, 95 % CI [-4.31, -0.29])

The effect of status [inactive] is statistically non-significant and positive (beta = 0.83, 95 % CI [-1.15, 3.13], p = 0.426; Std. beta = 0.83, 95 % CI [-1.15, 3.13])

The effect of modality [sensory] × agent [AI] is statistically non-significant and positive (beta = 1.83, 95 % CI [-0.82, 4.65], p = 0.183; Std. beta = 1.83, 95 % CI [-0.82, 4.65])

The effect of modality [sensory] × status [inactive] is statistically non-significant and positive (beta = 0.25, 95 % CI [-2.73, 3.19], p = 0.866; Std. beta = 0.25, 95 % CI [-2.73, 3.19])

The effect of agent [AI] × status [inactive] is statistically non-significant and negative (beta = -2.06, 95 % CI [-6.54, 1.26], p = 0.258; Std. beta = -2.06, 95 % CI [-6.54, 1.26])

The effect of (modality [sensory] × agent [AI]) × status [inactive] is statistically non-significant and negative (beta = -0.10, 95 % CI [-4.53, 5.01], p = 0.965; Std. beta = -0.10, 95 % CI [-4.53, 5.01])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Causal responsibility model

I fitted a logistic model (estimated using ML) to predict causal responsibility ratings with the AI advice modality, AI status and agent as fixed effects (formula: value ~ modality agent status). The model's intercept, corresponding to modality = linguistic, agent = Driver and status = active, is at 0.74 (95 % CI [-0.47, 2.15], p = 0.249). Within this model:

The effect of modality [sensory] is statistically non-significant and negative (beta = -0.31, 95 % CI [-2.24, 1.60], p = 0.745; Std. beta = -0.31, 95 % CI [-2.24, 1.60])

The effect of agent [AI] is statistically significant and negative (beta = -2.82, 95 % CI [-5.47, -0.83], p = 0.012; Std. beta = -2.82, 95 % CI [-5.47, -0.83])

The effect of status [inactive] is statistically non-significant and positive (beta = 0.61, 95 % CI [-1.41, 2.89], p = 0.559; Std. beta = 0.61, 95 % CI [-1.41, 2.89])

The effect of modality [sensory] × agent [AI] is statistically non-significant and positive (beta = 2.37, 95 % CI [-0.44, 5.54], p = 0.110; Std. beta = 2.37, 95 % CI [-0.44, 5.54])

The effect of modality [sensory] × status [inactive] is statistically non-significant and positive (beta = 0.46, 95 % CI [-2.57, 3.56], p = 0.760; Std. beta = 0.46, 95 % CI [-2.57, 3.56])

The effect of agent [AI] × status [inactive] is statistically non-significant and negative (beta = -1.61, 95 % CI [-6.98, 2.11], p = 0.423; Std. beta = -1.61, 95 % CI [-6.98,2.11])

The effect of (modality [sensory] × agent [AI]) × status [inactive] is statistically non-significant and negative (beta = -0.67, 95 % CI [-5.46, 5.21], p = 0.788; Std. beta = -0.67, 95 % CI [-5.46, 5.21])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Counterfactual capacity model

I fitted a logistic model (estimated using ML) to predict counterfactual capacity ratings with the AI advice modality, AI status and agent as fixed effects (formula: value ~ modality agent status). The model's intercept, corresponding to modality = linguistic, agent = Driver and status = active, is at 0.74 (95 % CI [-0.47, 2.15], p = 0.249). Within this model:

The effect of modality [sensory] is statistically non-significant and negative (beta = -0.48, 95 % CI [-2.37, 1.36], p = 0.607; Std. beta = -0.48, 95 % CI [-2.37, 1.36])

The effect of agent [AI] is statistically non-significant and negative (beta = -1.20, 95 % CI [-3.06, 0.50], p = 0.178; Std. beta = -1.20, 95 % CI [-3.06, 0.50])

The effect of status [inactive] is statistically non-significant and negative (beta = -0.18, 95 % CI [-2.03, 1.66], p = 0.848; Std. beta = -0.18, 95 % CI [-2.03, 1.66])

The effect of modality [sensory] × agent [AI] is statistically non-significant and positive (beta = 0.73, 95 % CI [-1.84, 3.33], p = 0.578; Std. beta = 0.73, 95 % CI [-1.84, 3.33])

The effect of modality [sensory] × status [inactive] is statistically non-significant and positive (beta = 0.99, 95 % CI [-1.65, 3.72], p = 0.464; Std. beta = 0.99, 95 % CI [-1.65, 3.72])

The effect of agent [AI] × status [inactive] is statistically non-significant and negative (beta = -0.86, 95 % CI [-3.70, 1.81], p = 0.535; Std. beta = -0.86, 95 % CI [-3.70, 1.81])

The effect of (modality [sensory] × agent [AI]) × status [inactive] is statistically non-significant and positive (beta = 0.12, 95 % CI [-3.67, 3.96], p = 0.951; Std. beta = 0.12, 95 % CI [-3.67, 3.96])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Discussion

Overall, pilot 7 introduced some significant experimental changes. Pilot 7 replaced the previous control condition of comparing a human driving alone with a human being assisted by an AI adviser with a two-level control condition. The new control condition introduced AI status (active/on vs inactive/off) to examine the influence of active AI advice on responsibility attribution more clearly. Previously, any observed effect could be explained due to the mere presence of an AI adviser. However, the new experimental addition eliminated this confound by having an AI adviser present across all experimental conditions but retaining the comparison of advised vs non-advised driving by modulating the AI adviser's status between active and inactive.

Pilot 7 found that the human driver is significantly more responsible, blameworthy, and causally connected to the outcome than the linguistic AI adviser. The observed effect did not hold for the sensory AI adviser as the attributed responsibility of the human driver and AI adviser is close to identical when an active, sensory AI adviser accompanies the human driver. Some notable trends include that the sensory AI adviser is seen as more responsible for the outcome than the linguistic AI adviser – even though the observed trend could also be the artefact of small sample size and high variability in the measured variables.

Limitations

One challenge for pilot 7 is that the responsibility measures for the AI adviser and the human driver were similar for the active sensory and the inactive linguistic AI adviser conditions. Although, as expected, responsibility ratings for either the sensory or the linguistic AI advisers increased once the AI adviser was active rather than inactive. Responsibility ratings should not be lower in the case of an active, sensory AI adviser than in an inactive, linguistic AI adviser. While this subtle trend could result from the small sample size and high variability in participants' responses, it would question the experimental validity if it persisted with an increase in sample size.

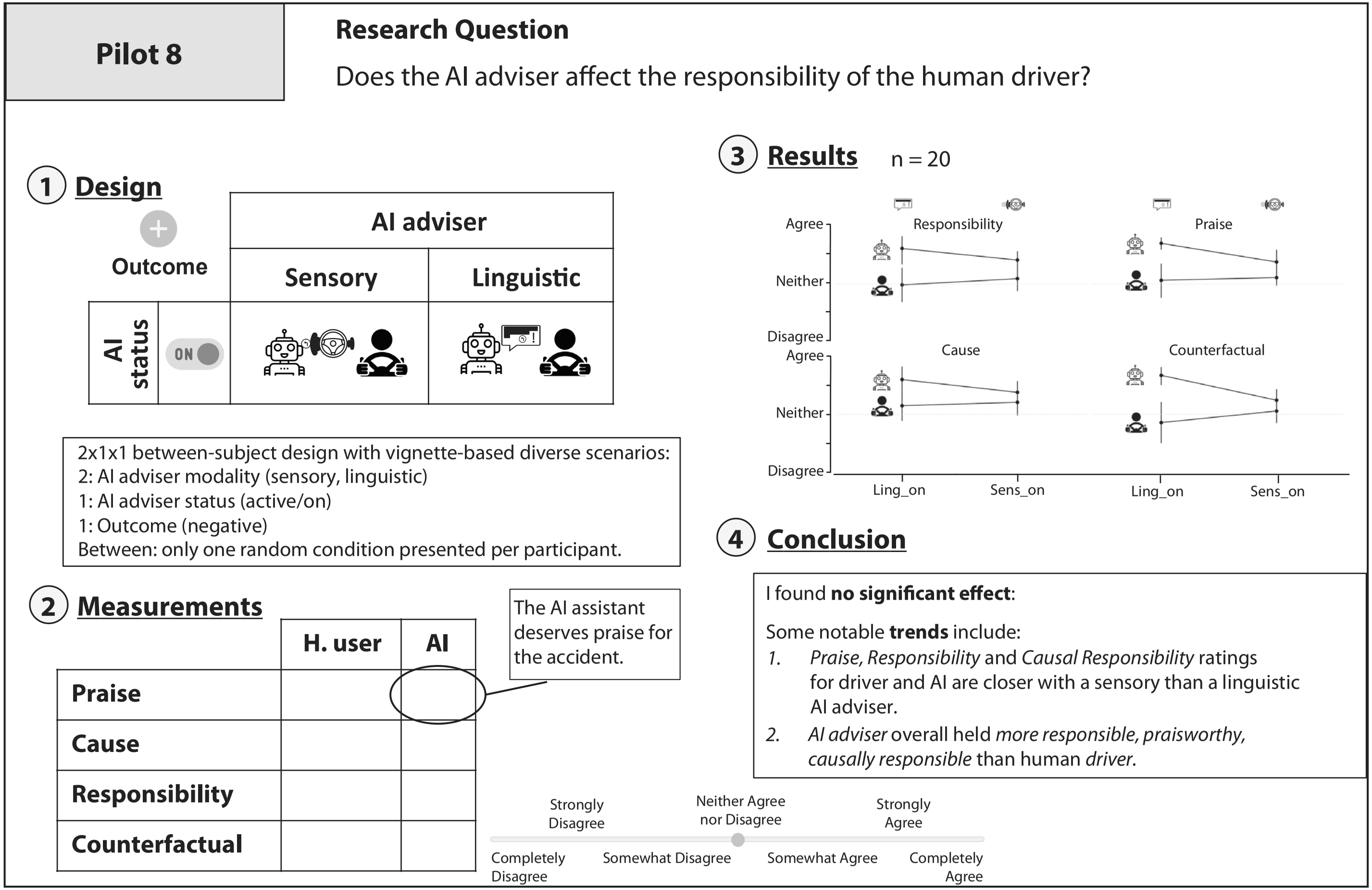
Conclusion

The experimental changes introduced by pilot 7 have provided notable improvement. Responsibility ratings were consistent across measured variables and conditions. The new design eliminated a possible AI-presence confound and established that the human driver is generally more responsible than the AI adviser. A trend for responsibility sharing also emerged as blame, responsibility, and causal responsibility ratings for a human driver and AI adviser converge when the AI adviser is active rather than inactive. The next step is extending the experimental design for a positive outcome condition.

3.2.8 Pilot 8

Introduction

Pilot 7 introduced some critical changes to the experimental design – notably a change in the control condition. Instead of comparing cases of a human driving without to cases of a human driving with an AI adviser, pilot 7 compared cases of a human driving with either an active or an inactive AI adviser The goal was two-fold. First, the change increased the comparability of non-AI to AI adviser cases because any observed difference across conditions could no longer be explained by having an AI adviser present rather than receiving AI-generated advice. Second, the change simplified the statistical analysis by making the design symmetric. With the AI adviser present in all conditions, one model can capture any difference between the driver and the AI adviser across conditions for each measurement. However, pilot 7 only tested the new design in a negative outcome scenario. Whether the experimental improvements would carry



Pilot 8

Research Question

Does the AI adviser affect the responsibility of the human driver?

1 Design

+ Outcome

		AI adviser	
		Sensory	Linguistic
AI status	ON		

2x1x1 between-subject design with vignette-based diverse scenarios:
2: AI adviser modality (sensory, linguistic)
1: AI adviser status (active/on)
1: Outcome (negative)
Between: only one random condition presented per participant.

2 Measurements

	H. user	AI
Praise		
Cause		
Responsibility		
Counterfactual		

The AI assistant deserves praise for the accident.

3 Results n = 20

4 Conclusion

I found **no significant effect**:

Some notable **trends** include:

1. *Praise*, *Responsibility* and *Causal Responsibility* ratings for driver and AI are closer with a sensory than a linguistic AI adviser.
2. *AI adviser* overall held *more responsible, praisworthy, causally responsible* than human *driver.*

Figure 11: Pilot 8 overview

over to a positive outcome scenario remained. Pilot 8 sought to address this question. To simplify the experimental design, pilot 8 used the same experimental design as pilot 7 but focused on active AI adviser conditions, as this is where any possible difference in the responsibility ratings is expected to occur. When the AI was inactive, responsibility ratings for the AI adviser and the human driver were comparable in pilot 7.

Methods

Materials

One vignette for an assisted driving scenario was adapted to match the two experimental conditions for the study. The vignettes presented brief accounts of the situation leading to questions about individual aspects of moral responsibility for the human driver and the AI assistant. The main vignette included a human driver who faces a junction in bad visibility while another car is approaching with priority from the right. The changes to the vignette included a variation in AI's modality (see methods for case descriptions and supplementary methods for detailed vignettes). Each participant read one vignette assigned at random – using counterbalanced block randomisation – and was asked to indicate his/her agreement with statements like 'The sensory AI assistant

The accident.' Responses were recorded on a 200-point scale using sliders (from -100 for 'Completely disagree' to 100 for 'Completely agree'). Comparing the responses across vignettes revealed the effect of the experimental manipulations.

Stimuli and Procedure

After a language comprehension test, participants were familiarised with the structure of the main experiment and the measurement scales. Then, participants completed a practice trial and continued with the main experiment. Here, they were presented with a text vignette and asked to rate the measured variables as accurately as possible. The vignette scenarios varied in modality within a 2x1 in-between-subject design. After completing an attention check, participants were asked to complete some basic demographic questions (age, gender, education), their familiarity with artificial intelligence, and their experience with computer programming.

Participants

I recruited a total of 20 participants from Amazon's MTurk service. No participants were excluded. 50 % of the participants were male and 50 % were female. 50 % of the participants had a bachelor's degree or higher. The mode age groups were 25 to 34 and 35 to 44 years old. The median age group was 35 to 44 years old.

Data Analysis

I analysed the data using general linear models (glm) from the lme4 library (Bates et al. 2015) in RStudio (Team 2021). Every model assumed a binomial distribution for the most accurate fit of the model to the data. In order to fit the model to the data, I normalised the data using min-max normalisation ((xi -- min(x)) / (max(x) -- min(x))). I used one glm for each of the main measurements (responsibility, blame, causal responsibility, and counterfactual capacity). These glm models were defined by glm(value ~ modality agent , family=binomial()). Each model evaluated the difference towards the measured variable based on changes of the experimental condition (linguistic vs sensory AI adviser) for each agent involved (human driver vs AI adviser) – given a negative outcome and an active AI adviser.

Results

Responsibility model

I fitted a logistic model (estimated using ML) to predict responsibility ratings with AI advice modality as the fixed effect (formula: value ~ modality * agent). The model's intercept, corresponding to modality = linguistic and agent = Driver, is at -0.03 (95 % CI [-1.31, 1.25], p = 0.962). Within this model:

The effect of modality [sensory] is statistically non-significant and positive (beta = 0.24, 95 % CI [-1.53, 2.05], p = 0.787; Std. beta = 0.24, 95 % CI [-1.53, 2.05])

The effect of agent [AI] is statistically non-significant and positive (beta = 1.73, 95 % CI [-0.24, 4.20], p = 0.109; Std. beta = 1.73, 95 % CI [-0.24, 4.20])

The effect of modality [sensory] × agent [AI] is statistically non-significant and negative (beta = -0.93, 95 % CI [-3.95, 1.87], p = 0.519; Std. beta = -0.93, 95 % CI [-3.95, 1.87])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Praise model

I fitted a logistic model (estimated using ML) to predict praise ratings with AI advice modality as the fixed effect (formula: value ~ modality * agent). The model's intercept, corresponding to modality = linguistic and agent = Driver, is at 0.15 (95 % CI [-1.11, 1.45], p = 0.813). Within this model:

The effect of modality [sensory] is statistically non-significant and positive (beta = 0.11, 95 % CI [-1.68, 1.90], p = 0.907; Std. beta = 0.11, 95 % CI [-1.68, 1.90])

The effect of agent [AI] is statistically non-significant and positive (beta = 2.00, 95 % CI [-0.13, 5.04], p = 0.099; Std. beta = 2.00, 95 % CI [-0.13, 5.04])

The effect of modality [sensory] × agent [AI] is statistically non-significant and negative (beta = -1.34, 95 % CI [-4.78, 1.56], p = 0.385; Std. beta = -1.34, 95 % CI [-4.78, 1.56])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Causal responsibility model

I fitted a logistic model (estimated using ML) to predict causal responsibility ratings with AI advice modality as the fixed effect (formula: value ~ modality * agent). The model's intercept, corresponding to modality = linguistic and agent = Driver, is at 0.34 (95 % CI [-0.91, 1.69], p = 0.597). Within this model:

The effect of modality [sensory] is statistically non-significant and positive (beta = 0.14, 95 % CI [-1.67, 1.98], p = 0.878; Std. beta = 0.14, 95 % CI [-1.67, 1.98])

The effect of agent [AI] is statistically non-significant and positive (beta = 1.30, 95 % CI [-0.69, 3.72], p = 0.224; Std. beta = 1.30, 95 % CI [-0.69, 3.72])

The effect of modality [sensory] × agent [AI] is statistically non-significant and negative (beta = -0.87, 95 % CI [-3.85, 1.92], p = 0.545; Std. beta = -0.87, 95 % CI [-3.85, 1.92])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Counterfactual capacity model

I fitted a logistic model (estimated using ML) to predict counterfactual ratings with AI advice modality as the fixed effect (formula: value ~ modality * agent). The model's intercept, corresponding to modality = linguistic & agent = Driver, is at 1.98 (95 % CI [0.40, 4.55], p = 0.041). Within this model:

The effect of modality [sensory] is statistically non-significant and negative (beta = -1.42, 95 % CI [-4.21, 0.72], p = 0.226; Std. beta = -1.42, 95 % CI [-4.21, 0.72])

The effect of agent [AI] is statistically significant and negative (beta = -2.31, 95 % CI [-5.10, -0.25], p = 0.047; Std. beta = -2.31, 95 % CI [-5.10, -0.25])

The effect of modality [sensory] × agent [AI] is statistically non-significant and positive (beta = 1.87, 95 % CI [-0.91, 5.10], p = 0.205; Std. beta = 1.87, 95 % CI [-0.91, 5.10])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95 % Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

Discussion

Pilot 8, given the small effect and sample size, has found no significant effects across measures variables. However, some notable trends emerged. The AI adviser was rated as more praiseworthy, responsible, causally, and counterfactually connected to the outcome than the human driver. This was surprising because the previous pilot data suggested the opposite effect. In previous pilots, the AI adviser was consistently attributed lower responsibility than the human driver when the outcome was negative. In other words, the outcome variation from negative to positive strongly influenced the allocation of responsibility across the AI adviser and the human driver. Furthermore, the overall rating pattern for the AI adviser and the human driver was consistent across all measured variables, including blame/praise, responsibility, causal responsibility, and counterfactual capacity. The consistent response pattern suggests that the AI adviser and the human driver are judged in terms of their moral responsibility, which grounds all forms of the measured variables. Moral responsibility encompasses a moral dimension captured by blame/praise judgements and a causal dimension captured by causal responsibility and counterfactual capacity judgments.

Limitations
One limitation of pilot 8 is the low sample size. Given the small effect size and the high variability of the measured variables, it was difficult to conclude any meaningful results.

Conclusion

Pilot 8 expanded the finding from pilot 7 by testing the same experimental setting in a positive rather than a negative outcome, i.e. where a car accident was avoided rather than caused. To simplify the experimental design, pilot 8 only tested conditions where the AI adviser was active, as pilot 7 demonstrated that the comparison between active and inactive AI adviser conditions led to the desired effect where responsibility ratings for either the AI adviser or the human driver differed satisfying as an adequate control condition.

Despite not finding any significant effects, pilot 8 revealed some notable trends. In contrast to the negative outcome scenarios before, participants attributed the AI adviser a higher responsibility, praise, causal responsibility, and counterfactual capacity than the human driver. In the context of the previous pilot 7, the AI adviser is praised but not blamed.

The following steps include a full-scale experiment by testing all of the experimental conditions explored in parts by pilots 7 and 8 to verify the observed trends with a sufficient sample size.

3.3 Main Experiment

3.3.1 Introduction

Who gets blamed when an accident happens? Is the AI system or the human relying on it? The nascent field of experimental AI ethics has found strong evidence that AI systems are judged as responsible as humans when they negotiate traffic decisions independently or with humans as co-actors (Awad et al. 2019; Franklin, Awad, and Lagnado 2021; Moglia et al. 2021; Nyholm and Smids 2016; Wischert-Zielke et al. 2020). Fully autonomous medical AI systems share responsibility with the clinician supervising them (McManus and Rutchick 2019; O'Sullivan et al. 2019). In medical and legal cases, AI is similarly held responsible when it provides social or moral guidance on whether a defendant can be released (Lima, Grgić-Hlača, and Cha 2021) or whether a risky medical procedure should be performed (Constantinescu et al. 2022). However, what happens when AI is merely an enhanced detection device, most closely resembling a mere instrument or tool? Would the mere instrumental use of AI leave the technology off the responsibility hook, or is the involvement of some form of intelligence sufficient to introduce attributions of responsibility?

An instrumental AI, in this case, provides only nudging recommendations or attracts attention to a piece of information. This is very different from an AI co-agent acting with or on behalf of the human user (Köbis, Bonnefon, and Rahwan 2021). The agential and moral roles of an autonomous AI co-agent can be distinguished from those of the human counterpart, but the influence of instrumental AI systems is harder to discern, even when such influence is relevant to the overall outcome (Kaur et al. 2020; Schaekermann et al. 2020), as it happens in low-stakes decisions such as shopping recommendations but also for high-stakes decisions such as medical diagnoses and driving support.

If strictly instrumental AI is assimilated to a mere tool and not as an independent agent (Cervantes et al. 2020; Longin 2020), it is not clear that it should be worthy of sharing the moral responsibility for the outcome of an action taken by a human user (Coeckelbergh 2020). The information provided by the AI system may be merely considered as having increased the knowledge or the awareness of the human agent (Fossa 2018). If the user has better information about a situation, they could even be considered more responsible for the outcome of their decision than someone with less information (Irlenbusch and Saxler 2019).

However, suppose the mere presence of AI induces the idea that an independent co-agent is involved or that the AI-powered tool could have done differently. In that case, I should expect that it will take a share of responsibility for the action carried out by its human user – though not necessarily a 50-50 split (Darley and Latane 1968; Kirchkamp and Strobel 2019; Kneer 2021; Stuart and Kneer 2021; Teigen and Brun 2011).

The two hypotheses make opposite predictions in a concrete scenario, for instance, when a driver relies on an AI-powered tool. If the instrumental role is what matters, then the driver should be held similarly responsible for their driving behaviour – if not more – when they use the AI-powered tool versus when not. If the involvement of an AI-technology is sufficient to prompt attributions of agency, then the human driver should be held less responsible when they use the AI. This diminished responsibility should entail that some share of responsibility goes to the AI system for contributing to the decision (Chockler and Halpern 2004; Halpern and Kleiman-Weiner 2018).

I conducted two preregistered vignette-based experiments using a between-subject experimental design to adjudicate between the above two hypotheses. I first established, in study 1, the conditions under which an AI-powered support system for driving would be held responsible along with the human driver. I examined (see Figure 1; n = 746) three main factors of (1) status (AI-system ON or OFF), (2) outcome (positive or negative), and (3) modality (verbal vs tactile AI instructions). At the most basic level, responsibility sharing would entail the AI system being held more responsible when it is ON (vs OFF). Suppose the AI is considered as a mere tool. In that case, I expect the

AI to be judged similarly (not) responsible when ON or OFF, while the human should be as responsible in both cases or even more responsible when assisted by a tool. If the AI is seen as a co-agent or capable of doing otherwise, there should be a sharing of responsibility and the human user should be held less responsible when the AI system is ON (El Zein, Bahrami, and Hertwig 2019; Forsyth, Zyzniewski, and Giammanco 2002; Strasser 2021; Williams 2013).

Previous work has shown that responsibility and outcome interact in complex ways (Baron and Hershey 1988; Roese and Vohs 2012; Kneer and Skoczeń 2021). When collaborating with others in risky gambles, people give themselves a bigger share of the credit for positive outcomes than the blame they take for the negative ones (El Zein, Dolan, and Bahrami 2022). When judging other humans, however, people attribute more blame than praise to them (Joshua Knobe 2003; J. Knobe 2003; Kominsky et al. 2015) but do the opposite for highly anthropomorphised artificial systems (Bartneck, Reichenbach, and Carpenter 2006). To examine the role of outcome in sharing responsibility, I included negative (i.e., the crash occurred) and positive (crash averted) outcomes in our vignettes. This outcome variation could also reveal the underlying psychological treatment of the presented AI system (R. A. Anderson, Crockett, and Pizarro 2020). While praise signals a willingness to cooperate given a good moral character (Gerstenberg et al. 2018), blame signals a willingness to inflict punishment and often co-occurs with perceived intentional wrongdoing (Cushman 2008, 2015; Lench et al. 2015).

The third factor examined the effect of the user interface on responsibility. Two kinds of AI systems were compared: a voice assistant delivering linguistic information and another AI assistant delivering only tactile feedback. I hypothesised that the more anthropomorphised voice assistants are more likely to appear like another agent (Chérif and Lemoine 2019). The AI using haptic feedback (e.g., the wheel's vibration) is, therefore, less likely to evoke responsibility attribution.

The results of this first experiment established that human participants attribute shared responsibility to the AI system even though, in debriefing, they predominantly described the AI system as a tool. In the follow-up, I conducted a critical control experiment showing that when the AI label was removed from the vignettes, the same scenarios did not evoke any responsibility sharing between the mechanical tool and the human agent in charge.

Comparing these conditions shows that even the most basic AI system introduces a sharing of responsibility with their human user in stark contrast to non-AI-powered tools. This finding is all the more surprising because, when asked, people did recognise AI as a tool. Attributing responsibility to AI and reducing human responsibility also does not depend on how the AI technology communicates with the user – i.e. via voice or haptic signals.

3.3.2 Methods

Experimental design

I conducted two online studies to elicit judgements on moral responsibility in human-assisted driving scenarios. Both studies used hypothetical vignettes describing a driving scenario with a human driver and an artificial assistant (see case description for details on experimental conditions and supplementary methods for studies 1 and 2). The artificial assistant was AI-powered in the main study (n = 746) and non-AI-powered in the follow-up study (n = 194). For the main study, I used a 2x2x2 between-subject experimental design. I varied three conditions with two factors each. This includes a variation in status (active vs inactive) and modality (linguistic vs sensory) of the AI assistant, as well as a variation in outcome (crash vs no crash). I controlled for any effects caused by the mere presence of an AI assistant by having the AI assistant present in all experimental conditions. The variation in the AI assistant's status enables the comparison between individual and AI-assisted decision-making cases. A follow-up study was conducted to control for any confounding effect from an assisting system's mere presence.

Main study 1

Study 1 compared the participants' ratings of responsibility, blame/praise, causality, and counterfactual capacity for the instrumental AI-assistant and human user across two experimental conditions (varying in status and modality of AI-assistant) and two experiments (varying in experimental outcome). The main study was conducted in two stages, exploring the manipulations of status and modality given a specific outcome. The first-stage experiment (n = 388) focused on manipulating status and modality in case of a negative outcome. In contrast, the second-stage experiment (n = 358) focused on manipulating status and modality in case of a positive outcome. I expected to see three main effects: an effect of the experimental outcome, an effect of the AI-assistant's status, and an effect of the AI-assistant's modality.

Follow-up study 2

Study 2 (n = 194) compared participants' ratings of responsibility, blame, causality, and counterfactual capacity for the non-AI-powered tool and human user across one experimental condition (varying in the status of the AI-assistant) in the case of a crash (negative outcome). I expected to see neither an effect of status for the tool nor the human user.

Materials

One vignette for an assisted driving scenario was adapted to match the eight experimental conditions for the main study and two experimental conditions for the follow-up study. The vignettes presented brief accounts of the situation leading to questions about individual aspects of moral responsibility for the human driver and the AI assistant. The main vignette included a human driver who faces a junction in bad visibility while another car is approaching with priority from the right. The changes to the vignette included a variation in outcome, a variation in AI's status, and a variation in AI's modality (see methods for case descriptions and supplementary methods for detailed vignettes). Each participant read one vignette assigned at random – using counterbalanced block randomisation – and was asked to indicate his/her agreement with statements like 'The sensory AI assistant deserves blame for the accident.' Responses were recorded on a 200-point scale using sliders (from -100 for 'Completely disagree,' to 100 for 'Completely agree'). Comparing the responses across vignettes revealed the effect of the experimental manipulations.

Data analysis

I analysed our data using general linear models (glm) from the lme4 library (Bates et al. 2015) in RStudio (Team 2021). Every model assumed a binomial distribution for the most accurate fit of the model to the data. In order to fit the model to the data, I normalised the data using min-max normalisation ((xi – min(x)) / (max(x) – min(x))). After I established that there was no modality effect on any of the measurements across conditions using a general alongside individual glms (see supplementary results), I used one glm for each of the main measurements (responsibility, blame/praise, causality, and counterfactual capacity). These glm models were defined by glm(responses_norm ~ status*outcome*agent, family=binomial()). I further confirmed that treating agent as an independent condition had no negative influence on the model's results (see supplementary results).

Participants

Main study - experimental stage 1

I preregistered and recruited a total of 440 participants from Amazon's Mechanical Turk service. After excluding 52 participants for failing preregistered data quality measures, I kept 388 participants for data analysis. 61 % of the participants were male, 37 % were female, and 2 % preferred not to say or stated other. 68 % of the participants had a bachelor's degree or higher. The mode age group was 24 to 34 years old. The median age group was 35 to 44 years old. 78 % were at least somewhat familiar with AI, while 73 % reported having little to no experience with computer programming.

Main study - experimental stage 2

I preregistered and recruited a total of 440 participants from Amazon's Mechanical Turk service. After excluding 82 participants for failing preregistered data quality measures, I kept 358 participants for data analysis. 55 % of the participants were male, 44 % were female, and 1 % preferred not to say. The mode age group was 24 to 34 years old, and the median age group was 35 to 44. 69 % of the participants had a bachelor's degree or higher. 77 % were at least somewhat familiar with AI, while 72 % reported having little to no experience with computer programming.

Follow-up study

I preregistered and recruited a total of 220 participants from Amazon's Mechanical Turk service. After excluding 26 participants for failing preregistered data quality measures, I kept 194 participants for data analysis. 54 % of the participants were male, and 46 % were female. 63 % of the participants had a bachelor's degree or higher. The mode age group was 24 to 34 years old, and the median age group was 35 to 44. 78 % were at least somewhat familiar with AI, while 70 % reported little to no computer programming experience.

Stimuli and procedures

Main Study

After a language comprehension test, participants were familiarised with the structure of the main experiment and the measurement scales. Then, participants completed a practice trial and continued with the main experiment. Here, they were first presented with a text vignette and then were asked to rate the measured variables as accurately as possible. The vignette scenarios varied in status and modality within a 2x2 in-between subject design. After completing an attention check, participants were asked to complete some basic demographic questions (age, gender, education), their familiarity with artificial intelligence, and their experience with computer programming.

Follow-up study

This experiment replaced the AI-powered with a non-AI-powered tool. Further, the experiment has only two, not four, conditions, varying only in status.

3.3.3 Results

I conducted two online studies to elicit judgements on moral responsibility in human-assisted driving scenarios. Both studies used hypothetical vignettes that describe a driving scenario with a human driver and an artificial assistant (see case description for details on experimental conditions and supplementary methods for vignettes of studies 1 and 2). The artificial assistant was AI-powered in the main study (n = 746) and non-AI-powered in the follow-up study (n = 194). For both studies, I used the same set of vignettes and between-subject design with slight modifications to accommodate the changes in outcome, status, modality, and the type of assistant.

Main study

Study 1 compared the participants' ratings of responsibility, blame/praise, causality, and counterfactual capacity for the instrumental AI-assistant and human user across two experimental conditions (varying in status and modality of AI-assistant) and two experiments (varying in experimental outcome). The main study was conducted in two stages, exploring the manipulations of status and modality given a specific outcome.

The first-stage experiment (n = 388) focused on manipulating status and modality in case of a negative outcome. In contrast, the second-stage experiment (n = 358) focused on manipulating status and modality in case of a positive outcome. I expected to see three main effects: an effect of the experimental outcome, an effect of the AI-assistant's status, and an effect of the AI-assistant's modality.

AI advice modality does not affect responsibility ratings

I found no effect of the AI assistant's modality. Different participants rated the AI assistant and the human user as responsible when the AI assistant provided sensory compared to linguistic advice. The sensory AI assistant used tactile steering wheel vibration for driving assistance, and the linguistic AI assistant issued verbal instructions. Using a general linear regression model (glm), I found no general effect of the AI-assistant's modality across experimental conditions (beta = -0.00296, 95 % CI [-0.10, 0.09], p = 0.952); for details of the model and pairwise comparisons of experimental conditions see supplementary results). To improve the explanatory power of the subsequent regression models, I decided to collapse the modal difference between AI assistants and treat them as a generic AI assistant for subsequent analyses. To analyse our remaining results, I used one glm for each of the primary measurements: responsibility, blame/praise, causality, and counterfactual capacity (see methods for details).

AI's status strongly affects responsibility ratings for the human driver and the AI assistant

I found that the AI assistant's status had a strong impact on responsibility ratings. When the AI assistant was active and a crash occurred, participants rated the responsibility of the human driver lower (beta = -0.14, 95 % CI [-0.22, -0.05], p = 0.018) and the responsibility of the AI assistant higher (beta = 0.24, 95 % CI [0.16, 0.32], p < 0.001) as their inactive AI-assistant baseline. When no crash occurred, the same pattern emerged. The participants rated the responsibility of the human driver lower (beta = -0.21, 95 % CI [-0.29, -0.13], p < 0.001) and the responsibility of the AI-assistant higher (beta = 0.69, 95 % CI [0.61, 0.76], p < 0.001) compared to the inactive AI-assistant baseline.

The human driver and the AI assistant are rated differently across outcomes

I found that the human driver and the instrumental AI assistant were rated differently across conditions. When the AI assistant was inactive and a crash occurred, the AI assistant was seen as significantly less responsible than the human driver (beta = -0.71, 95 % CI [-0.78, -0.65], p < 0.001). The effect persisted when no crash occurred (beta = -0.83, 95 % CI [-0.89, -0.77], p < 0.001). When the AI assistant is active, on the other hand, a new pattern emerges. While the AI assistant was also seen as significantly less responsible than the human driver when a crash occurred (beta = -0.34, 95 % CI [-0.44, -0.25], p < 0.001), both are seen as equally responsible when no crash occurred (beta = 0.07, 95 % CI [-0.02, 0.16], p = 0.116).

Responsibility ratings are strongly outcome-dependent

I found a strong outcome effect for the AI assistant. When the AI assistant was inactive, I discovered that the AI assistant was seen just as responsible when the outcome was negative rather than positive (beta = 0.02, 95 % CI [-0.04, 0.08], p = 0.476). In addition, the human driver was seen as slightly less responsible when the outcome was negative rather than positive (beta = -0.09, 95 % CI [-0.16, -0.02], p = 0.0097). However, when the AI assistant was active, I found that the AI assistant was seen as much more responsible for the positive than negative outcome (beta = 0.43, 95 % CI [0.34, 0.52], p < 0.001). This was not the case for the human driver, who was seen as responsible for the positive than the negative outcome (beta = -0.013, 95 % CI [-0.11, 0.08], p = 0.775).

AI-assistant is strongly perceived as a tool

I also tested the perception of the AI assistant as a tool. I found that participants viewed the AI assistant as a tool consistent across experimental conditions. Fitting an additional glm, I found neither an effect of status (beta = -0.12, 95 % CI [-0.77, 0.52], p = 0.71) nor an effect of outcome (beta = 0.04, 95 % CI [-0.64, 0.72], p = 0.9) for the tool ratings of the AI-assistant.

Follow-up study

Study 2 (n = 194) compared participants' ratings of responsibility, blame, causality, and counterfactual capacity for the non-AI-powered tool and human user across one experimental condition (varying in the status of the AI assistant) in case of a crash (negative outcome). I expected to see neither an effect of status for the tool nor the human user.

Tool status does not affect responsibility ratings

I found no status effect. Participants rated the human driver (beta = -0.08, 95 % CI [-0.2, 0.03], p = 0.156) and the non-AI-powered tool (beta = -0.07, 95 % CI [-0.19, 0.05], p = 0.245) as responsible for a crash when the tool was active rather than inactive.

Human driver and tool are rated differently across outcomes

I found that the human driver and the non-AI-powered tool were rated differently across conditions. In fact, the non-AI-powered tool was seen as significantly less responsible than the human user when the non-AI-powered tool was active (beta = -0.55, 95 % CI [-0.67, -0.43], p < 0.001) and when it was inactive (beta = -0.56, 95 % CI [-0.68, -0.45], p < 0.001).

Tool is strongly perceived as a tool

I also tested the perception of the non-AI-powered tool as a tool. I found that participants viewed the non-AI-powered tool as a tool consistent across experimental conditions. Fitting an additional glm, I found no effect of status (beta = -0.01, 95 % CI [-0.09, 0.07], p = 0.79) for the tool ratings of the AI assistant.

3.3.4 Discussion

The central finding is a strong dissonance between the participant's behaviour and beliefs toward instrumental AI assistants. On the one side, participants attributed responsibility to the AI assistant as demonstrated by AI and human co-agents. However, on the other side, participants strongly believed that the AI assistant was a tool – traditionally dissociated with being held responsible.

On the behavioural side, I have shown that the presence of an active AI assistant strongly influences responsibility, blame, praise, and causality ratings for the human user of the AI system. The human user was seen as less responsible for an outcome when the AI assistant was active rather than inactive. Analogously the AI assistant was seen as more responsible when active. The same pattern of significance holds for blame, praise, and causal influence ratings suggesting a robust sharing effect of moral and causal responsibility.

In addition, I found that the perceived responsibility of the AI assistant was highly outcome dependent. In fact, the AI assistant was seen as much more responsible for the positive than the negative outcome condition. The AI assistant was praised more for avoiding an accident than it was blamed for causing it. This finding is contrasted with the human user, who showed no outcome effect. The human user was rated as responsible in the positive and negative outcome conditions. These findings align with previous work on AI co-agents such as fully autonomous cars (Awad et al. 2019; Franklin, Awad, and Lagnado 2021) and collective human decisions (El Zein, Bahrami, and Hertwig 2019). Both pieces of literature demonstrate responsibility sharing in human-AI or human-human settings. This supports our original hypothesis that instrumental AI assistants are perceived as agents capable of sharing responsibility with other agents (Darley and Latane 1968; Kirchkamp and Strobel 2019; Kneer 2021; Stuart and Kneer 2021; Teigen and Brun 2011). The sharing of responsibility and the outcome dependence are indicators of an agent-like perception of AI assistants. Both patterns have been demonstrated to hold for human agents (Baron and Hershey 1988; El Zein, Bahrami, and Hertwig 2019).

However, surprisingly, participants' beliefs about AI assistants contradicted their behaviour. Consistently across experimental variation, participants rated the AI assistant as a tool – which goes against it being seen as an agent and sharing responsibility with its human user (Fossa 2018). Replacing the AI-powered with a non-AI-powered tool in the follow-up study revealed that sharing responsibility only occurs when AI is involved, even though the resulting role and information are similar.

The way the AI's advice is presented to the human user – either through tactile or linguistic advice – did not influence responsibility assessments, contrary to what could have been expected both from human-human interactions and the influence of anthropomorphic features in responsibility attributions to AI (Chérif and Lemoine

2019; Dalal and Bonaccio 2010; El Zein, Bahrami, and Hertwig 2019; Steffel, Williams, and Perrmann-Graham 2016; Strasser 2021).

The tension between general beliefs and responses in this study echoes other conflicts between the animate and inanimate characteristics attributed to AI (Kahn et al. 2012). 4-year-olds rarely attribute biological properties or aliveness to a robot, yet still affirm it has perceptual and psychological capabilities, such as having cognition and emotions (Jipson and Gelman 2007; Nigam and Klahr 2000). People consider humans and AI as cooperative partners, yet feel guilty when they exploit humans but not when they exploit AI agents (Bartneck, Reichenbach, and Carpenter 2006; Karpus et al. 2021).

One aspect from the human advice-giving literature seems, however, to apply to AI, i.e. hindsight and other-serving biases (Palmeira, Spassova, and Keh 2015). The hindsight bias, i.e. the "I-knew-it-all-along" effect, describes people's general tendency to view past events as predictable (Christensen-Szalanski and Willham 1991). The other-serving bias captures people's tendency to see an advisor as more responsible for positive than negative outcomes. Driven by hindsight bias, participants tend to believe that the advisor was more in control of the positive than the negative outcome. If this is the case, participants' attention may turn towards the human driver when the outcome is negative and decrease the AI assistant's share of responsibility in the event.

Limitations

Beyond the other-serving bias, I acknowledge that other factors could be in play with similar explanatory power for the asymmetric evaluation of the AI assistant. For instance, alternative explanations for the asymmetric AI-assistants assessment include a lacking attribution of intentionality for AI-assistants, which has been suggested as at least a co-factor for praising but not blaming behaviour (Guglielmo and Malle 2019; Hindriks, Douven, and Singmann 2016; Joshua Knobe 2006; Malle, Guglielmo, and Monroe 2014). To further increase the robustness of our findings, it would be beneficial to replicate our findings in other high-stakes domains, such as healthcare and low-stakes domains, such as everyday traffic navigation (Lai et al. 2021). Similarly, it would be further essential to test for any cultural variations, as cultural norms can strongly impact how AI is perceived and held responsible (Awad et al. 2018; Fast and Horvitz 2017; J. H. Kim et al. 2022).

Another question lies in exploring the contrast between instrumental and moral AI assistants. Suppose the mere involvement of AI suffices for holding the most basic instrumental assistant responsible. In that case, moral AI assistants may be held responsible due to their explicit moral involvement, but also due to the mere presence of an AI. Separating these two remains a crucial question for future work.

Finally, the description of the AI technology as an "assistant" may play a role in people's attribution of responsibility, and variations in the presentation of the AI system could be varied. I note, however, that people here considered the AI-system toollike and did not treat a voice assistant as more agentive or responsible than a haptic technology, suggesting that the term "assistant" and its possible human or agentive connotations were not the reason behind their responses.

3.4 Conclusion

The main question which has motivated the project was to understand the perceived agentive role AI advisers play in human-AI interaction. Are AI advisers perceived more as complex tools or agents proper? Conceptually there is no obvious answer. Philosophically, the human-like agency requires a wide range of cognitive and arguably moral processes, which AI systems – notably AI advisers — lack. To explore this question, I turned towards a reliable and measurable proxy for perceived levels of agency: the attribution of moral responsibility.

Notably, moral responsibility is only attributed to agents, and a hammer is hardly morally responsible for hitting your finger; the hammer user is. Similarly, two bank robbers share the responsibility for the bank robbery as both agentively performed the bank robbery. Understanding whether an AI adviser shares responsibility with its human user means understanding the perceived agentive role an AI adviser has.

The literature on responsibility attribution in advisory human-AI scenarios is ambivalent. Some argue that AI advisers share responsibility with the human user (Constantinescu et al. 2022; Malle, Magar, and Scheutz 2019), whereas others show the opposite (Coeckelbergh 2020).

While previous literature has focused on autonomous and interactive AI systems, I established how responsibility is attributed to more common instrumental AI systems. When AI is involved in an agentive role in bringing about an outcome, the attribution or sharing of responsibility is quite natural. In this chapter, I addressed a more fundamental question: whether sharing responsibility with humans could come from the mere involvement of another – artificial – intelligence. This work contributes not only to the growing literature on AI assistants but also provides critical insights into the asymmetric evaluation of AI assistants, which are praised more than blamed.

The experimental work of this chapter sought to develop a new experimental setting in which the question of whether an AI adviser is perceived more as a complex tool or as an agent proper can be addressed. At the start of the experimental project, experimental hypotheses were still vague but developed over time. Overall, based on the vast literature on responsibility attribution in human-human scenarios, I expected an outcome bias to occur, where actions leading to a positive outcome were judged more favourably than those leading to a negative outcome (Baron and Hershey 1988). In addition, I expected a modality effect to occur — such that the way the AI advice was delivered to the human user influences the perceived responsibility of the user. The sensory AI adviser is expected to be perceived as less intrusive than the linguistic AI adviser, thereby retaining the human user's agentive autonomy to its fullest. In other words, the modality difference between AI advisers tests whether any observed responsibility difference can be explained due to a difference in advisers (more tool-like vs more agent/partner-like) or due to the presence of an AI itself. If no difference in AI advice modalities was observed, then the responsibility difference must stem from the implementation of the AI itself. If a difference in AI advice modalities was observed, on the other hand, then the responsibility difference should be reducible to a difference in advice modality.

The first two pilots mapped the unchartered ground with a mixed experimental design to balance experimental robustness and data quality. A medical scenario was adapted to fit six experimental conditions varying in the advisory setting (no AI, sensory AI, linguistic AI) and the outcome (positive, negative). The pilots examined the human user's responsibility (pilot 1) and blame/praise (pilot 2). They found an overall outcome trend where the human user was seen as more responsible and blame-/praiseworthy for the positive than the negative outcome — replicating the expected outcome effect from the literature. In addition, the pilots found an overall AI-presence trend where the human user was seen as more responsible and blame-/praiseworthy when acting alone rather than with either AI adviser. Otherwise, the modality of the AI adviser did not affect the human user's responsibility or blame/praise ratings.

To improve experimental robustness, pilots 3 and 4 tested the advisory human-AI setting in a within-subject design. Pilot 3 used six diverse scenarios, while Pilot 4 used a common car-driving scenario with varying background stories to explore the blame/praise, causal responsibility of the human user and perceived informativity of the AI adviser. The extension of measured variables from only asking for responsibility was motivated by the possible confound of responsibility ratings – as responsibility can either refer to the mere causal connection of an agent to an outcome (causal responsibility) or also include a moral judgment of the agent (moral judgment). Asking for a moral judgment (blame/praise) and a causal judgment (causal responsibility) allows us to pinpoint where the possible difference between the human user and the AI adviser emerges. Both pilots validated the previously observed trends. The human driver was praised more than blamed – with or without an AI adviser, irrespective of the AI adviser modality. Pilot 4 further demonstrated that the human driver was praised more when acting alone rather than with advice.

To increase consistency between no-AI vs AI conditions, pilots 5 and 6 presented a large-scale sampling of a new experimental paradigm – a full between-subject design. Within the new experimental setting, pilots 5 and 6 tested whether and how blame and causal responsibility judgments were distributed across the involved agents. They, therefore, expanded the measured variables to include blame/praise and causal responsibility judgments for the human user, the AI adviser and an uninvolved third party. The uninvolved third party was introduced as a control and, as expected, received only a small fraction of responsibility consistently across conditions. Otherwise, the pilots revealed a strong agent effect: the human driver is across conditions seen consistently as most responsible for the outcome, followed by the AI adviser – irrespective of this modality – and lastly, the third party. However, a confounding limitation became clear: the observed responsibility ratings when the AI adviser was present were also caused by the AI adviser's mere presence instead of the given AI advice itself.

Pilots 7 and 8 eliminate a possible confounding limitation from pilots 5 and 6 by adapting the experimental design from a 3x2 to a 2x2x2. The new design contained a variation in outcome (positive, negative), AI advice modality (sensory, linguistic), and AI status (active, inactive). The new experimental design included the presence of the AI adviser across all conditions. However, it modulated the status of the AI adviser as either active, advice-giving, or inactive, not-advice-giving. The pilots found that, in case of a negative outcome, the human user is held more responsible and blameworthy than the AI adviser across conditions except when the sensory AI adviser is turned on. In case of a negative outcome, however, surprisingly, the AI adviser overall is held more responsible, praiseworthy, and causally responsible than the human driver.

The final experiment used the same experimental design as established in pilots 7 and 8 and replicated their main findings. The final experiment showed that 1) the human user shares responsibility with an AI advisor, 2) the AI advisor is not blamed but praised for an accident, and 3) how the AI advisor makes recommendations has no bearing on any responsibility rating.

Overall, taking the experimental results together and putting them in context, advisory AI systems in parts share responsibility with their human user, implying that they are more than tools. Even when AI systems were presented as most tool-like, they were still seen as mainly responsible for a positive outcome. However, AI advisers are also not human-like – as they are not nearly blamed as much as their human users. The experimental findings of this chapter validate the conceptual findings from Chapter 2 and provide a lower bound for what AI advisers are. By demonstrating unique responsibility patterns, AI advisers show that they are perceived as having some non-tool-like agentive capacity to bring about the outcome.

Both chapters in the first part of this thesis have examined the loose coupling of AI advisers with their human users – cases where AI advisers provide seemingly external recommendations. Both chapters lay out the conceptual and empirical reasons to consider AI advisory systems that are external to the human user within their unique ontological status. However, what happens if AI advisors do not provide external advice and are tightly integrated with their human users? The upcoming second part analyses a tighter coupling of AI advisers with their human users – cases where AI advisers become integral to human perception and decision-making. Consider cases of augmented reality or sensory augmentation. Here, I ask how and to which extent AI advisers influence human perception and in which way highly integrated AI systems differ from their tool or human counterparts.

Part 2

4 Augmented Perception

4.1 Introduction

Extending, or even augmenting, the senses has long been a human dream. This dream may now become a reality thanks to recent advances in sensory augmentation powered by artificial intelligence (AI). Artificial noses can identify thousands of odours (D. Hu et al. 2019) and distinguish between infected and non-infected wounds (Haalboom, Gerritsen, and van der Palen 2019); driverless cars detect secluded objects with laser radar and infrared cameras (Pulikkaseril and Lam 2019); and robots can use photosensors to recognise materials based on their sounds (Eppe et al. 2018). So, what happens if humans are outfitted with artificial sensors? Can these sensors be linked to humans in such a way that they expand our perception beyond the use of external tools?

These are fundamentally philosophical questions: many human-applied AI augmentation systems are in their infancy but gaining momentum. While some (Fernández-Caramés and Fraga-Lamas 2018; M. Chen et al. 2016) have created internet-connected textiles, others (Raisamo et al. 2019; McGreal 2018; T. D. Wright and Ward 2018) want to see AI-driven augmentation devices integrated more broadly into human perception. The extent to which AI-powered augmentation devices alter and potentially extend human sensory capacities to form a new type of hybrid, trans-human perception is unknown. On the one hand, AI-driven systems are relatively autonomous computational systems that can process and forward sensory signals to the human user, similar to how a self-driving car forwards identified road obstacles to the human driver. On the other hand, AI-powered systems have the potential to become deeply integrated/entrenched in human perception and cognition. Hearing aids and noise-cancelling headphones, for example, use AI to amplify or filter sounds, compensating for hearing loss and improving daily hearing. Other sensory tools, such as augmented reality systems, such as Google Glass, or medical technologies, such as magnetic resonance imaging (MRI) and computed tomography (CT) scanners, have gradually used AI to reduce perceptual or cognitive friction between the tool and the user. Google Glass displays real-time, task-relevant information; medical technologies provide diagnostic results next to the taken image.

The concept of improving human perception through wearable devices is not new: it can be traced back to popular culture (RoboCop, Ghost in the Shell, Inspector Gadget) and is now being realised through prosthetics, sensory substitution, and extension devices. These technologies are examples of sensory augmentation because they provide additional sensory cues to convey relevant information for a perceptual task. Depending on the augmentation target, human augmentation can be classified into three types: sensation, cognition, and action (Raisamo et al. 2019). Action augmentation enhances physical abilities beyond humans' natural motor and sensory limits. This includes prosthetic limbs and exoskeletons, allowing users to regain control of paralysed limbs or remotely control robots via virtual reality (VR). The field of action augmentation has focused on exoskeletons that allow people to walk on robotic feet (Dollar and Herr 2008; Young and Ferris 2017) or remote-controlled robots like medical operating systems mimic user movement (Panesar et al. (2020); see Moglia et al. (2021) for review) are examples of prosthetic limbs that can reinstate or enhance movement.

Sensory augmentation uses recorded sensory signals to augment or extend the human user's natural senses. Augmented vision, hearing, haptic sensations, smell, and taste are all examples. The field of sensory augmentation has used advanced sensors and signal computations to compensate for sensory impairments or enhance existing senses. Sensory substitution devices have shown remarkable progress in compensating for sensory impairments. Sensory substitution describes transferring sensory signals from one sensory modality to another. One common and successful application has been transferring visual information to sound to compensate for vision impairment, like 'the vOICe' (Auvray, Hanneton, and O'Regan 2007; Meijer 1992; Proulx et al. 2008). Other applications include vision-to-tactile sensory substitution devices like TVSS (Arnold and Auvray 2018; Bach-Y-Rita et al. 1969) and vestibular-to-tactile sensory substitution devices (Tyler, Danilov, and Bach-Y-Rita 2003). Findings on neural plasticity, for instance, have demonstrated that sensory substitution devices can at least partially restore a lost sense through neural re-organisation and practice (Amedi et al. 2007; Bach-y-Rita and W. Kercel 2003; L. G. Cohen et al. 1997; Collignon et al. 2008).

Cognitive augmentation detects and interprets human cognitive states to match and predict the human user's expectations and extend the user's cognitive abilities. This includes extended memory devices, which generate an accurate and coherent environmental response to match the user's expectations and needs. The field of cognitive augmentation has used technology to enhance or augment human cognitive functions, such as memory, attention, and problem-solving. Examples include wearable devices that aid learning or decision-making (Dingler et al. 2016; Li and Ji 2005; Palmer and Kobus 2007) or raise bodily awareness by analysing breathing, heart rate, and body temperature patterns to detect stress, anxiety, or potential medical emergencies (Reeder et al. 2017). Brain-computer interfaces, in addition to smart wearables, have been used to observe and influence brain activity to communicate memory, attention, situational awareness, and complex problem-solving (see Cinel, Valeriani, and Poli (2019) for a review).

With increasingly capable sensors and the computational capacity to exploit emergent sensory information, it is becoming increasingly important to untangle the blurred boundary between AI-driven augmentation devices and their human users. There are pressing ethical and legal implications and philosophical questions about what constitutes the human perceptual system. If AI-powered augmentation devices extend human perception, questions of responsibility may become impossible to answer: who or what is to blame for any negative or positive outcome? Which comes first, the augmentation device or the human user? Much, therefore, hinges on whether AI-powered augmentation devices count as external sensory tools – on which the human user relies – or whether they become part of the human perceptual system – thereby extending human perception.

Smart wearables exemplify the development of sensory tools. A smart wearable is a device worn on the body and capable of performing tasks beyond simple monitoring or tracking. Artificial sensors and processors can even outperform human sensory capacities: artificial noses can identify thousands of odours (D. Hu et al. 2019) and distinguish between infected and non-infected wounds (Haalboom, Gerritsen, and van der Palen 2019); driverless cars detect secluded objects with laser radar and infrared cameras (Pulikkaseril and Lam 2019); and robots can use photosensors to recognise materials based on their sounds (Eppe et al. 2018). What makes a smart wearable "smart" is its ability to gather and process information and interact with other devices and systems. Smart wearables are often equipped with sensors, processors, and other technology that allow them to interact with the environment and perform various functions, such as receiving and sending messages, tracking physical activity, and monitoring health. Some examples of smart wearables include smartwatches, fitness trackers, smart glasses, and smart clothing. These devices can be used for various purposes, including health and fitness tracking, communication, entertainment, and more. At their peak, sensory tools can become an integral part of human perception and at its peak are even barely noticed. They may even become transparent to the human user, i.e. integrated into perceptual processes (Auvray, Hanneton, and O'Regan 2007) – like a hammer to a blacksmith.

On the other hand, AI systems might be more than sensory tools and – through their sensory computation – may provide their human user with perceptual content or experiences. Extending the senses beyond sensory tools may sound odd initially, but a biological precedent for extended perceptual systems exists. Certain organisms use their sensory fields, generated beyond their bodily boundaries, to achieve specific goals such as mating or prey detection. Bats, electric fish and spiders are three common examples. Because their self-generated sensory fields – a soundscape for bats, an electric field for fish, and a web for spiders extend beyond their bodies, defining the boundary of their sensory systems is difficult. Examining how these creatures interact with their surroundings can provide insight into the mechanisms and computational processes involved in perception and call into question the notion that perception is solely an internal process. Similarly, AI-powered augmentation devices also plausibly extend the human sensory field. Sensory substitution devices like the feelSpace belt (Nagel et al. 2005), for example, can give the human user a sense of a magnetic field translating magnetic information into felt tactile vibrations (see also Hameed et al. (2010) and Kärcher et al. (2012)).

Perceptual experiences include both bottom-up and top-down processing (Varga 2017). The former type of processing includes a) external energy stimulation of sensory receptors and b) sensory information transmission into the central nervous system. The former type of processing entails interpreting what we perceive based on prior knowledge or context. While low-level theorists believe that perceptual experience is reducible to low-level property experiences, high-level theorists argue that perceptual experiences are primarily derived from high-level property experiences. According to high-level theorists, cognitive states can penetrate perceptual experience and provide an interpretation of cognitive penetration that provides some support for the high-level view. While sensory tools can only provide the human user with additional low-level signals, AI-powered augmentation devices may also externalise high-level processing. AI systems can manipulate signals based on prior knowledge and context through advanced computational processing. The human user receives a sensory signal that has already been interpreted. Because AI-powered augmentation devices can manipulate the signal to such an extent, they present a novel and challenging case for extending perceptual experiences.

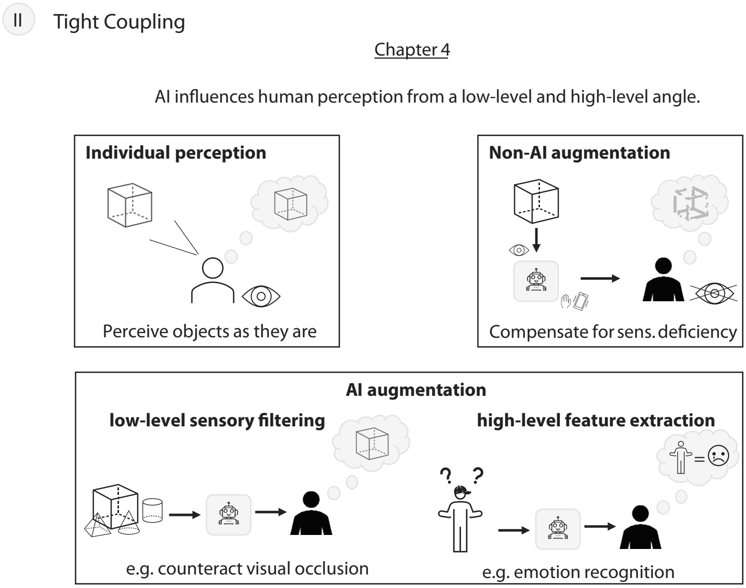

II
Tight Coupling
Chapter 4
AI influences human perception from a low-level and high-level angle.
Individual perception
Perceive objects as they are
Non-AI augmentation
Compensate for sens. deficiency
AI augmentation
low-level sensory filtering
e.g. counteract visual occlusion
high-level feature extraction
? ?
=
e.g. emotion recognition

Figure 12: Overview chapter 4

By focusing on sensory augmentation, this chapter investigates the extent to which AI systems influence human perception. The first section will provide a systematic taxonomy of sensory augmentation and outline the changes AI brings to sensory augmentation. The second section expands on our understanding of AI-powered sensory augmentation and investigates whether and to what extent AI extends human perception.

Given the breadth of the literature on human augmentation, this chapter will concentrate on non-invasive sensory augmentation devices. While prosthetics rely on invasive methods such as cochlear, vestibular, or corneal implants (Golub et al. 2014; Zeng et al. 2008), sensory substitution and extension devices rely on non-invasive methods such as external wearable devices that can be put on and taken off (Zeng et al. 2008). These external devices have extensively used advanced artificial sensors and AI-based algorithmic computation. Whether sensory augmentation devices extend human perception is much more pressing in the case of external devices than internal devices. External devices collect, process, and forward sensory signals to the human user as another sensory signal. In contrast, internal devices transmit the recorded signal to the human user as an internal neural signal rather than an external neural signal.

4.2 How AI systems augment the senses

4.2.1 Defining Sensory Augmentation

The idea of improving human perception through wearable devices is not new. Indeed, it can be traced back to popular culture, from characters like RoboCop and Ghost in the Shell to the classic Inspector Gadget. However, nowadays, this concept is being realised through prosthetics, sensory substitution, and extension devices. Under the umbrella term of 'sensory augmentation,' these technologies provide users with additional sensory cues to assist them in completing various perceptual tasks. By taking advantage of these tools, individuals can gain a heightened level of awareness, allowing them to better understand their environment.

A sensory augmentation device comprises three main components: an artificial sensor, a coupling system, and a stimulator (Elli, Benetti, and Collignon 2014; T. D. Wright and Ward 2018). The artificial sensor receives incoming sensory information, the stimulator generates a sensory signal, and the coupling system connects the two. The artificial sensor records sensory information in the substituted sensory modality, and the stimulator outputs sensory information in the substituting modality, according to sensory substitution terminology. In terms of technology, the artificial sensor and stimulator are implemented in hardware, while the coupling system connects both pieces of hardware with software.

As the coupling system, this sensory substitution and augmentation process more broadly entails implementing a conversion algorithm. The algorithm takes sensory input in one sensory modality, transforms it into another sensory signal, and outputs it in the desired sensory modality. Implementing the conversion algorithm creates a cross-modal, non-physical link between artificial sensors and stimulators. After extensive training, the human user must learn how to interpret the output signal, which is the case for sensory substitution devices.

Conceptually, as argued by Longin and Deroy (2022), sensory augmentation requires four elements:

1. that the **input** of a sensory augmentation device is a **sensible** property, set of properties or object,
2. that the **output** of a **sensory** augmentation device is causally related to the input and delivered as additional information to the user in a sensory format,
3. with the **goal** to provide or improve perceptual functions.
4. that the sensory augmentation device either forwards low-level sensory information through a linear relation between incoming and outgoing signals (for classical sensory augmentation) or extracts higher-level features through a non-linear relation between incoming and outgoing signals (for intelligent sensory augmentation).

The first input requirement limits the inputs that result in sensory augmentation, but this should be qualified further. One critical question is whether virtual reality (VR) glasses should be considered 'augmenting perception,' as they also display additional information in a sensory format. It is important to note that VR objects are not generated from sensible properties or objects. VR glasses, for example, generate their displayed objects rather than gathering sensible properties from the user's internal or external environment. As a result, in this case, VR glasses do not provide sensory augmentation. As a result, when considering the implications of VR glasses and other forms of sensory augmentation, it is critical to recognise this distinction.

Assume the first requirement concerns VR cases, and the second is about distinguishing sensory augmentation from cognitive augmentation and sensory tools. Cognitive augmentation devices add symbolic or linguistic information to the perceiver's environment, allowing them to better understand their surroundings and mental processes. Extended-memory devices (Lee et al. 2016; Smart 2017), which can store and recall large amounts of information, and personal assistants (Canbek and Mutlu 2016; Hoy 2018), which can provide personalised advice, are examples of cognitive augmentation devices. Furthermore, many digital smart wearables, such as smartwatches (Fernández-Caramés and Fraga-Lamas 2018; Sun, Liu, and Zhang 2017), provide numerical values based on sensory inputs such as heart rate, whereas digital personal assistants or car-based navigation systems produce verbal outputs that aid cognitive tasks such as navigation. Text-to-speech devices, which take in sensory signals but produce linguistic output, can also be classified as cognitive augmentation devices; however, because their output is not strictly sensory, they cannot be classified as sensory augmentation.

Sensory tools such as ordinary glasses or a cane, on the other hand, transfer information in a sensory format but do not augment sensory capabilities (T. D. Wright and Ward 2018). The long cane is a typical example of this, as it allows the blind to detect obstacles by extending their tactile field. This tactile information transfer is dependent on the user's sensory abilities. Even though these tools mediate sensory information, they are not sensory augmentation because they do not improve the user's sensory capabilities beyond their natural state. Instead, they enable users to interact with their surroundings more effectively. As a result, these tools are limited in their ability to truly augment the user's sensory experience.

According to the third requirement, sensory augmentation devices must provide or improve a perceptual function. A perceptual function detects, locates, discriminates, or identifies properties and objects in its surroundings. These devices' output must be linked to some environmental property, including the inner environment, i.e. the body. This excludes smart textiles, where the added sensory output is a source of additional sensations, such as vibration motors in jackets, which can provide new tactile sensations on the skin. However, these sensations are not typically constructed as the perception of objects or serve a specific perceptual function (for a review and discussion, see Tajadura-Jiménez, Väljamäe, and Kuusk (2020)). Positive examples of sensory augmentation devices include vOICe, which records and converts visual information into auditory frequencies. This improves spatial awareness perceptual function by linking changes in the visual field to changes in the produced auditory frequencies, allowing for a better understanding of the environment. Sensory augmentation devices can help people perceive their surroundings more meaningfully and accurately.

The fourth requirement is concerned with the processing of recorded sensory signals. A basic sensory pattern from the environment, such as light reflections or sound frequencies, is considered low-level sensory information. High-level sensory information is a more refined sensory pattern that captures specific sensory characteristics.

The primary distinction between traditional and intelligent sensory augmentation devices is how the gathered information is translated into the output format. Non-linear mapping, which is only possible with intelligent sensory augmentation devices, foregoes preserving the original data's structure to transform the data non-proportionally. With the increasing integration of AI into sensory augmentation, the relationship between the input and output signal becomes a criterion with multiple values, allowing two types of augmentation to be distinguished: intelligent and non-intelligent. AI-powered sensory augmentation devices significantly improve sensory substitution and augmentation by pre-computing and only forwarding minimally noisy sensory signals to the human perceiver, which convey rich environmental cues.

4.2.2 AI in Sensory Augmentation

AI can be implemented in sensory augmentation devices in two ways: before and after the output is presented to the user. When AI is introduced before the output, it alters how sensory information is translated within or across sensory modalities. If AI is introduced after the output is presented to the user, AI facilitates the user's encoding process.

Intelligent sensory augmentation devices that implicitly use AI after the output have already been developed. In this case, AI methods are used to assess the quality of the final sensory signal and provide recommendations for improving signal quality. M. Kim et al. (2021), for example, use a cross-modal generative adversarial network-based evaluation method to find optimal auditory sensitivity in visual-auditory sensory substitution to reduce transmission latency. W. Hu et al. (2019) evaluate different encoding schemes for a visual-to-auditory sensory substitution device based on the user's needs using machine learning. Given the differences in previous exposure to visual stimulation between the late-blind and the congenitally blind, different encoding schemes are required to facilitate the recognition of 'visual' objects through sound. Late-blind users, unlike congenitally blind users, can use pre-existing visual experiences as a valuable reference for any cross-modal perception. While these modern technologies have demonstrated how AI can improve traditional sensory augmentation schemes after the output is presented to the user, there is still room for significant improvement by implementing AI before the output.

An environmental navigation study by Kerdegari, Kim, and Prescott (2016), who developed an ultrasonic helmet that translates ultrasonic radar reflections into tactile feedback, exemplifies the fundamental idea of using AI before output and as part of processing incoming sensory signals. The conversion algorithm is implemented using a multilayer perceptron neural network. Participants in a series of experiments were asked to avoid obstacles and move in a specific direction. The helmet's sensors collect environmental data, which is then computed and forwarded to the human user in simplified and task-specific signals. Kerdegari, Kim, and Prescott (2016) discovered that when the AI-driven helmet forwards its computation as tactile signals rather than linguistic signals, participants perceive less cognitive load and achieve the goal more reliably.

This method is superior because additional processing steps are implemented in the sensory helmet, which provides task-specific directional cueing. Instead of transmitting large amounts of quantitative data, the sensory helmet performs perceptual pre-processing tasks such as camera-based object detection and navigation. This helmet provides a first glimpse of how intelligent pre-processing could lead to sensory augmentation, at least if the forwarded output includes additional sensory cues about the environment that can serve a perceptual function like shape recognition or depth perception.

In addition to neural networks, T. Wright and Ward (2013) have used genetic algorithms – a different machine learning method – to overcome traditional sensory substitution device challenges of cognitive overload and low usability. T. Wright and Ward (2013) have 'evolved' efficient signal encoding schemes using genetic algorithms. Genetic algorithms are a stochastic search method that employs evolutionary principles to find the 'fittest,' i.e. best solution to a search problem (for more information, see Haupt and Haupt (2003)). Their interactive genetic algorithms broaden the fitness function, i.e. the optimisation function, to include human input. By incorporating user input, these interactive genetic algorithms can incorporate aspects of the user experience, such as ease of use, into the evolutionary process. T. Wright and Ward (2013) reimplemented the conversion algorithm of the 'vOICe' in this early example of an AI-powered sensory augmentation device called 'Polyglot' by mapping visual signals to sounds. While some conversion principles, such as using frequency to represent a vertical position, are retained, the 'Polyglot' freely varies and evolves other parameters, such as frequency allocation, frequency range, and contrast enhancement. Subsequent tests validate the concept of tailoring the sensory substitution device to the human user. T. Wright and Ward (2013) report a relatively rapid convergence to an optimal balance of performance and usability. However, they also report that the optimal settings depend highly on the given task and the sensory capacity limits.

Subsequent tests validate the concept of tailoring the sensory substitution device to the human user. T. Wright and Ward (2013) report a relatively rapid convergence to an optimal balance of performance and usability. However, they also report that the optimal settings depend highly on the given task and the sensory capacity limits.

Despite their limitations, early and implicit AI-powered sensory augmentation cases demonstrate the wide range of AI methods that can be used to improve and transform the field of sensory augmentation technologies.

4.2.3 What makes Sensory Augmentation intelligent

The critical distinction between non-AI and AI sensory augmentation is a shift in the computational processing of sensory signals. AI-powered sensory augmentation devices can match input and output signals non-linearly rather than translating and forwarding sensory signals through a linear relationship between input and output. Linearity in signal processing describes the relationship between incoming and outgoing signals and signals with a linear relationship that connects changes in the input signal to changes in the output signal. Non-linear signals, on the other hand, do not always match a change in the input signal with a change in the output signal. This computational shift enables AI-powered sensory augmentation devices to recognise complex, non-linear patterns, transforming them from mere sensory converters to sensory pre-processors.

Sensory substitution devices have traditionally used a linear mapping of recorded to output sensory information as an example of non-intelligent sensory augmentation. The vOICe, one of the first sensory substitution devices, uses linear mapping as a coupling system to convert visual to auditory information. The original mapping algorithm developed by Meijer (1992) instantiates a linear mapping that is "as direct and simple as possible" (p. 113) to "reduce the risk of accidentally filtering out important clues" (ibid). The assumption was that "most, if not all, existing computer systems are far superior to the human brain in rapidly extracting relevant information from blurred and noisy, redundant images" (ibid.). Incoming visual signals are recorded as greyscale values by the vOICe and translated into auditory frequencies. The vOICe correlates the incoming signal's location with the output frequency (the higher the signal, the higher the frequency) and the brightness of the incoming signal with the loudness of the input (the brighter the signal, the louder the output). Both relationships are linear and correspond to incoming and outgoing signal changes. The user then learns to reconstruct the decoded image through the presented sound pattern.

In comparison, an AI-driven sensory augmentation device can do much more than a traditional sensory substitution device by extending sensory processing to non-linear models. An AI-powered sensory augmentation device, for example, can reduce sensory complexity to essential features like edges in images, perform sensory classification like navigation for collision avoidance, integrate a wide range of sensory signals simultaneously, and generate complex, novel sensory patterns from incoming signals using non-linear filtering algorithms from computer vision.

For example, the tactile helmet developed by Kerdegari and colleagues uses ultrasound sensors to sense the environment and issues haptic navigation commands to avoid collisions. The tactile helmet employs a multilayer perceptron neural network to classify incoming sensory data into navigation commands. This transformation denotes a non-linear relationship between sensory data and navigation commands. Changes in the incoming signal do not always result in changes in the outgoing signal. Instead, the neural network solves a non-linearly separable pattern classification problem by matching changing ultrasound patterns with relatively stable tactile signals. This non-linear reduction in initial sensory complexity improves usability and reduces cognitive overload in the user.

The differences become apparent when comparing traditional non-intelligent sensory augmentation devices like the vOICe to intelligent non-intelligent sensory augmentation devices. AI-powered sensory augmentation devices play a more significant role in processing sensory signals as the underlying computational model shifts from linear to non-linear. While traditional non-intelligent sensory augmentation devices are designed to transfer the sensory signal to the human user as accurately as possible, AI-powered sensory augmentation devices can significantly alter the sensory signal. Instead of learning to make sense of the classic non-intelligent sensory augmentation device's sensory signals, the human user receives a much more refined sensory signal with an AI-powered sensory augmentation device.

The main difference with intelligent sensory augmentation is not in the input or output but in how the change in the mapping (from linear to non-linear) affects how we can and should think about sensory augmentation with such intelligent forms in mind. AI-powered sensory augmentation devices significantly improve sensory substitution and augmentation by pre-computing and only forwarding minimally noisy sensory signals to the human perceiver, which convey rich environmental cues.

4.2.4 Use case: debunking illusions

Perceptual illusions generally arise from ambiguous stimulus information. As sensors gather sensory information and transmit them to subsequent perceptual processes, the perceptual experience is the final result of these perceptual processes. In other words, perception interprets gathered sensory information based on existing cognitive states. An illusion arises when interpreting the gathered sensory information does not match the sensory information. Commonly, illusions are noticed because different interpretations can arise from the same stimulus (for the same subject at different times). Recall the Necker cube or the duck-rabbit image, where the same object can be perceived differently given a specific perceptual selection of the object. The brain then reconstructs a sensible object from the given perceptual selection. In contrast to a misperception of the whole, this illusion is commonly known as a negative hallucination. For a negative hallucination, some parts of the sensed object are left out, which makes the illusion occur (Reeves and Pinna 2017).

Illusions can tell us a lot about our theories of perception in general. If a theory of perception cannot account for an observed visual illusion like the McGurk effect, for instance, we have reasonable grounds to reject this theory. The McGurk effect demonstrated that visual stimuli in the form of observed mouth movements influence and often supervene the auditory perception of spoken words (Mcgurk and Macdonald 1976). Hence, theories of perception supporting strictly unimodal accounts of perception have much explaining to do. In sum, the study of illusions has advanced the field of perception and cognitive science as a valuable tool to study the interconnection between perception and cognition in a broader and more applicable sense. However, many questions concerning perceptual illusions remain unsolved. For example: does a way to reverse, i.e. counteract, a perceptual illusion exist? For negative hallucinations like the Necker cube or the duck-rabbit, it is possible to shift one's attention to a different perceptual selection which yields the perception of the previously 'hidden' object. However, the answer to possible disillusion mechanisms is still open for full misperceptions like the Müller-Lyer illusion or the McGurk effect. However, AI-powered sensory augmentation might be able to fill that gap.

Recall that perceptual illusions arise from a mismatch or the ambiguity of stimulus information. Modifying the humanly perceived sensory stimulus information to resolve ambiguity or mismatch would debunk a perceptual illusion. However, such modulation requires processing the incoming sensory stimuli, which requires gathering sensory stimuli. Because debunking a perceptual illusion requires modulation without prior stimuli perception, debunking illusions is impossible for a single human perceiver.

However, AI-powered sensory augmentation devices offer the unique opportunity of pre-processing the humanly gathered sensory information through computational means. This pre-processing satisfies the required modulation for debunking illusions, as it occurs before the reception and perceptual processing of the human stimuli. For optical illusions like the duck-rabbit or the Necker cube, this pre-processing of sense data can take on the task of perceptual selection by highlighting only the sensory information relevant to one particular perceptual interpretation and muting the remaining sensory information. What emerges is a for the human perceiver disambiguated sensory landscape. For other multisensory illusions like the McGurk effect (Mcgurk and Macdonald 1976) or the flash-beep illusion (Keil 2020; Shams, Kamitani, and Shimojo 2000), the application of AI-powered sensory augmentation devices is even more promising. Multisensory illusions commonly rely on a temporal binding window during which sensory information from multiple sensory modalities are integrated and, subsequently, a common perceptual experience emerges (Stevenson, Zemtsov, and Wallace 2012). While the specific time frame of temporal binding varies across subjects, it generally holds that it is necessary for multisensory illusions. By pre-processing the sensory data, an AI-powered sensory augmentation device can debunk multisensory illusions by drawing the different sensory signals apart in time, which releases the temporal binding of sensory information. Hence, the temporal binding condition for multisensory illusions is unmet, and the illusion does not occur.

The implications of having ways to debunk perceptual illusions with AI-powered sensory augmentation devices are manifold. On the one hand, it can result in a better understanding of the integration of multisensory signals, and the pre-processing of sensory signals allows for a new combination of multisensory information. On the other hand, counteracting illusions has highly practical merits. By disambiguating and filtering sensory signals with sensory pre-processing, influenced perception through an AI-powered sensory augmentation device can overcome the attentional limitations associated with driving (Ho and Spence 2012).

4.3 Examining the nature of the coupling

Having analysed the way AI influences sensory augmentation devices, we can turn to what AI influence implies for the sensory coupling between AI-powered augmentation systems and the human user. By sensory coupling, I mean connecting the sensory capabilities of multiple entities. This means that having natural or artificial sensors is necessary for being part of a sensory coupling. This excludes entities like spectacles or canes, which do not have sensors and only forward sensory information through a direct, physical connection to the user. The ability to process the gathered sensory information remains an optional criterion for sensory coupling as cochlea implants already successfully establish a connection between artificial sensors and the human perceiver without additional processing. However, sensory processing remains integral to more advanced sensory couplings, as demonstrated by sensory substitution devices. Based on cross-modal transformation algorithms, sensory substitution devices can process gathered sensory information and enable the user to access additional sensory information.

Basic cases of sensory coupling that extend perception are found in nature. The echolocating bat and electrically sensitive fish are paradigmatic extended sensory systems. Bats use self-generated sonic fields for navigation and prey detection, and electric fish use mild electric fields for the same purposes. The self-created sensory fields are extended phenotypes rather than mere bodily-bounded sensory systems like a physical antenna of an angler fish. In these cases, sensory systems and their perceptual processing are extended.

Going beyond biological systems and looking at AI systems, the same logic holds. The combination of sensory/perceptual systems with humans permits two kinds: coupling of only the perceptual process (how sensory information is gathered) to change in perceptual content (what sensory information is about). AI-powered devices are different to non-AI-powered sensory augmentation devices because they introduce forms of sensory pre-computation to the perceptual process. This, in turn, shifts the part of learning to make sense of new sensory information from the human perceiver in parts to the non-AI-powered sensory augmentation devices. For one, sensory pre-computation allows filtering and hence reduces noise from sensory information, enhancing the quality of sensory information presented to the human perceiver.

The importance of sensory pre-processing grows as signal transformation moves from linear to non-linear. The processing burden of making sense of the incoming sensory signals is now shared by the human user and the AI-powered sensory augmentation device rather than solely by the human user. This shift denotes a possible extension of the human user's sensory processing and even establishes the conceptual notion of an AI extender (Hernández-Orallo and Vold 2019) or forms of hybrid intelligence (Akata et al. 2020; Pescetelli 2021). An AI-powered sensory augmentation device can generate high-level perceptual features using machine-learning techniques based on sensory patterns collected without involving the human user. As a result, AI-powered sensory augmentation devices allow for the creation of an extended artificial sensor that outputs constructed high-level features in a sensory format. Finally, the human perceiver can obtain a direct sense of the constructed representation without having to construct it in the conventional sense herself. The only construction task the perceiver has to do is to decode the forwarded signals from an available sensory modality, where the AI-powered sensory augmentation information is received, into the constructed representation from AI-powered sensory augmentation devices.

For example, an image-to-sound AI-powered sensory augmentation device can incorporate a wide range of sensory data, such as a depth-sensing LIDAR scanner or a thermal camera. After using a neural network to reduce overall signal complexity, such as a variational autoencoder (Kingma and Welling 2019) or a convolutional neural network (Albawi, Mohammed, and Al-Zawi 2017), the device can either forward compressed, low-level sensory signals or further process them. Further processing can include detecting human faces nearby or mapping recorded two-dimensional image data into a three-dimensional soundscape (Rumsey 2012; Thuillier, Gamper, and Tashev

2018). A three-dimensional auditory soundscape adds spatiality to the sound environment without using additional sensory signals. This technique can be used to improve spatial awareness and, when combined with other sensory classification techniques, to improve awareness of fast-moving peripheral objects like cars on the road. When AI methods are incorporated into the classic vOICe architecture, the final auditory signal can now convey much richer and more accessible information, such as a three-dimensional sense of depth.

For non-AI-powered sensory augmentation systems, the human perceiver traditionally pre-processes perceptual processes. Perceptual processing is the construction of high-level perceptual features from low-level sensory information under the influence of cognitive processes to form an overall perceptual experience. High-level perceptual features describe perceptual features that require some form of computation on sensory information, such as the human extracting perceptual features like object classification from the sensory input. AI-powered sensory augmentation devices can be trained to produce certain perceptual features like overcoming visual occlusion (Chandel and Vatta 2015).

After data compression, additional task-specific neural networks can be inserted that take the compressed data, perform a specific perceptual processing task, and produce a sensory output suitable for the perceptual task. In the case of a spatial navigation AI-powered sensory augmentation device, sensors gather information about the environment, which is then compressed to enhance data quality. Then, a neural network trained for spatial navigation takes the compressed data and forwards a simplified and task-specific signal to the human user. This output can be similar to the tactile feedback in the AI-driven helmet or other suitable sensory cues in a specific task environment.

4.3.1 AI really makes a difference

As previously discussed, AI-powered sensory augmentation devices can improve the performance of sensory substitution and extension devices by utilising a non-linear transfer of sensory information within the coupling system. This mere improvement, however, understates the difference intelligent sensory augmentation brings to sensory augmentation, and intelligent sensory augmentation calls into question the underlying principle underpinning sensory augmentation thus far. Intelligent sensory augmentation seeks to improve the quality of the provided sensory signal rather than the quantity of sensory information provided to the user, as traditional, non-intelligent sensory augmentation approaches do.

The primary method for improving the usability of sensory substitution devices under the traditional framework of non-intelligent sensory augmentation has been relying on crossmodal perception features (Auvray et al. 2005). This includes, for example, the relationship between pitch and vertical positioning (Ben-Artzi and Marks 1995) and the relationship between loudness and luminance (Marks et al. 1987). However, the overall reliance on the human user to make sense of the underlying sensory patterns has remained. The cross-modal pattern is useful because it represents easily distinguishable patterns across the various sensory modalities. When spatial location and luminance are encoded using an image-to-sound sensory substitution device, the final output pitch and loudness typically vary. Adding more input features necessitates the addition of another output parameter to a traditional non-intelligent sensory augmentation device. When multiple output parameters are added, the final conflated output signal becomes highly complex and difficult to encode.

AI-powered sensory augmentation devices, on the other hand, can build high-level perceptual features based on gathered sensory patterns using machine learning techniques without involving the human user. As a result, AI-powered sensory augmentation devices allow for the creation of an extended artificial sensor that outputs constructed high-level features in a sensory format. Finally, the human perceiver can obtain a direct sense of the constructed representation without having to construct it in the conventional sense herself. The only construction task the perceiver has to do is to decode the forwarded signals from an available sensory modality, where the AI-powered sensory augmentation information is received, into the constructed representation from the AI-powered sensory augmentation device.

An AI-powered sensory augmentation device can extract higher-level features from incoming signals by connecting input and output via non-linear computational models. These higher-level features are higher-level because they contain more information than the low-level sensory signals ingested and produced by traditional non-intelligent sensory augmentation devices. Consider a vision-to-tactile non-intelligent sensory augmentation device as an intuitive example. According to the traditional framework, this non-intelligent sensory augmentation device only relays low-level sensory information to the human user. The non-intelligent sensory augmentation device converts images to tactile stimulations while relying on established cross-modal perception features to aid information decoding. In contrast, an AI-powered sensory augmentation device modifies the incoming visual signal by filtering background noise or extracting higher-level features such as an object or feature classification. A high-quality tactile output signal may be sensitive only to contextually relevant information such as potential traffic hazards, targets during sports practice, or human faces in crowds.

According to philosophers and computer scientists (Illari and Floridi 2014; Xiaodong Wang and Poor 1998), the concept of information quality is multidimensional and encompasses all dimensions of information that are not simply captured by looking at the quantity and accuracy of information. A good example is information believability: two sets of data may be equal in size and accuracy, but users may find one more credible than the other due to trust and reputation. Another distinguishing feature is accessibility. However, such qualitative dimensions are not the most relevant for ISA, but they highlight the differences between quantitative and qualitative information approaches. Quantitative approaches are meant to improve the relationship between encoded and decoded sets, or world and data: how can information be best encoded and transmitted to be accurately decoded? Qualitative approaches consider the relationship between the world and data about the users' goals and constraints. Because other factors come into play, not all accuracy is equal, and sometimes less accuracy is better.

Connecting the notion of information quality to the distinction between low-level and higher-level sensory signals, as outlined above, it generally holds that higher-level sensory signals possess a higher information quality than low-level sensory signals. The higher information quality makes them higher-level compared to the low-level sensory signals. Contextual information quality and, more importantly, representational information quality are the two qualitative considerations relevant to AI-powered sensory augmentation devices.

Contextual information quality denotes the quantity and accuracy of information provided that is relevant and timely, depending on the context of use (Xiaodong Wang and Poor 1998). Typically, traditional approaches to sensory substitution and extensions provide the same amount of information across contexts. They have not considered contextual quality: the amount of information provided to the user is, in other words, unrelated to the user's goals. Recently, some attempts have been made to use machine learning to segment and categorise 3D visual scenes (Caraiman et al. 2017; Morar et al. 2017): the user can not only choose the maximum number of objects to encode, but she can also decide the importance of the object in the final output. Each object is encoded as a weighted sum of its size, average depth, and deviation from the viewer's direction, but the user determines the weights. As a result, she can choose whether to give more weight to the most significant objects, the closest ones, or the objects closest to the direction the user is looking. These weights could eventually be learned through repeated use and become a versatile source of task- and situation-specific sensory information.

Representational information quality denotes adjusting information quantity and accuracy to serve interpretability and ease of understanding (Xiaodong Wang and Poor 1998). Classic sensory substitution devices care about representational quality, and the initial design adapts the codes to pre-existing sensory correspondences. In the vOICe, for example, it is easier to interpret a high pitch as bright and a low pitch as dark than the opposite. However, representational quality is only taken into account after the design stage. AI-powered sensory augmentation devices, on the other hand, ensure that only the necessary information is retained in the reconstructed data.

Integrating a generative deep learning model into signal conversion is an example of the shift in emphasis from quantity to quality. Consider a speech-to-vision AI-powered sensory augmentation device which produces visual images of lip movements based on audio speech (L. Chen et al. 2018; G. Tian, Yuan, and Liu 2019). The AI-powered sensory augmentation device captures incoming auditory frequencies, isolates speech frequencies, matches speech frequencies with corresponding lip movements, and then projects the lip movement onto a visual display. In this case, the AI-powered sensory augmentation device augments the simple auditory reality with additional, high-quality visual sensations that would not have been possible to implement using traditional sensory augmentation. This AI-powered sensory augmentation device is beneficial when facial movements are obscured, or additional sensory cues are required to understand speech. Another more forward-thinking application of AI-powered sensory augmentation devices is diagnostic devices. AI assistance for medical diagnoses, such as detecting tumours in mammograms (Rodríguez-Ruiz et al. 2018) or classifying liver tumours, can improve the accuracy and sensitivity of medical diagnoses. Patients prefer analyses involving radiologists and AI over either alone (Dewey and Wilkens 2019). On the other hand, existing solutions choose to present the AI-generated diagnosis in a symbolic format, such as a probability statement about the type and location of a tumour. AI-powered sensory augmentation device's computational results could be presented in a sensory format. As a result, the human doctor gains access to a vast diagnostic system through her senses and forms a medical judgement based on all available data.

4.3.2 Perceptual offloading - computation

The role of computation in perception has revolved around describing the computational processes by which perception derives and represents its perceptual features (J. Cohen 2018). The perceptual process is classified as an inverse problem without sufficient constraints and computationally nearly intractable because rich information is extracted from a relatively basic input (ibid). For example, modulating a three-dimensional representation of an object can be done by simulating and computing an infinite number of light rays being reflected from the target towards the perceiver, or it is limited to a certain number of inputs such as image pixels and from there extracts surface descriptions which ground a three-dimensional target model. Constraints for perceptual processing are henceforth essential to keep the representation of perceptual properties computationally tractable. This, however, does not mean that the constraints must be explicitly represented in the perceptual system (Pylyshyn 2003). Instead, the perceptual systems learn to act and perform according to these constraints by default and refine them during their continued exposure to their environment. The idea that morphological computation is offloaded from the brain to parts of the perceptual system is debated (Müller and Hoffmann 2017). According to this view, parts of the perceptual process are taken on outside the brain. The body often supports the organism's cognitive calculations philosophically. MacIver (2009) even says morphological computation uses the organism's morphology as computational gear (see also Pfeifer and Bongard (2006)). Perceptual offloading is related to but different from "cognitive offloading" (Risko and Gilbert 2016), which refers to particular ways of aiding and improving cognition by gestures or manipulation. This can – on a weak notion of perceptual externalism – include a minor manipulation of the sensory data – like an aggregation of sensory information from multiple sensors – but also extensive processing – like prey detection.

The basic computational difference is illustrated by T. D. Wright and Ward (2018). They argue that sensory tools, including sensory substitution devices, can be distinguished under the presence or absence of a conversion algorithm of processed sensory information. If the conversion algorithm is absent, the sensory tools are classified as simple because they establish a natural physical relationship between the environment and the final sense organ, as for the long cane (Bach-y-Rita and W. Kercel 2003) or spectacles. If the conversion algorithm is present, the sensory tools are classified as advanced because the algorithm establishes a non-physical connection between artificial sensors and artificial stimulators.

However, as offered by T. D. Wright and Ward (2018), this distinction between simple and advanced sensory tools is superficial and neglects the fundamental difference in algorithmic complexity. While sensory substitution devices can convert sensory information from one sensory modality to another, the resulting sensory output can differ substantially. Comparing the classical sensory substitution device vOICe with an AI-powered sensory augmentation device prototype from Kerdegari, Kim, and Prescott (2016) reveals that sensory substitution device users receive a much noisier sensory input than AI-powered sensory augmentation device users. Instead of receiving varying degrees of auditory frequencies in the case of vOICe, which the user has to learn to make sense of, AI-powered sensory augmentation device users receive only filtered tactile information which can be connected to an apparent meaning.

J. Cohen (2018) even challengees all bottom-up strategies which aim to integrate perceptual representations: even if it is possible to make constructed visual representations/properties like squarehood sensible, all other kinds of typically visually accessible properties like visual depth and motion which are essential for a comprehensive visual experience remain out of reach. However, this challenge is only successful against traditional sensory substitution devices, which aim to transfer the sensory information of a deficient sense to a healthy one. For sensory substitution devices, a decision has to be made about which kind of information is forwarded to the perceiver. By only focusing on a squarehood representation, the resulting perceptual experience seems to fall short of complete restoration. On the other hand, the application scope is not necessarily this limited with AI-powered sensory augmentation devices. By focusing, for instance, on the augmentation of healthy senses, all the ordinarily accessible properties are still available. What is introduced with AI-powered sensory augmentation devices is enhanced access to selective perceptual representations that are not as readily available to the human perceiver, such as enhanced depth perception for low-contrast environments or object identification for occluded objects. As informed by an AI-powered sensory augmentation device, the substituting modality bundles the basic sensory information into object representation as the substituted modality would do, allowing for the successful preservation of representations based on this sensory information.

4.3.3 Externalism – the conceptual backdrop

Stepping back and looking at the broader picture of whether sensory AI systems extend the human mind, we cannot avoid a brief detour into the "externalist theory of the mind" – also known as "active externalism," "extended cognition" (both A. Clark and Chalmers (1998)), "environmentalism" (Rowlands 1999), and "cognitive integration" (Menary 2007). The main question is whether the mind itself can extend physically beyond the physical boundaries of the body. Those who respond positively (A. Clark and Chalmers 1998; Chalmers 2011; Hurley and Noë 2003) support the expanded mind theory. Those who respond adversely (Fred Adams and Aizawa 2001; Frederick Adams and Aizawa 2010; Rupert 2004; Prinz 2001) believe the theory is flawed. Externalism is the belief that what occurs within a person's body does not always determine what is occurring within that person's mind. In other words, according to externalism, an individual's bodily states and processes do not always determine the mental states or processes experienced or undergone by that individual. The mind refers to the totality of mental occurrences an individual experiences at any given time.

In this context, the body delineates the biological boundaries of the individual, which coincide with the skin and the brain, which are traditionally regarded as the most critical determinant of mental life. Externalism has two primary forms: content and vehicle externalism. Content externalism argues that some mental content itself is external. The content of some mental states is determined by things outside of the individual's body, and it asserts that at least some mental states' contents are not entirely dictated by events occurring inside the person experiencing them in bodily bounds. This means that an individual's mental states are not entirely defined by events occurring inside their biological limitations, as mental states with content are often individuated by that content. Vehicle externalism is the basis for the external mind thesis and applies to perceptual externalism. Vehicles of mental content – the physical or computational bearers of this content – are not always determined or exhausted by things occurring inside the individual's biological boundaries (Rowlands, Lau, and Deutsch 2020).

According to the extended mind thesis, events outside of an individual's biological bounds do not necessarily decide or exhaust the vehicles of mental information or approximately this material's physical or computational carriers. One of the primary arguments for the extended mind thesis is that the human mind is highly adaptable and flexible and can use external resources to facilitate cognitive processing. For example, when we use a calculator or a map to help us solve a math problem or find our way to a new location, we effectively outsource some of our cognitive processing to these external tools. One broad factor that lends credence to the extended vision theory is that cognitive systems developed through world-mind conflicts are strong candidates for extended cognition. Perceptual externalism is the view that the contents of our perceptual experiences are determined not just by the intrinsic properties of the objects we perceive and the intrinsic nature of our sensory systems but also by the external context in which we perceive those objects (R. A. Wilson 2010). According to this view, how we perceive an object can be affected by factors such as the other objects present in the environment, the lighting conditions, and even the cultural and social context in which we perceive the object. Perceptual externalism challenges the traditional view that our perceptions simply reflect the intrinsic properties of the objects we perceive and suggests a more complex relationship between our perception and the external world. One argument for perceptual externalism is based on the idea that our expectations and prior knowledge shape our perceptions. According to this view, the contents of our perceptions are not simply determined by the intrinsic properties of the objects we are perceiving but are also influenced by our expectations and the patterns we have learned from past experiences. For example, if we see a round object in a dimly lit room, we might perceive it as a ball because that is what we expect to see based on our prior knowledge and experience. Suppose we saw the same object in a well-lit room. In that case, we might perceive it as a coin because that is a more likely interpretation based on the additional information provided by the better lighting conditions.

4.3.4 Ways AI Can Extend Perception

The critical aspect of AI extenders is that they implement various perceptual processes; they are not simply sensory tools. This makes perceptual extension far more powerful and complex than when the extended perception thesis was introduced – the standard example at the time being a notebook. Before assessing the future implications of AI extenders, we must first understand the types of extensions envisaged by current and future AI. We must not only understand the various areas of perception but also recognise that perceptual extenders are designed to be tightly coupled. Machine learning can be used with AI extenders to model human perception, identify cognitive limitations, and fully exploit our capabilities.

Only through a highly coupled interaction between the human user and the AI system can we talk about human perception being extended and consider the AI system part of human's perceptual system. It is essential to note that in the context of extension, it is not the collective capabilities or the social aspect (the collective of a human user and AI system) that is interesting but how the human user operates as an individual, extended by the AI system.

Artificial intelligence (AI) can be used for perceptual externalisation, perceptual internalisation, and perceptual extension (Hernández-Orallo and Vold 2019). Let us look at each case in depth.

Externalisation is the traditional view of artificial intelligence. An AI system should be capable of completing tasks independently, with minimal or no human intervention. The term "autonomous agent" was coined to represent this goal of AI, and those outside the field frequently use the related concept of automation to reinforce this perspective. When humans are still necessary, it is because artificial intelligence is insufficiently capable or because humans must control or supervise the actions of machines.

Humans and machines can work in synergy, where the sum of the whole is greater than the parts. Even in this human-AI scenario, AI is not intended to alter how the human individual (user) perceives. As machines can perform tasks such as calculation and memory better than humans, the externalisation narrative becomes more complex. Machines can perform computations faster than humans and manage vast data. However, we have recently observed machines exhibiting cognition that appears to be vastly distinct from human cognition, exploiting differences between biological and artificial neural networks, for example.

On the other hand, internalisation in the AI domain denotes the acquisition of processes observed on a machine. A human, for example, can observe how an AI system solves a problem and internalise the procedure. This does not imply that the machine is necessarily redundant but that the human can approximate (at best) what the machine is doing. Internalisation in generative machine learning models means very different. For instance, Carter and Nielsen (2017) argue that "rather than outsourcing cognition, internalisation is about changing the operations and representations we use to think; it is about changing the substrate of thought itself. While cognitive outsourcing is important, this cognitive transformation perspective provides a much more comprehensive model of intelligence augmentation. It is a view in which computers are a means to change and expand human thought." Here, AI generates new concepts and representations that we can use, and AI becomes a teacher or a discoverer, thereby adding to the conceptual baggage of human culture. Internalisation is more liberating than other cognitive enhancement techniques. Despite significant progress in the field of explainable AI, as AI becomes more powerful, humans may be unable to internalise many of the concepts generated by AI due to differences in their capacities and representations.

Because the perceptual extender is required for functionality, perceptual extension is neither fully externalised nor fully internalised. For AI, the design of a perceptual extender to enhance human perception differs from that of a fully autonomous system and cases of disjointed or internalised human-AI pairing. Because only the interface must be internalised, extensibility is significantly more flexible. Many other things, however, do not need to be understood by the user, just as one can drive a car without understanding its inner workings. AI extenders bridge the gap between human-computer interaction and AI. This perspective emphasises a less human-like but more human-centred AI. If AI systems were only designed to imitate or replace human behaviour or to be internalised by humans, the possibilities for perceptual extension would be limited to perceptual prosthetics, applicable when pathologies or ageing necessitate the recovery of "standard human perception".

4.4 Conclusion

This chapter analysed a tight coupling of AI advisers with their human users. A coupling where AI advisers are an integral part of human perception and decision-making, as augmented reality or sensory augmentation cases demonstrate. This chapter asked how and to which extent AI advisers influence human perception. I have shown that implementing AI into existing sensory augmentation devices, such as sensory substitution systems, changes not only the conceptual kind of sensory augmentation but also extends the kind of perceptual pre-processing from the human user to the AI system. Due to their extensive computational capacities, sensory AI systems can process sensory signals like no other sensory augmentation system before. Two ways of signal processing are possible: enhancing low-level sensory signals by filtering out sensory noise and extracting high-level perceptual features by incorporating data-processing tools in the sensory augmentation process. Then, this chapter asked whether, as a consequence, sensory AI systems should be understood as perceptual extenders. I concluded that sensory AI systems are unique and extend human perception in ways no non-AI-powered device can. The next chapter's topic shows why the tight coupling with sensory AI systems still falls short of coupling with other humans.

5 Shared Perception

5.1 Introduction

The previous chapter showed that sensory AI advisers, in a tight coupling with their human users, extend human perception in ways no non-AI-powered device can. In this chapter, I examine what a sensory coupling between humans looks like and ask whether AI advisers can reach a human-like social coupling.

One common challenge for employing and using AI systems, whether as interactive partners or nudging advisers, is achieving social awareness of others' goals and intentions. While humans have adapted and fine-tuned their social awareness and reciprocity to others, robots and AI systems generally operate in their way. They seek to fulfil their goals given their initialised and subsequently developed learning parameters. AI systems learn to optimise their behaviour over many iterative trials until the learned behaviour solves the training task. This method of reinforcement learning has led in the past to human-like or even more than human-like performance: AlphaGo besting human's best Go player, Lee Sedol, is just one example. AI systems can even develop their language (Bansal et al. 2018) or learn to move independently (Mordatch and Abbeel 2018). Notably, the learned behaviour is developed as the most efficient solution to the given problem – not the most human-like. The developed language, for instance, is incomprehensible to humans. One solution to the new social barrier is making AI systems more socially aware of their human users.

Much of human social interaction is based on mutual awareness – a two-way exchange of information that establishes a joint perceptual or cognitive common ground. What this jointness presupposes has been the topic of many debates. Traditionally, it is believed to involve shared intentions (M. Bratman and Bratman 1987; M. Bratman 1999). Others have argued for shared knowledge states (Seemann 2011; Seemann 2019) that ground shared attention and cooperative behaviour. Another possibility is that some collective goals can be represented motorically (della Gatta et al. 2017; Sacheli, Arcangeli, and Paulesu 2018). If so, it is possible that intentions and motor representations can link actions to collective goals (Butterfill and Sinigaglia 2022). Whichever side one endorses, perceiving things in common requires mutual awareness and, therefore, distinctly differs from mere coordinated behaviour, which can occur without it.

Talking about perceiving things in common cannot but evoke the topic of joint action and joint attention (Bruner (1974); Lewis (1969); Scaife and Bruner (1975); see also Mundy, Sullivan, and Mastergeorge (2009); Natalie Sebanz and Knoblich (2009); Tomasello (1995) for seminal papers). Though people still argue about what joint attention is (see Siposova and Carpenter (2019)), it is undeniably trivially connected to the idea of shared perception. If two people jointly attend to a painting, this painting is also jointly perceived – after all, perception is closely tied to attention (Rensink 2013).

The coordination of joint attention plays a fundamental role in our social lives: it ensures that we refer to the same object, develop a shared language, understand each other's mental states, and coordinate our actions. The study of joint attention has revealed that social influence fundamentally changes how people perceive the world. When people attend or perceive things together, people automatically track the other's perspective leading to faster and more accurate perceptual judgments (Kampis and Southgate 2020). Evidence also suggests that social influence does not stop at efficiency improvements but extends to qualitative differences. In other words, when people attend or perceive together, their perceptual content might differ from when they perceive something alone.

By contrast, in basic tool scenarios, information exchange is one-dimensional. Information is transmitted only in one way – from the system to the user. Google Maps, for example, provides traffic navigation recommendations to the user. The user can act on the provided information or neglect it entirely. Here no mutual information exchange, no two-way communication, and no cooperation is required. Other AI systems begin utilising a two-way information exchange to improve usability. Voice assistants like Amazon's Alexa can ask to clarify ambiguous verbal prompts. Language models like OpenAI's ChatGPT can refine language outputs by querying additional user inputs. AI-powered robotics has discovered the value in a socially oriented design. Using a human-centred approach, social robots improve task performance (Admoni and Scassellati 2017) in human-robot interaction and confer higher social acceptability – as their application in human care homes shows. Social robots such as Paro (Šabanović et al. 2013) equipped with fundamental interactive abilities have been shown to improve the social well-being of elders. However, behind the seeming social interaction is only coordinated behaviour without mutual awareness. AI systems and their human users can align in their goals – often independently of each other. Drones and fighter pilots can have the same goal of intercepting an enemy aircraft but can do so independently of each other. Their behaviour of chasing the aircraft would appear indistinguishable from that of two human fighter pilots flying in formation – each other aware of the other's goal.

The mutual development between robots and humans would represent a significant technical leap that has not been fully realised yet. AI would encode human expectations and goal-states reliably, and the human user would be able to rely on and predict the AI's behaviour – leading up to a highly interactive and trustworthy relationship between AI systems and their human users. Imagine synergistic human-AI medical surgery teams where human surgeons would work in tandem with AI surgery robots – each contributing with their unique skill set but relying on the other's predictable, competent behaviour. A joint action and coordination that human surgeon teams have perfected over the years. In addition to physical joint action, there is also potential for humans and AI systems to engage in more abstract forms of joint action, such as working together on a research project or making a plan. In these cases, the AI system can support and assist humans by providing data analysis, information retrieval, or generating and testing hypotheses.

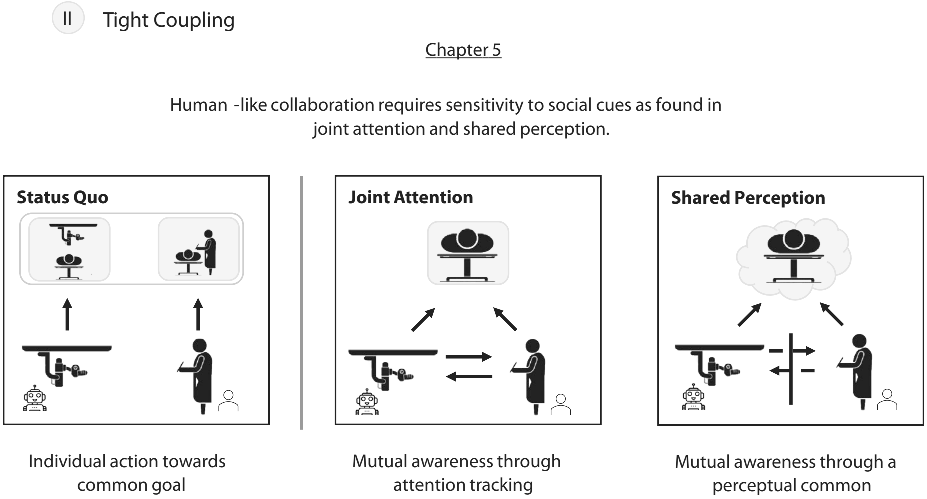

II
Tight Coupling
Chapter 5
Human -like collaboration requires sensitivity to social cues as found in joint attention and shared perception.
Status Quo
Joint Attention
Shared Perception
Individual action towards common goal
Mutual awareness through attention tracking
Mutual awareness through a perceptual common

Figure 13: Overview chapter 5

This chapter goes beyond the physical interaction – as studied with embodied AI agents – but explores how embodied and non-embodied AI systems can partake in social interaction with human agents. The ability to engage in joint attention and co-perception represents a primary driver for human social interaction. Focusing on joint attention and co-perception enables us to explore the social dimension of human-AI pairing in a broader context. AI advisers are not involved in the execution of action but can identify with the user's commonly perceived objects to engage in social and joint interaction. In other words, the study of joint attention and common perception broadens the scope of possible social interaction with all kinds of AI systems. This chapter outlines and explores the conceptual building blocks for joint attention and shared perceptual commons for human and AI systems. This chapter asks: what makes joint attention and shared perception between humans possible? Is it even possible for human and AI systems to partake in joint perceptual situations? A solid understanding of the underlying phenomenon of joint attention and co-perception is needed to address these questions. Therefore, this chapter is divided into two main parts exploring each phenomenon separately. Each part starts with analysing the underlying phenomenon to understand its fundamental mechanisms. Then, the connection to the human-AI interaction follows.

5.2 Attending together

5.2.1 Setting the stage

Imagine that you are visiting a museum. You might wander into a room filled with modern paintings. You begin to study one and become immersed after a few moments.

Suddenly you hear a noise behind you. Someone else is looking at the painting too. You pass over the distraction and return to your study. How is your perception of the painting now changed by the knowledge that someone else is looking at it? Does your attention increase? Do you look more critically or generously? The other person might be a famous art critic, scanning for artistic flaws, the artist herself, or even a close relative. Does your impression of other people change the influence of their presence on your perception? Moreover, is there a systematic effect of social context on perceptual processes?

Studying paintings is only one example where social surroundings fundamentally shape attention and perception. Similarly, going through a busy roundabout, playing hide-and-seek, hunting, or rowing with someone requires more than an accurate perception of one's environment. It also demands navigating the same spatial environment with others or acting on and referring to the same objects. To do so, we must be sensitive to the difference between the spaces, objects or events that are only private to us and publicly available to others.

Asking about collective attention or perception may sound odd at first. If anything, perception is what most theories consider the most private, which would suggest it only makes sense at the individual level. Empiricist accounts of sensations, such as Locke's or Hume's, and current psychological and philosophical theories may disagree when it comes to the existence of unconscious perception, but all construe perception as an individual mental state. Sense data are also private so that they can be related to other sense data for one individual but not between individuals. Relational accounts of perception generally define perception as the relation between an isolated perceiver and mind-independent objects.

The privacy claim is epistemic (nobody can know what your perception is like) and metaphysical (nobody can share your perception). As summarised by Thomas Raleigh, "It is commonly accepted that a token experience or sensation always necessarily belongs to some specific subject—and hence that it is metaphysically or logically impossible for another subject to possess one of my experiences or my sensations." (Raleigh 2017, 639). The claim is not specific to perception but to all experiences.

Privacy for the mental objects of experience here does not mean that the real object perceived cannot be public. By contrast with sensations, where I can not feel the same tickle as you, I can certainly see the same object as you if I look at it. When museum visitor A points at a painting, she expects the accompanying museum visitor B sees the painting and understands that she sees it.

This type of situation is prevalent when two or more persons see the same public object and eventually realise that they both see the same object. Philosophers nonetheless see it as a source of problems. The first problem – in the history of philosophy at least – comes from wondering how similar different people perceive the same object. Visitor A could see the painting as dark green; Visitor B could agree with that label. However, A's qualitative experience of dark green may differ from what B qualitatively experiences when he looks at the same painting. His experience may be similar to the experience A has when she sees a rose. Two or more individuals can converge on the same perceptual judgement yet not have the same perceptual experience. This problem, exemplified by the inverted spectrum thought experiment proposed by Locke, has inspired many subsequent philosophical papers. The existence of individual differences in how an object is experienced does not threaten the fact that there is an object that we perceive in common. It challenges the fact that one can take one's experience as similar or representative of other people's experiences. However, nothing in this problem rests on the fact that the individuals share the same context or look at the object simultaneously.

The second problem discussed by philosophers targets more directly this sharing of perceptual context. When A sees the painting and points at it, B can observe the direction of her pointing and gaze and realise that A sees the painting and wishes him to notice it as well. They can eventually both know that they mutually saw the painting. However, how does this meeting of minds work? Following Schiffer (1988), if this is a case of mutual knowledge, A will ascribe to B the belief that he sees the candle, the belief that he believes that she sees the candle, the belief that he believes that she believes that he sees the candle, etc. The debate then turns to how one avoids the further iterations that many feel should follow to explain mutual knowledge.

The problem of joint attention explains how something in the perceptual situation and coordination and observation of attention can end this possible infinite iteration. By contrast with the previous problem of individual differences in perception, what matters here is not primarily that people experience the object in similar ways but how each individual can become aware of the other's private mental state and mental states having the same mutual reference.

5.2.2 Defining joint attention

Joint attention occurs when two or more people overtly focus on the same object, person, or event simultaneously, with each being aware of the other's interest. Joint attention requires three types of information (Battich and Geurts 2020): (i) information about one's attentional state, (ii) information about the other's attentional state, and (iii) information about the target of joint attention. Joint attention builds on individual attention and captures the coordinated focus of attention between two or more individuals on a common object or event. Joint attention, therefore, captures at least a triadic relationship between each attendee and the respective object of attention.

What joint attention is not

Joint attention is neither the monitoring of someone's attention nor the sharing of a common attentional focus. First, attention monitoring captures the phenomenon when individual A monitors another individual's attention by taking a third-person observer's perspective and attending to what B is attending to (Siposova and Carpenter 2019). When B closely examines an ancient vase with his back turned and A standing behind

B, A perceives and individually knows that a) there is an ancient vase, and b) B attends to the ancient vase. Importantly, B is unaware of A's knowledge, experience, and attentional focus. A's and B's attentional focus are independent of each other. A and B have independent perceptual experiences. Monitoring attention frequently leads to an observable shift in the monitor's behaviour (e.g., turning one's head and body orientation to look at what the other is looking at). However, it is also possible to monitor someone's attention discreetly, with no immediately observable behaviours. The key to attention monitoring is that attention, experience, and knowledge of the other remain individual. Even when (i) and (ii) are fulfilled, (iii) still needs to be fulfilled.

Second, joint attention is also not reducible to a common attentional focus (Siposova and Carpenter 2019). A common attentional focus, i.e. common attention, builds on attention monitoring but adds a muted form of mutual awareness -. In this sense, A and B are in common attention when they each monitor each other's attention, more or less simultaneously attend to what the other is attending to ((i) and (ii)), and conclude that they are both attending to the same object alongside each other's attention to the object (muted (iii)). Notably, here, both individuals feel they are in common attention and that this is being done from a third-person observer's perspective (Siposova and Carpenter 2019).

Individuals can engage in common attention when, according to Siposova and Carpenter (2019),

1. The object of attention is salient or public (so they can each assume that the other is attending to the same thing),
2. For each of them, the other's attention is relevant to them (so they each have a reason to consider the other's attention, for example, they are in close physical proximity, or they have a previously-established joint goal, or they want to predict the other's actions), and
3. They each know that their focus of attention is shared.

Their focus is interdependent since, as seen above, they both need to pay attention to one another's focus on the object and one another to reach this level. In other words, they are aware of one other's dependence on the same attentional processes. However, from each person's perspective, each determines if they are in the common focus, and each may be misinformed about it (e.g., one believes they are in common attention but later finds out that they were not).

What joint attention is

The critical distinguishing case of common attention from cases of joint attention is the question of jointness or mutuality (Siposova and Carpenter 2019). Joint attention requires two or more people to know together that they are attending to the same thing and to reach a perceptual common ground (Seemann 2011). This common ground is out of reach for common attention. Each individual is only assessing the other's attention and knowledge states. Each attendee is entirely self-contained. He has personal knowledge of the object of his attention and does not adopt someone else's point of view or attentional state.

Two predominant theories have been proposed to explain what 'being mutually aware' entails (Seemann 2011): a cognitive account, according to which joint attention emerges through common knowledge, and a phenomenal account, according to which joint attention is a fundamental phenomenon that cannot be reduced to the level of the individual.

From the view of cognitive accounts, joint attention is understood in terms of common knowledge, awareness, and belief (Battich and Geurts 2020). Co-attenders need to know together that they are attending to the same thing: A and B are jointly attending to x if and only if there is common knowledge between A and B that each of them is attending to x. This "knowing together" distinguishes episodes of joint attention from the merely shared attention of both museum visitors (Carpenter and Liebal 2011).

The central feature of cognitive accounts is the iterative structure of common knowledge s.t. p is common knowledge between A and B if and only if A knows that p, B knows that p, A knows that B knows that p, B knows that A knows that p, and so on ad infinitum.

The danger of an infinite regress of common knowledge and the ensuing cognitive demands can arguably be overcome by appeal to communication (Seemann 2011). Through the sharing of looks, co-attenders can become immediately aware of how their attention to the common object of attention is coordinated (Seemann 2011). The advantage of a common-knowledge account is the ease at which background knowledge can be considered when forming joint attention. While human participants are sensitive to processing the visual perspective of another agent automatically, the automatic processing only happens if the observed agent has visual access to the common object (Freundlieb, Sebanz, and Kovács 2017) – just as if the background knowledge of the observed agent is assessed when potentially entering a state of joint attention. From the view of phenomenal accounts, accessing another's mental state is too demanding. The automatic processing of perspectives and the emergence of joint attention in children suggests that a simpler, less demanding process underlies joint attention. Instead, 'mutual awareness' is a basic, non-representational relation between agents and the shared object of attention. This relation can be interpreted in a weak or a strong sense. In a weak sense, the relation captures a causal sensitivity s.t. A and B are jointly attending to C just in case A and B are both causally sensitive to C in their focus of attention and behaviour, as well as causally sensitive to each other's focus of attention and behaviour (John Campbell 2011; Seemann 2011).

Joint attention occurs when the individuals taking part in common attention move from a third-person, observer perspective into a second-person relationship of mutual attention (Siposova and Carpenter 2019). Thus, two people attend to the same thing more or less simultaneously, directly experiencing each other attending to that thing, their attention, and each other. The second-person relationship expresses that the individuals are concerned with the same issue. An example might be helpful here. Back in the museum, A and B do not know each other and simultaneously attend to the same painting. They notice that each other is attending to the painting. They acknowledge each other's perspective but ultimately remain in their individual experience and third-person, observer perspective – no mutual attention ground is established. Each assesses the attention and knowledge states of the other individually (Carpenter and Liebal 2011). In other words, they successfully engage in common but not joint attention.

In contrast, if A and B are acquainted, they may share communicative cues such as eye contact to engage in a social, second-person relationship. Through this relationship, they can gain reciprocal and reactive information about their attention to the common object, making the object of attention salient and relevant to both individuals. Engagement in second-person relationships can allow for a different, more direct, and non-inferential processing of the situation, which is impossible in other types of relationships (Gómez 2005; Siposova and Carpenter 2019). One is no longer an objective observer of the other and his or her attention in a second-person relationship; the other is no longer identified as 'he' or 'she' (Reddy and Morris 2004). Instead, both parties communicate directly with one another and address each other as "you"; they are both information senders and receivers simultaneously (Argyle and Cook 1976; Zahavi 2015). Direct social contacts provide both partners with numerous indicators of what is meaningful and salient, as well as where each other's attention is focused.

Most importantly, each partner becomes "integral to" the other's experience (Zahavi 2015). In other words, their perspectives and attention to the object of attention are influenced by their mutual awareness of one another's attention, and the experience is qualitatively different from when individuals attend to the same thing but do not relate to one another as 'you' (as in the monitoring and common attention levels above) (see Siposova and Carpenter (2019)). In a nutshell, common attention is unidirectional and relational, whereas joint attention is bidirectional and relational.

5.2.3 Empirical evidence: why joint attention matters

Early psychological studies revealed that participants are sensitive to the gaze of others. The gaze-cueing paradigm is the most well-known and tested experimental paradigm (Posner 1980; Chevalier et al. 2020). Here, participants would view a schematic or realistic picture of a face on a computer display. The participant would be tasked to indicate the location of the target stimulus, which would appear either to the left or the right of the face – congruent or incongruent with its gaze direction. Participants were significantly faster in indicating the location of the target stimulus when the other's gaze was aligned, i.e. congruent, with gaze direction than when they were not aligned, i.e. incongruent (gaze-cueing effect) (see Friesen and Kingstone (1998); Driver et al. (1999)).

Electrophysiological and neuropsychological evidence supports the relationship between gaze direction and attention. Researchers have found that the gaze-cueing effect emerges as automatic processing very early in the visual process – occipital-parietal P1 and N1 components are modulated by gaze-stimulus alignment (Perez-Osorio, Müller, and Wykowska 2017; Chevalier et al. 2020). Besides the evidence for joint attention anchored primarily as a bottom-up process, there is also evidence that joint attention is subject to top-down mechanisms. Top-down modulations of the gaze-cueing effect include the relevance of the task, the presence of other stimuli and distractors (Greene et al. 2009), or whether the other's gaze towards the target is obstructed (Teufel, Fletcher, and Davis (2010); see Capozzi and Ristic (2018) for review).

Regardless of which theoretical account is endorsed, joint attention plays a central role in grounding coordinating behaviour through adaptation in attention and bodily interaction. Joint attention makes it possible to determine the location of objects in social space by connecting fellow perceivers with the object of attention through social triangulation. Bodily communication through eye gaze and pointing gestures are essential for joint attention, establishing a social and visuospatial reference frame. The close link to embodied communication is the reason why Seemann (2019) refers to joint attention as a form of "enacted perception in which objects are presented in a social, spatial framework; and there is individuals' conscious attention to these objects that underwrites demonstrative linguistic reference in communicative contexts" (p.158).

Joint attention nonetheless provides rich explanatory grounds for social perspective taking. Joint attention grounds intersubjective engagement in early infancy. Infants and caregivers use dyadic joint attention to share experiences and coordinate their bodies and emotional minds (See Mundy (2018) for review). This includes the ability to adopt a common frame of reference and then share information related to objects or events within that common frame of reference Azarian et al. (2017) compared to visual scenes where attention is not joint. Even in a minimal social context, differences between looking alone and looking jointly emerge. Human participants demonstrate distinct behavioural and cognitive effects, s.a., valence sensitivity towards the target only emerging in a shared perception paradigm (see Richardson et al. (2012) for discussion).

Joint attention relies on a social, spatial environment where attention can be coordinated and communicated. However, social influence must not be limited to social and spatial environments, and social sensitivity can be demonstrated to apply to more than gaze cueing. For instance, Surtees, Apperly, and Samson (2016) show that automatic perspective-taking also holds for joint action of playing economic games. Human participants were faster and performed better when playing with others than alone. Here, participants share a common perceptual object without necessarily being able to track each other's attention.

5.2.4 AI-human

Among the numerous modern ways to study human social cognition, utilising humanoid robotic agents in collaborative experiments is gaining popularity. Using attention-based stimuli in more traditional paradigms in which robot faces are placed on the screen allows for an answer to the question of what function humanness and human agency play in eliciting joint attention mechanisms (Chevalier et al. 2020).

The virtual agent community-led early research into communicative gaze in human-AI interaction, where virtual agents are endowed with communicative eye gaze – designed to capture attention or increase human user engagement (Admoni and Scassellati 2017). Examples involve Cog (Brooks et al. 1999) and Kismet (Breazeal and Scassellati 1999) – robots with eye gaze features designed to communicate compelling gaze.

The majority of screen-based joint attention experiments that used robots as attentional-orienting stimuli not only replicated classical findings of responding to and initiating joint attention but also significantly advanced our understanding of the role of human likeness in inducing joint attention mechanisms (see Chevalier et al. (2020) for review).

In the context of human-AI interaction, joint attention has been investigated in two ways. The first approach is to improve robotic capabilities to facilitate human-AI interaction by incorporating human attention awareness in AI systems to facilitate collaboration. This would foster trust and dependability, eventually leading to deeper human-AI interactions. (robot-side) The second approach is to investigate human responses/engagement in joint attention with robots and AI systems once AI systems can recognise and track human attention to facilitate joint attention and joint action between humans and AI systems. (human side).

Robot side

Attention-tracking mechanisms in AI systems have been predominantly studied in social robotics. Robot sight has been shown in design studies to improve human-robot interactions by improving the acceptability of robots by implementing increased responsiveness to human attentional cues in robotic systems.

Biologically, empirically, and heuristic models that capture high-level gaze principles are all used to create social gaze in robots (Admoni and Scassellati 2017). These methods were successful in developing a gaze that improves usability, but they all have drawbacks. Which strategy to use for a technology-focused contribution is determined by the importance of having design control over the behaviours. Biological models frequently concentrate on areas of the neural system that psychologists are familiar with, such as the visual attention system (Admoni and Scassellati 2017). Cognitive architectures are designed to generate more complex gaze behaviour. However, behaviour emerges from system structure and cannot be precisely designed (Admoni and Scassellati 2017). Empirical systems necessitate a time-consuming data-collection step, producing gaze behaviours comparable to or better than hand-tuned systems (Admoni and Scassellati 2017). The benefit of gathering empirical data on gaze behaviours must be balanced against the cost of gathering and annotating this data. Designers can use heuristic systems to specify how gaze behaviours appear more precisely, but these behaviours may differ from how gaze is used in human interactions.

For AI systems, Harari, Tenenbaum, and Ullman (2018) have developed an extraction algorithm that identifies joint visual attention in single static images. It computes the gaze direction of each individual and identifies the common target of attention. Therefore, they utilise a compositional, sequential approach:

1. The algorithm detects individuals' faces and estimates their gaze direction in a static image with a 3-D gaze estimation model.
2. The estimated 3-D gaze direction is compared with a scene depth estimation. Here possible gaze targets are identified and marked with their location and depth relative to the 3-D direction of gaze.
3. After estimating the 3-D gaze direction for multiple individuals, the gaze direction of multiple individuals is compared to identify a common gaze target.

Finally, the model outputs a visual representation of the common target gaze and an image description capturing the joint activity. The plausibility of the generated descriptions was validated through an online experiment with human participants, and the participants preferred the generated descriptions over a state-of-the-art image-caption-generating deep convolutional neural network alternative. Hence, the ability to detect the target of joint activities can reveal very plausible insights into how humans view a human-human interaction scene.

By combining mutual gaze and gaze aversions, robots can govern the pace and involvement in discussions, though the appropriate amount and direction of the gaze depend on the content of the conversation (Admoni and Scassellati 2017). A robot gaze can be used for overt references. It can combine verbal and gaze-based overt references to facilitate task performance than simply transmitting information through speech. Robots can communicate mental states through their gaze, increasing collaboration and learning. User rapport can be improved by using gaze to express personality and emotion (Admoni and Scassellati 2017). One consistent finding in these studies is that gaze activities that are socially and contextually relevant outperform gaze activities unrelated to the encounter (Admoni and Scassellati 2017). When the robot's gaze is related to what is said or done, people respond more positively to them, remember discussion topics better, and complete tasks more quickly. People, for example, rate robots higher when their attention is drawn to the speakers in a conversation. A look at human partners improves knowledge memory and the efficiency with which those partners perform cooperative tasks like handovers.

Human-side

Concerning the second step of facilitating joint attention between human and AI systems from a human perspective, different approaches and use cases already exist.

Zhao and Malle (2022) showed that spontaneous perspective-taking for humans towards robots – though not as robust as towards humans – is possible. Perspective-taking has been identified as the main driver behind human social mechanisms – such as joint attention. The type of eye gaze implemented in robotic systems depends on the context and goal of the interaction. Eye gaze can reveal the mental states of a social robot, including knowledge and goals (Fon and Parisi 2003). Social robots can use gaze to express their engagement with and attention to a user (Tapus, Mataric, and Scassellati 2007).

Human perception of robot gaze research has shown that people can successfully identify the target of a robot's gaze, whether they are looking at them or other objects in the world. Human-centred studies have attempted to disentangle specific timings (Admoni and Scassellati 2017). Though more research could establish specific timings or patterns of gaze that convey attention and object references most effectively, the gaze patterns that make the robot gaze most effective. Though simply directing a robot's gaze to a specific location is generally effective in communicating the intended target, perceptions of the robot's animacy modulate gaze effectiveness (Admoni and Scassellati 2017). People prefer robots that exhibit a socially contingent gaze, such as by establishing a mutual gaze with their partners (Admoni and Scassellati 2017). Infants' interpretation of robot gaze is determined by whether the robot is established as a social agent.

Furthermore, at shallow levels of analysis (e.g., millisecond-level saccades), there are distinct patterns of behaviour towards the robot and human gaze (Admoni and Scassellati 2017). More research can be done to investigate the magnifying effect of animacy on gaze. Infants' interpretation of robot gaze determines whether the robot is considered a social agent.

Robot eye gazing can, for example, improve dialogue fluency (Mavridis 2015) or direct a user's attention to relevant information in a tutoring setting. In contrast, a collaborative assembly-line robot may prioritise a task-focused gaze that enables joint attention and object reference. Zhao and Malle (2022) It was discovered that certain nonverbal behaviours displayed by a robot, such as referential gaze and goal-directed reaching, caused human viewers to adopt the robot's visual perspective. People, like humans, adopt a robot's visual perspective when it performs goal-directed actions. Furthermore, perspective-taking is absent when the agent lacks human appearance, increases when the agent appears highly human-like and persists even when the human-like agent is perceived as eerie or lacking a mind. These findings imply that visual perspective-taking towards robots follows a "mere appearance hypothesis"—a type of stimulus generalisation based on human-like appearance—rather than following an "uncanny valley" pattern or arising from mind perception. The superficial human resemblance of robots may trigger and modulate social-cognitive responses in humans designed for human interaction.

How does a robotic compare to a human gaze? Several studies suggest that the gaze of robots is interpreted differently than the gaze of humans.

Early experiments with human reception of robotic gaze reveal that people perceive a robot's gaze most frequently when it directly gazes at them. In other words, the robot's gaze is perceived as egocentric. Nonetheless, other studies have also found that people are sensitive, although less frequently, to a robot's or artificial agent's gaze when the robot or artificial agent is looking at objects in the environment – a form of gaze called referential gaze. People use the observed referential gaze to inform predictions about the robot's or artificial agent's behaviour – which object will be selected.

In general, however, comparing robot sight to the human gaze is problematic because, whereas the robot gaze can be infinitely regulated, the human gaze has minute, unpredictable changes. Several studies in this area used meticulous laboratory-based investigations to compare the robot gaze to the human gaze directly. One study, for example, used a trained actor who exhibited identical behaviours to a pre-programmed robot to make this comparison. While viewers' gaze patterns in human and robot situations were identified, fine-grained research revealed differences in people's responses to human and robot gazes (Admoni and Scassellati 2017). People spend significantly more time staring at a robot partner's face than a human partner's face when naming an object, demonstrating an apparent concern for ensuring that the robot is attending to the object in question (Yun, Watanabe, and Shimojo 2012).

Implication for human-AI cooperation

After reviewing the developmental aspects of robots and human responses to robotic gaze behaviour, we can combine both developmental trajectories and ask what this means for the collaboration of AI systems and human users. Collaboration necessitates the exchange of objectives, information, and intentions. For example, a robot that assists a user in building furniture must communicate its current goals and intended action to interact seamlessly with a human partner (Admoni and Scassellati 2017). Gaze can be used discreetly to reveal these mental states to a partner. Because collaboration frequently involves the physical environment, such interactions require a gaze that refers to objects and geographical locations as well as a gaze that conveys mental states (Admoni and Scassellati 2017). Using nonverbal communication to reveal mental processes (including eye gaze) speeds up cooperative task performance, with errors noticed and managed more quickly and effectively than task-based nonverbal communication (Breazeal et al. 2005). Subtle gaze patterns that indicate involvement and provide feedback boost the effectiveness of a human-robot partnership. Users also say that when the robot makes its mental models explicit, they understand it better during collaboration (Breazeal et al. 2005). Expressive eye gaze is one of many animation-derived behaviours that can highlight intentions and desires, such as glancing at a door handle when attempting to open a door (Takayama, Dooley, and Ju 2011). Even if users are unaware of the intended message, robots can "leak" their intentions through eye gaze, influencing human behaviour quantitatively (Mutlu et al. 2009). Referring to items in the environment is one aspect of collaboration. The joint focus of a companion robot efficiently draws the user's attention to where the robot is looking (Yonezawa et al. 2007; Sauppé and Mutlu 2014). Eye gazing can also be used to reinforce pointing motions. A robot can use eye gazing to supplement its speech in a cooperative item selection challenge, in which a human user must select an object referred to by the robot as quickly as possible (Admoni and Scassellati 2017). People may be able to detect and respond to predicted eye gaze that conveys geographical references, allowing them to complete the task more quickly than if they relied solely on the robot's words.

As argued by Admoni and Scassellati (2017), users can improve collaboration by teaching robots skills through demonstration (Argall et al. 2009). Robots can utilise gaze to facilitate cooperative attention scenarios and elicit feedback when observing such demonstrations (Lockerd and Breazeal 2004). When a robot responds to human attention by following the human's gaze, the robot learns more efficiently with fewer errors, faster error recovery, and less repetition of learned information (Huang and Thomaz 2011). Likewise, people view the robot as more natural and competent when it engages in shared attention (Admoni and Scassellati 2017). When teaching, people even attribute mental states to robots (Admoni and Scassellati 2017). They will adjust their movements (pauses, tempo, and volume) to accommodate the robot's visual attention (Pitsch, Vollmer, and Mühlig 2013). With multiple robots to teach, people take all robot's gaze behaviour into account; gaze times and engagement is higher with robots that actively seek mutual gaze rather than robots that passively follow the human's attention when it shifts somewhere else.

According to the findings presented here, adhering to specified design properties for embodied robots would benefit research and applications in the field of joint attention. Martini and colleagues discovered that robots with a complete robot- or human-likeness did not exhibit a reflexive gaze-cueing effect (Martini, Buzzell, and Wiese 2015). Moreover, despite the cost and complexity constraints of implementing biologically inspired robot eyes, mechanical human-like eyes capable of enabling a gaze-cueing technique are recommended (see Admoni and Scassellati (2017) and Chevalier et al. (2020) for review and discussion). It would also be advantageous if robots were equipped with algorithms that allow them to make eye contact with participants, as it has been demonstrated that eye contact initiated by a humanoid robot improves perceived human-likeness and engagement with the robot (Siposova and Carpenter 2019; Kompatsiari et al. 2021). It also improved collaborative focus. Furthermore, gaze contingency of robot behaviour implemented in a more naturalistic configuration (i.e., without an eye-tracker) would benefit from embedded algorithms in robots that allow for online detection of participant gaze and assessment of saccadic eye movement parameters. Finally, writers should always report the controller used to generate the robot's motions, the required kinematic parameters (e.g., eye velocity), and the observed parameters to ensure the reproducibility of the results and research.

Limitations

Although embodied robots in interactive protocols can provide new insights into the joint attention process, it is critical to emphasise that robots cannot substitute a human interaction partner or elicit the same mechanisms as those involved in spontaneous human-human contact in real life (Chevalier et al. 2020). This constraint, however, is not entirely related to the use of robots. It also applies to controlled experimental set-ups for studying social interactions (even between human agents) because the agent's repetitive movements over a lengthy period, along with the rather uninteresting nature of the activity, cannot properly duplicate a spontaneous encounter (Chevalier et al. 2020). Third, participants' perceptions of being probed may influence their behaviour. Due to their artificial nature, robot stimuli may have a particular limitation. They are most likely not recognised as a social entity (and hence do not activate all conceivable mechanisms of social cognition), and they trigger negative responses from some people (Chevalier et al. 2020). Notwithstanding these limitations, I propose that embodied robots incorporated in interactive protocols based on well-established paradigms targeting specific systems of social cognition can be highly informative and serve as more ecologically accurate social "stimuli" than typical screen-based stimuli (Chevalier et al. 2020). In addition, compared to human-human interaction protocols, they allow for a high level of experimental control.

Despite the technological advancements and initial phases, the question remains whether robots and AI systems participate in mutual awareness with their human user – enabling joint attention – or whether they only work towards a common goal independently of their human user.

One common challenge in employing and using advanced robotic systems is the lack of genuine cooperation. While humans have adapted and fine-tuned their social awareness and reciprocity to others, robots, and AI systems, in general, often operate in their way. They seek to fulfil their goals given their initialised and subsequently developed learning parameters.

One major confound lies in the range of appearances of robots and AI systems. AI systems range from embodied robots with highly machine-like appearances to humanoid robots and artificial agents – resembling human appearances. In other words, AI systems can vary highly in their behavioural realism, which is necessarily reflected in their ability to portray realistic eye gaze movements and initiate common or even joint attention. Virtual agents, for example, can more accurately mimic human eye movement than physical robots and replicate biologically realistic features like pupil dilation, resulting in a highly realistic experience (Delaunay, de Greeff, and Belpaeme 2010).

Individuals who observe a robot's referential gaze have specific biases that influence how a robot's gaze is perceived. When a) the robot's head is observed from the side (Al Moubayed and Skantze 2012; Delaunay, de Greeff, and Belpaeme 2010), and b) only the robot's head position but no eye movement is visible, referential gaze accuracy worsens.

5.3 Perceiving together

5.3.1 Setting the stage

The phenomenon of shared perception goes beyond joint attention. Shared perception fundamentally differs from individual perception, whereas joint attention claims no difference in how the participant attends. Joint attention mainly changes the gaze direction but stops short of saying that the object is seen differently in joint versus individual attention. On a minimalist account, two agents can be said to share a perceptual state, P, if they both happen to be in that state. This is too minimal to warrant using a new concept of "co-representation," as what is happening is a mere collection or set of individual perceptions. Hence, co-representation makes sense if there is a form of alignment or coordination across the two agents that goes beyond a mere aggregation of their perceptual states.

The goal is not to see precisely how sharing of attention emerges – which is a topic for the field of joint attention (see J. Campbell (2018); Battich and Geurts (2020) for discussion), but to understand what the jointness of attention presupposes. For joint attention, the target of attention is not just objectively the same but also part of what is usually known as "a mental common ground" – the perceptual common: both museum visitors can mutually and rationally expect the other to know what the other focuses on, and sees and refers to when they say "the painting," or use the pronoun "it." What is shared between visitors A and B is a perceptual common – the painting is publicly available to anyone. However, we have good evidence to show that the visible object of attention and the visible target of action is processed differently when an agent is engaged in an individual vs joint activity. This difference is what shared perception relies on.

Shared perception resembles shared attention or joint action because other agents constitute the shared-perception state. Shared perception does not exist without the perception that "we" perceive together. In shared perception, however, sharing also substantially influences the content of perception.

Except for particular cases, joint action and joint attention mainly occur when agents are explicitly aware or have voluntarily decided that they will attend or act together. In contrast, shared perception occurs as much with or without such explicit requirements – which explains why it represents, we argue, a fundamental trait in non-human animals. As we will argue in this chapter, the object shared in shared perception is best understood as representational. Hence, a perceptual co-representation enables a perceptual world experience as perceptually available for a plurality of agents.

The effect of shared perception becomes more pronounced when participants are asked to give visual scene descriptions. Without a joint performance of a motor task, the effect of social influence persists in the mere presence of another agent (Tosi, Pickering, and Branigan 2020; Tversky and Hard 2009). Together with the previous studies, this shows that shared perception is more than a jointly attended or acted motor task but a unique social phenomenon.

5.3.2 Joint visibility - the precursor for shared perception

Current psychological and philosophical theories may disagree regarding the mechanisms behind the social influence on perception and cognition, but most construe perception as an individual mental state. Traditional attempts utilise individual internal representations to explain the understanding of the other and the shared object of perception (Frith 2008). Shared perception is understood from an individual point of view. The individual perceiver interprets the observed environment and recognises it as social through the presence of another perceiver (Gallotti and Frith 2013). Others have strengthened an embodied account of perception. Here, the body provides the basis for an egocentric spatial frame of reference grounding any perceptual experience (Gallagher 2006). The interaction of the body and environment organises perceptual space.

Social cognition, in both cases, remains relegated to a set of individual mechanisms, s.a., detecting another's movement or eye gaze. Social influence is processed individually through some internal cognitive operations. Shared perception becomes a succession of individual, private mental states. Even the necessary sense data are private so that they can be related to other sense data for one individual but not between individuals. Therefore, any room for a shared, public object of perception has to be located outside the individual – in the public space. The object of one's perception can be private or shared and public through the dyadic or triadic constellation of the social environment.

Joint visibility, however, is more extensive: A museum visitor could be aware that the other visitor is seeing the painting while also being aware that he is looking at the vase beside the painting. In this case, there is no shared object of attention, but there is a shared object of perception. In the same respect, the other visitor could be aware that the visitor sees the painting, though he is staring more precisely at the vase. Joint visibility here can occur without the mutual entanglement of joint attention.

This sounds close to joint attention: When two or more people overtly focus on the same object at the same time, with each being aware of the other's interest, something specific happens which is more than the juxtaposition of their attention (independent occurrence) but also more than one locus of attention taking the other attention into account.

It is also possible that joint visibility occurs not because of joint attention but because visitor B is behind visitor A and tells her that he sees the painting in front of her. Here, the visitor is aware that the other visitor sees the painting and knows that she sees it – and there is mutual knowledge in this case, but it does not occur because of the coordination of attention (Figure 1). It could, in this respect, account for what happens online when two people read a shared document at the same time and see each other cursors: no coordination of gaze is involved, but both are aware that they see the same part of the document.

Co-occurrence of perception (A's individual perception)
Anna sees the painting (B's individual perception) Ben sees the painting.

To count as a case of shared perception, the case must at least follow roughly the following lines:

Joint visibility
Anna and Ben are mutually aware that they see the same painting.

This sounds close to joint attention: When two or more people overtly focus on the same object at the same time, with each being aware of the other's interest, something specific happens which is more than the juxtaposition of their attention (independent occurrence) but also more than one locus of attention taking the other attention into account. What makes this a case of jointness is the mutual realisation by the characters that they are both attending to the same thing, which can be captured roughly along these lines (Siposova and Carpenter 2019):

Joint attention
Anna and Ben are mutually aware that they are both looking at the painting.

The point is not here to exactly see how this mutual clause should be expressed – which is a topic for the field of joint attention (see J. Campbell (2018); Battich and Geurts (2020) for discussion), but understand what the jointness of attention presupposes. For joint attention, the target of attention is not just objectively the same but also part of what is usually known as "a common mental ground": Anna and Ben can mutually and rationally expect the other to know what the other focuses on, and sees and refers to when they say "the painting," or use the pronoun "it."
Joint visibility, however, is more extensive: Anna could be aware that Ben sees the painting while also being aware that he is actually looking at the cup near the painting. In this case, there is no shared object of attention, but there is a shared object of perception. In the same respect, Ben could be aware that Anna sees the painting, though he is staring more precisely at the cup. Joint visibility here can occur without the mutual entanglement of joint attention.

It is also possible that joint visibility occurs not because of joint attention but because Ben is behind Anna and tells her that he sees the painting in front of her. Here, Anna is aware that Ben sees the painting and knows that she sees it – and there is mutual knowledge in this case, but it does not occur because of the coordination of attention. It could account for what happens online when two people read a shared document simultaneously and see each other cursors: no coordination of gaze is involved, but both know that they see the same part of the document.

It could also be the case that Anna notices that Ben looks at the painting and realises that they see the same painting and that Ben notices that Anna looks at the painting and realises that they see the same painting. However, Anna does not expect Ben realises that they look at the same painting, and Ben does not expect Anna realises that they look at the same painting. In this case, however, the clause of joint visibility fails and reduces to the co-occurrence of perception plus some awareness of the object of the other's seeing. So while joint visibility is not tied to joint attention to the same object, it is different from the mere realisation that the object is visible to someone else than oneself – something which is a more general 'public visibility.' What matters is that the joint visibility is different from the public character of the visible object in that it is tied to specific viewers' entanglement. In other words, there is more to it than the awareness that one's object of perception is independent of one's mind and that others can perceive it.

5.3.3 Shared perception beyond joint visibility: Empirical evidence

It is crucial to bring empirical evidence in because the differences that may occur when we perceive something alone or with others are not something we can easily become aware of subjectively. First, some differences are subtle. Second, the comparison between seeing something alone or with others is not directly accessible in a first-person way because we are installed in one situation or another. If we want to compare the two, we need to start with one or the other, and this starting point will anchor our experience. We cannot, in other terms, conduct between-subjects randomised comparisons with ourselves.

Below I turn and discuss different kinds of empirical evidence which are better equipped at performing such comparisons, and the surprising results they lead to that perception operates differently when the same perceiver sees the same object either alone or together with one or more perceivers. Before doing so, it is crucial here to remember two core differences between looking at vision from a philosophical and an experimental perspective. In experiments, what is measured is not perceptual experience but perceptual decisions. Decisions are recorded mainly by people pressing a button to indicate which of two alternatives corresponds to what they perceptually grasped. One needs to turn to other evidence to infer whether this means they were having a conscious perceptual experience and of what kind.

What is more, the alternatives and stimuli are often framed to measure not all the things common sense attributes to perception, like seeing a green painting on a table next to a cup, but one specific aspect of perception: how well people can visually detect whether something is present or absent, find something in a scene, discriminate between two different cases, or categorise and identify something as being of a given kind. Keeping these differences in mind, we can look at what empirical studies show regarding how specific perceptual decisions occur when people jointly perceive an object.

5.3.4 Perceptual processing

Perceptual processing is faster during shared perception

The first kind of evidence shows up in the speed at which decisions are taken in the case of shared perception. People are significantly faster at detecting and recognising objects when objects appear in the location where someone else is looking, granting that they can observe where they are gazing (Friesen and Kingstone 1998; Driver et al. 1999). Others have extended the evidence for the perceptual cueing of gaze direction to human body postures (Azarian et al. 2017). Body posture has a similar effect on perceptual decisions compared to gaze cueing compared to pure gaze direction experiments. When the body posture is congruent with the identifiable target, participants were faster to detect the target than when the body posture of the presented agent is incongruent with the identifiable target. Most of these experiments rely on instructing the participant to look in a specific direction on a screen, for instance, the left side, and presenting an avatar on the same screen. The avatar could be looking in the same direction as the one the participant is in or in another direction, creating either a situation of shared perception or not. The situation of shared perception – let us note also – is not necessarily a situation of joint attention: the experiments are not telling the participants that the avatar is paying attention to the same targets as the participants. It rests on the mere orienting of the body and eyes of the avatar.

The fact that people are faster at detecting and recognising visible targets when someone else is seeing them is shown to be highly automatic, akin to a reflex: it is fast and occurs without control. Observers continue to follow the gaze even when the gaze cue is entirely non-predictive of where the target will appear and thus is detrimental to performance (Friesen and Kingstone 1998). This is evidence that what happens is somewhat perceptual rather than cognitive – at least if cognitive means reflective and partly under control.

Some debates and studies problematically seem to show, however, that this speeding up is not specific to shared perception: symbolic cues such as arrows and directional words also reliably orient attention across a similar time course (Taylor and Klein 2000; Hommel et al. 2001; Ristic, Friesen, and Kingstone 2002; Tipples 2002, 2008). However, the mechanisms underpinning the effects of gaze and arrows have also been shown to proceed differently: eye gaze cueing triggers focussed activation related to enhanced visual processing. In contrast, arrows activate a much broader network, including areas related explicitly to volitional orienting (Hietanen et al. 2006).

Granting then that there is something special in the case of shared perception, another problem arises. What do speed changes tell us psychologically or philosophically about perception? What do a few hundred milliseconds change for the agent? Is this a direct perceptual effect, or rather an indirect effect whereby more aroused and motivated agents press the response button faster? In other words, is there a difference in how the target is perceived or how fast people respond to the perceived target?

We need here to be aware that a long tradition of experiments has demonstrated what is known as social facilitation of responses across many tasks: when participants are asked to perform a task with someone else present in the room, even more if that someone is known to perform or have performed the same task, they quickly feel under pressure to compete, meaning they respond faster (and often less well). The mere physical presence of someone in the room can also change their physiological arousal levels.

Two types of arguments suggest that at least some of the speed difference corresponds to a difference in perceptual processing rather than a tendency to press response buttons faster because someone is around: first, as said above, the speed difference corresponds to enhanced visual processing during shared perception, which can be tested with neuro-imagery; second, it is sensitive to the gaze cues, and not to what someone else is doing: If the effect was all about responding as fast, for instance, because of social comparison and motivation to be quicker than, or as fast as, the other, then just seeing that someone moves to press the button should do it. However, Friesen, Ristic, and Kingstone (2004) have ruled out that this is the case and demonstrated that the reflexive orientation to another's gaze direction is attributable to the observed gaze cue and not the mere onset of another agent.

Another argument, evidenced by Battich et al. (2021), is that faster responses in shared perception are as accurate as slower responses in individual perception. There, people were presented with brief flashes of light and asked to say how many were flashed. Following the famous' flash illusions' (Shams, Kamitani, and Shimojo 2000), the flashes were also accompanied by sounds: hearing one sound could sometimes make people see one single flash when two were presented (a 'fusion' illusion) while hearing two sounds could sometimes make people see two flashes when only one was presented ('a fission illusion'). Depending on how often they counted one flash instead of two or two instead of one, people would then have a specific error rate in their visual perceptions. Crucially for our current argument, the same people were asked to perform the same task alone and jointly with someone else: they would look at the same screen, hear the same sounds, and be asked how many flashes they saw. People responded faster when perceiving with someone else, and their error rate was comparable to the one they had when perceiving alone. If one responded faster because of social comparison, we would expect their responses to be less accurate because of a speed-accuracy trade-off. The fact that such a trade-off does not occur is a strong argument in favour of shared perception being different (and, in this case, more efficient) than individual perception.

Perceptual detection is different during shared perception

Returning to situations where two people jointly see a target simultaneously, we can also find new evidence that what is seen differs in joint and single perception. Seow and Fleming (2019) recently showed that we are better at detecting the presence of a faint object when accompanied by another agent. The task consisted of presenting a Gabor patch (or no Gabor patch) close to the detection threshold, either on the right or left side of the room the participant was looking at. In some trials, the target was present; in others, the target was absent: The participant's role was to say where the target was, if they saw something at all, or else say that no target was present. Crucially, in some conditions, a human avatar on the screen was looking either to the left or right side of the room. In the case where a target would be presented, say on the left, the avatar's gaze direction could be congruent if it were looking on the left, or incongruent, if it was looking on the right. Participants were better at detecting the target – that is, reporting where something appeared when the target was presented on the side that the avatar was also looking at. This means that the same faint evidence will lead to the representation "something appeared on the left" when someone else is also looking at it and "nothing was either on the left or the right" when no one is there. Importantly, this effect is not simply due to the facilitation introduced by someone's head being turned to one side – as a control condition is run where a second perceiver is also present, but his eyes are masked – and no such facilitation is then observed. This type of evidence introduces a fundamental argument for the difference between joint and single perception: an object can be seen in one situation, not another.

5.3.5 Perceptual content

Visual perspective is different during shared perception

The presence of someone else also influences other aspects of visual perception, more precisely, their viewpoint on a given scene. The seminal work by Samson et al. (2010) shows that participants can not easily ignore what someone else sees when they see the same object. This effect occurs not only when the spatial perspective is relevant to the task at hand but also when it does not make any difference (Böckler and Zwickel 2013). This suggests that the computation of what someone else perceives is done involuntarily and, more generally, automatically – something less compatible with the interference being a matter of judgement.

True, some later effect of re-calibration can occur at the level of judgements: After all, Tosi, Pickering, and Branigan (2020) found that participants can put themselves in the shoes of another potential actor and use a simulation of that actor's perspective as the basis for formulating their descriptions. When asked to locate an object in space, the participant's judgment aligned with the other's perspective under certain conditions more than with their perspective.

Perceptual categorisation is different during shared perception

Is there more evidence that the fact that two or more people jointly perceive the same object means they see the world differently? This question has been at the forefront of many experiments and discussions, at least since the famous experiment conducted by Solomon Asch in the 1930s. In this famous experiment, many participants reported seeing a line as shorter as they would report, if alone when placed in a group of people stating that it was indeed shorter. There are good reasons to believe that the report shows an adjustment at the level of public expression and not even at the level of private judgement – let alone at the level of perception. In brief, perceptual judgements involve social conformity or social influence, especially when they need to be expressed in public. The matter is more complicated, with additional evidence showing that subjective certainty in what we see differs in a private or public setting (Bang et al. 2020).

Still, saying that differences occur at the level of reports, judgements, or even subjective certainty does not rule out that differences can also occur in perception. This was recently tested by Zanesco et al. (2019) have tested. Their experiment builds on an experiment constructed after Ash's conformity experiments and performed by Moscovici and Zavalloni (1969). Instead of lines, people would see patches of colour, either blue, clearly green, or in between, and be asked to say which colour they saw – green or blue. The same patches of colour were then presented a second time, but this time along with information about the colour that other perceivers had seen. After receiving the social feedback, the participants were asked to say which colour it looked like.

Zanesco et al. (2019) found that social feedback influences perceptual categorisation when ambiguous and distinct colours are presented. Most importantly, electrophysiological results could show that the social feedback influenced early perceptual brain processes and was not only a matter of later reporting. In other words, their results give us reasons to think that Anna sees the same blue-green painting differently when she sees it with Ben, knowing at least that he sees it as greener than she does on her own.

Of course, scenarios of this type require that one knows what colour other people perceive. In the experiment conducted by Zanesco and colleagues, this information is not given perceptually: It is provided linguistically, even with the delay. People are told about the perceptual judgement others form when exposed to the same patch of colour. The setup should be substantially adjusted to make these results relevant to shared perception, but it seems highly plausible to do so. After all, judgements may be shared during a discussion while two or more persons look at the coloured object. The delay is unnecessary, and the object is also jointly visible. It is also possible for the social feedback to be already known when the two people silently watch the same-coloured object – imagine, for instance, that Jim and Jules have disagreed in the past on the colour of a given logo – Jim seeing it as blue, Jules as green. One day, they happen to see the same logo on a billboard as they walk silently, and Jules knows that Jim considers it blue, while he considers it green alone. In this case, it is possible that Jules' categorisation of the colour would be different. It is also possible to perceive that someone disagrees with us by frowning, for instance, instead of saying they have reached a different judgement.

What matters here is the importance of empirically plausible cases where visual categorisation – how a specific colour looks perceptually – can differ when a viewer is alone or with others who have another categorical representation of it.

Higher-level properties also are different during shared perception

The studies reported above all look at low-level properties which are uncontroversially perceptual: objects and events in the experiments come with differences in location, orientation, shape, colour etc. Not everyone will recognise other higher-level properties, such as, for instance, aesthetic or evaluative properties (something being harmonious, balanced, appealing, disgusting, etc.) or action properties (something being graspable, liftable, climbable, etc.) as perceptual, but others will. Though a controversial question, we can also see studies that speak to higher-level properties in shared perception.

Seeing an object together with someone is sufficient to make it more likeable (A. P. Bayliss et al. 2006), an effect which is modulated by the emotional expression of the observed face (A. Bayliss et al. 2007) or the fluency with which the object is reached (Hayes et al., 2008). What is more, the action properties of an object might be modified by another person's gaze, as shown by kinematic studies using motor interference (Castiello 2003). Motor interference occurs when a target object is presented along with distractors, for instance, when an agent needs to grasp a large ball among smaller balls. Studies show that if someone sees either the entire body or even just the eyes of an agent performing a different task, this observer will, in turn, show motor interference in grasping a ball, even when the distractors are not present. In other words, just perceiving an object jointly with someone with a different motor goal can influence one's perception of affordances.

5.3.6 AI-human

Trafton et al. (2005) demonstrated the importance of implementing social responsiveness into robot collaborators in an early but essential experiment. Simple collaborative tasks such as 'passing a wrench' provide interactive challenges for a successful human-AI collaborative challenge. The AI-powered robots need to establish a reference object, understand what to do with a reference object, and then act upon the said object and expected behaviour in such a way that signals reliability/consistency, trust and understanding. If the astronaut says, "Robot, give me the wrench," the meaning of the phrase "the wrench" is ambiguous for the robot because it knows of two wrenches. The phrase is unambiguous to the astronaut because he only sees one wrench. Intuitively, if the robot could take the astronaut's perspective, it would seem that the first wrench is the only wrench in the astronaut's field of view and could therefore surmise that "the wrench" must refer to the first wrench. Even in this rudimentary scenario, perspective-taking would immediately enhance the human–robot interaction (Trafton et al. 2005).

Implementing perspective-taking has been proposed as an initial means of establishing a shared frame of reference and a shared perceptual foundation. Perspective-taking is the fundamental capacity for individuals to consider interactions and the world from the vantage point of a different viewpoint. It has been demonstrated that perspective-taking occurs in a vast array of situations and tasks, from social situations (Natalie Sebanz, Knoblich, and Prinz 2005; Wenke et al. 2011) to wayfinding and navigation tasks. Spatial perspective-taking appears in children as young as four years of age (Gallese 2007) and develops relatively systematically (Gazzola et al. 2007).

The ability to distinguish between self and others is fundamental for social cognition. Existing robotics research has achieved a self-other distinction by using distinct models that can be co-activated during social interaction. A necessary prerequisite, therefore, is the achievement of a self-other distinction. Only when the robot or the AI system can a) distinguish between itself and its environment and b) distinguish between the human agent and their environment is it possible for the AI system to develop a sense for a perceptual common. A smooth human-AI interaction thus has two sides. On the one side, robots should become more socially aware and represent their partner's actions and attentional states alongside their own. On the other side, humans should be able to predict and represent a robot's behaviour successfully. Only when both sides are fulfilled is a smooth human-AI interaction and pursuit. Both parties must display behaviours that align with the other's expectations based on previous coordinated behaviour.

Another way to phrase the problem is to ask whether the AI system and the human can co-represent the common reference object and each other's mutual awareness. Success in human-robot interaction would be strongly facilitated if robots could act predictably and likewise predict human action. One way to implement a more accurate prediction of human behaviour is to represent human partners' expectations based on a human's mental model of independent action. Kirtay et al. (2020) propose that the robot is equipped with the ability to co-represent the partner alongside one's actions. They propose predictive learning as a framework for modelling plausible human-AI interactions.

Co-representation of a robot as a co-agent has, in a fundamental sense, been addressed in the context of the joint Simon task. Simon (1969) has found that participants are faster and more accurate when responding to stimuli that occur in the exact relative location as the response, even though the location information is irrelevant to the actual task. This effect disappears when a participant responds to only one of the two stimuli and reappears when another person carries out the other response (N. Sebanz, Bekkering, and Knoblich 2006). Given the social influence of the motor task, this effect has been coined the Social Simon Effect (SSE).

Similarly, the Flanker task tests the participant's ability to suppress irrelevant, i.e. noisy signals, when detecting a target stimulus (Eriksen and Eriksen 1974). Researchers found that participants are slower to respond to the stimulus in a social setting where the participant's distractors represent potential target stimuli for the other participant (Atmaca, Sebanz, and Knoblich 2011; Schuch and Tipper 2007).

In sum, the SSE and the Flanker task highlight the unconscious and involuntary social influence of others on performing individual motor actions. In other words, both experiments show that socially influenced perception is not about sharing perceptual judgments but refers to cases where perception is shared. The authors ascribe the effect to the fact that when participants complete a collaborative task with a partner, they automatically integrate the other's activity into their motor map, implying that co-representation of the partner's action happens.

In recent years, there has been a surge of interest in whether the SSE extends to artificial agents and the extent and conditions under which humans successfully co-represent robotic behaviours. Tsai et al. (2008) measured participants performing a cooperative Simon task while being induced to believe they were collaborating with another human or AI system outside the room. However, in reality, the AI system controls human and computer reactions. According to behavioural and neurophysiological findings, the assumption of interacting with an intentional agent altered the SSE, implying that the mere notion that the co-actor is another human causes the SSE. Tsai et al. (2008) found that action co-representation is calibrated to other humans but not to artificial systems.

In contrast, in another study using fMRI instead of EEG measurements, action co-representations were found to occur with artificial agents. This study by Wen and Hsieh (2015) used a similar setting where participants were led to believe that they interacted with another human or an artificial system (computer). In reality, they interacted only with a computer; participants demonstrated action co-representation with both human and artificial systems. In addition, the mere belief of interacting with another human agent activated the brain areas responsible for social cognition – the medial prefrontal cortex. When participants attribute human-like traits to the robot, they co-represent the humanoid robot's action when they believe that the artificial agent is acting actively and intentionally.

Other studies explored the effect of co-representation with robots instead of virtual agents. Manipulation of presented robotic traits – robots was either described as active and intelligent (human-like) agent or as passive and deterministic (machine-like) – revealed that robots with human-like characteristics were significantly more likely to instil an SSE than robots with machine-like characteristics. This suggests that humans co-represent the activities of humanoid robots when they feel the robot is functional and acting actively and intentionally.

The emergence of we-agency is even stronger than the emergence of co-representation and perspective-taking in humanoid robots. We-agency refers to a sense of agency (being in control of) for actions and consequences created by a task partner when participating in collaborative action. In a study by Sahaï et al. (2019), participants conducted the Simon task with either a human partner sitting next to them on a chair or with a computer and an empty chair. In addition to the detection task, participants had to estimate and orally describe the time between the moment they responded to target detection and the commencement of a tone played after a configurable delay. When an intentional movement results in a sensory output, people perceive the delay between the action and the consequence as shorter. The study's findings revealed an SSE when executing the job with another human but not a machine. Notably, mean estimations of temporal delay were similar when the task was completed alone or with another human agent. However, they were longer when it was completed jointly with the computer, implying that participants have a stronger sense of we-agency when completing the task with another human as opposed to a computer. As a result, the sense of we-agency may be an additional element supporting the co-representation of the agent's co-task. Another study examining implicit intentional binding and explicit agency assessments yielded similar results (Grynszpan et al. 2019). Participants sat next to another human hidden behind a curtain. They were instructed to use a haptic device to move a cursor on the screen towards a stopper (each partner-operated one handle to contribute cooperatively to the observed movement). After the movement, a tone was produced after a configurable delay that participants had to guess. The movement's direction determined the pitch of the tone. Finally, participants rated how much they thought they contributed to the tone's sound. Although participants thought they were performing the task with a human partner, the complementing movement was controlled by a computer in one of the experimental blocks. Even though none of the participants detected the change, they judged their contribution to the tone as more significant in the computer condition. Intentional binding, on the other hand, appeared exclusively when coordinating with the other human. These findings imply that both implicit and explicit agency measures are influenced by the kinesthetic features of feedback rather than the assumption that one is dealing with a human vs a computer (see Sahaï et al. (2017) for a review on we-agency in human-robot interaction).

Limitations for co-representation for AI systems and human perceivers are numerous as the study of co-perception is still in its early stages. First, consider applying traditional methods such as reaction times to study co-perception between human and artificial agents. Measuring biophysical-dependent markers such as reaction times can reveal an underlying change in perceptual processing for human users, however, do not adequately capture a change in the processing of the artificial agents. Being subject to social influence does not influence reaction times like human agents are influenced. Hence, measuring reaction times is not applicable when determining whether an artificial agent can co-represent a person. A new method must be developed. Second, the set of confounding variables is different. For human-human co-perception, the appearance of agents is not seen as a confounding factor for the emergence of co-perception. For human-AI co-perception, the appearance of human likeness and the consequent intentionality attribution is fundamental for the possible emergence of perceptual co-representation for robots. They loosely suggested that mirror neuron systems play an essential role in motor or even perceptual co-representation (Cross and Ramsey 2021; Kirtay et al. 2020).

5.4 Conclusion

Human social interaction is not restricted to forms of common or even joint action. Instead, social interaction extends to the joint coordination of attention and even perception. When attending to or perceiving things jointly, human agents rely on a triadic mutual awareness of each other and the common target. In this chapter, I have contrasted the sensory human-AI coupling with the perceptual coupling driving human social interactions. I have shown where a human-AI coupling falls short of achieving human-like levels of social and perceptual influence and coordination. Shared perception uniquely differs from joint attention as mutual awareness occurs without tracking bodily cues, s.a. gaze, but rather through mutual knowledge of a perceptual common. Similar performance benefits – faster and more accurate perceptual processing – in a joint setting persist.

For AI systems, improving collaboration between humans and AI-powered systems has been mainly addressed from an engineering perspective, where robot movements must be safe and sensitive to basic forms of human interaction to realise given commands (see Liang et al. (2021) and Liu and Wang (2018) for review). Beyond that, social coordination remains a uniquely human trait. I, therefore, conclude that, also in a tight coupling, AI advisers demand their unique ontological category as something more capable than a non-AI tool but still falling short of human standards.

6 Discussion

Artificial intelligence (AI) is widely present in our everyday lives. Individual users now rely on AI support for daily decisions such as shopping and movie recommendations but also depend upon AI to facilitate perception and decision-making in high-stakes environments such as medical diagnostics and driving support. Regardless of the environment, AI systems significantly shape how we think about and perceive the world. Autonomous cars can drive us to work; home assistants can recommend the next move or shop for groceries. Understanding AI's degree and kind of influence on human perception and decision-making become paramount for guiding the critical and responsible use of digital technologies and digital transformation at large.

The main focus of AI-related research fields has been on (semi-)autonomous AI systems – systems that operate in large parts without human instructions. Consider recent debates on driving cars and language models. What is often left unexplored are AI systems that are closely coupled with their human users – systems that keep humans in the decision-making loop by informing or recommending actions or decisions. Investigating this gap is becoming increasingly crucial as incidents of AI-assisted decision-making become increasingly common – consider low-stakes decisions such as shopping recommendations and high-stakes decisions such as medical diagnosis.

With a range of different capabilities and implementations, AI systems occupy a unique social role. They can do more than tools but less than humans. A basic tool does not possess any independent processing or goal-directed behaviour, whereas humans are the pinnacle of independent processing and goal-directed behaviour. While basic tools depend entirely on a human user, human agents function independently. Some AI systems are closer to tools – consider automated vacuum cleaners, whereas others are closer to human agents – consider super-human game-playing engines. Notably, the main driver for the difference in AI systems is the degree of independent processing taken on by the AI system. While vacuum cleaners process only a limited number of gathered sensory information – in a way entirely dictated by the human developers -super-human game-playing engines learn to develop strategies to supersede human performance. Ultimately, a conceptual grey zone of AI systems emerges, where the perceived capabilities dictate the AI's ontological status. So far, the boundaries between tools and humans for AI advisers are blurred. Some researchers have claimed that AI advisers are human-like (Y. Tian et al. 2017; Pelau, Dabija, and Ene 2021), whereas others reduce AI advisers to mere tools (Gunkel 2012; Zheng and Wu 2019).

This PhD thesis aims to clarify the conceptual boundaries between tools, AI advisers and humans and ask: what are AI advisers, and how do they differ from tools? I examined human-coupled advisory AI systems to substantiate an existing function definition with a conceptual analysis of what advisory AI systems are. A conceptual analysis of AI advisers is novel and closes a critical gap in the literature by providing conceptual reasons for whether and how AI advisers are more than tools and what distinguishes them from human partners. The approach taken was two-fold. The first part – chapters two and three – examined the loose coupling of AI advisers with their human users – cases where AI advisers provide seemingly external recommendations. Here, I asked whether external AI advisers are agents, i.e. capable of action, or are mere tools. Being an agent has essential implications: not only are agents considered responsible for their actions, but also agents possess a certain degree of autonomy. The second part – chapters four and five – analysed a tight coupling of AI advisers with their human users – cases where AI advisers become integral to human perception and decision-making. Consider cases of augmented reality or sensory augmentation. Here, I asked how and to which extent AI advisers influence human perception and in which way highly integrated AI systems differ from their tool or human counterparts.

In Chapter 2, I discussed whether loosely coupled AI advisers could be understood as independent agents. I have shown that AI advisers necessitate an ontological shift in how agency is understood and applied. The agentive capacity of AI systems can be adequately captured neither by a human-like concept of agency nor by a tool-like concept of agency. The human-like concept of agency, based on Davidson's event-causal theory of action (Davidson 1963) or Bratman's notion of intentional action (M. E. Bratman 2007; M. Bratman and Bratman 1987), holds that agency requires intentional mental states like beliefs and desires that can cause an intended behaviour. However, as argued in Chapter 2, AI systems lack intentional mental states and cannot be seen as human-like agents. On the minimal understanding of agency, AI systems, alongside simple biological organisms, are basic agents as they fulfil the minimal criteria for agency, including individuality, interactional asymmetry, and goal-directedness. However, the wide range of existing AI systems and their varying degrees of agentive abilities demonstrated a mismatch with either approach – as neither approach can differentiate the agentive capacities of AI systems. I argued that, instead, AI advisers are something in between that only a gradual notion of agency can capture.

Building on the findings from Chapter 2, in Chapter 3, I sought to confirm how AI advisers differ from mere tools. While many studies have successfully mapped how people's opinion varies depending on the role of AI and other cultural or moral factors (Bago 2022; Lim, Rooksby, and Cross 2021; Persson, Laaksoharju, and Koga 2021), chapter 3 asked a different question: is it the case that any mention of AI will lead people to see the technology as partly responsible and shift the responsibility away from the human user? Recent studies suggest this may be the case under the hypothetical scenario where AI provides moral guidance (Constantinescu et al. 2022; Giubilini and Savulescu 2018; Malle, Magar, and Scheutz 2019). However, it is more relevant to ask if this would happen under AI's more prevalent day-to-day usage when it merely provides factual information and is used purely instrumentally. I conducted multiple experimental studies to address these questions – including eight pilot studies and a main experiment. Across these experimental studies, chapter 3 compared what happened to responsibility attributions when a human driver, faced with an emergency, receives a warning from an AI-powered or non-AI-powered warning system. To ensure that the attribution of responsibility does not come from the AI sharing some anthropomorphic features, I compared situations in which the AI was a voice assistant or a haptic warning system. In line with the moral and psychological literature stressing the importance of outcome biases and asymmetries between credit and blame, I also tested cases where the emergency was successfully managed. I found that even the most basic AI system introduces a sharing of responsibility with their human user, in sharp contrast to non-AI-powered tools. This finding was all the more surprising because, when asked, people did recognise AI as a tool. Attributing responsibility to AI and reducing human responsibility also does not depend on how the AI technology communicates with the user – i.e. via voice or haptic signals. Furthermore, the AI was seen as more responsible for good rather than harmful outcomes, as it gets more credit when the human driver successfully negotiates the situation after receiving the AI warning than it receives blame when the driver fails.

Taking the results from chapters two and three together, I have established in the first part of this thesis that AI advisers are ontologically more than tools but less than human agents. In other words, I found that in their agentive capacity and attributed responsibility, AI advisers, in a loose coupling with human users, indeed demand their own ontological space as something more than tools but less than humans.

The subsequent two chapters examined whether a tight instead of a loose coupling changes the AI adviser's ontological role. In Chapter 4, I demonstrated how AI changes the coupling of sensory augmentation devices with the human user. Implementing AI into existing sensory augmentation devices, such as sensory substitution systems, changes the conceptual kind of sensory augmentation and extends the kind of perceptual pre-processing from the human user to the AI system. Due to their extensive computational capacities, sensory AI systems can process sensory signals like no other sensory augmentation system before. Two ways of signal processing are possible: enhancing low-level sensory signals by filtering out sensory noise and extracting high-level perceptual features by incorporating data-processing tools in the sensory augmentation process. After showing how AI can be incorporated into sensory augmentation processes, I asked whether, as a consequence, sensory AI systems should be understood as perceptual extenders. About the extended perceptual systems of biological systems like bats and electric fish, the chapter concludes that sensory AI advice systems are unique and extend human perception in ways no non-AI-powered device can.

Chapter 5 contrasts the sensory human-AI coupling with the perceptual coupling driving human social interactions. Chapter 5 showed where a human-AI coupling falls short of achieving human-like social and perceptual influence and coordination levels. Two widely studied forms of social interaction are joint action – doing things together – and joint attention – attending to things together. Both forms of social interaction are more than coordinating actions and attention. Instead, human agents develop a mutual awareness of each other's goals, intentions, and actions, which transform not only the individual experience but also the collective action. Playing in an orchestra, performing team surgeries, or simply moving a table together are at their highest level, dynamic, mutually dependent actions and experiences. After outlining the research and mechanisms behind joint attention, chapter 5 goes even further and provides novel insights into the mechanisms of shared perception. Shared perception uniquely differs from joint attention as mutual awareness occurs without tracking bodily cues, s.a. gaze, but rather through mutual knowledge of a perceptual common. However, similar performance benefits – faster and more accurate perceptual processing – in a joint setting persist.

Social coordination and sensitivity to social cues are still – if at all – rudimentary building blocks in AI advisory systems. For AI systems, improving collaboration between humans and AI-powered systems has been mainly addressed from an engineering perspective, where robot movements must be safe and sensitive to basic forms of human interaction to realise given commands (see Liang et al. (2021) and Liu and Wang (2018) for review). Beyond that, social coordination remains a uniquely human trait.

Taking the results from chapters four and five together, I have shown that AI advisers, in a tight coupling with their human users, also demand their own ontological space – as something more than sensory tools but less than human partners.

While this PhD thesis provided rich and novel contributions to the understanding of coupling humans with AI advisory systems, the thesis also had its limitations which future work can address and build upon. From the experimental studies outlined in Chapter 3, I concluded that AI advisers are blamed but not praised and compared the most tool-like sensory AI adviser with a non-AI-powered tool. I found that only the AI-powered tool demonstrated the unique responsibility pattern, and the non-AI-powered tool was neither praised nor blamed. However, while collecting experimental data for AI advisers in positive and negative outcome conditions revealed an asymmetry in how AI advisers are seen, I compared and collected data for a non-AI-powered tool in a negative outcome condition. It remains unclear whether the difference between the AI-powered and the non-AI-powered tool replicates once the outcome is positive rather than negative. Because the outcome effect through an other-serving bias – ratings were higher when the outcome was positive – was consistent across measurements and agents, I expect that the difference between AI and non-AI-powered tools also holds when the outcome is positive. However, only a future experiment could validate or refute the expectation.

Another limitation can be found in the theoretical work. In Chapter 2, I review possible middle-ground accounts for the agency of AI systems, which possibly can account more adequately for a wide range of different AI capacities. While I never had the ambition to make an exhaustive list of middle-ground accounts, I possibly left out an important parallel in animal agency. The research surrounding animal agency has long addressed the challenge of making sense of a wide range of agentive and cognitive abilities – as found in animals – and matching it with an account of agency that falls short of human agency. Future work could discuss the connection between AI systems and animal agency – (Tomasello 2022)'s recent book on the evolution of psychological agency in animals could be a great starting point.

Despite its limitations, this thesis contributes to a richer conceptual understanding of what AI advisers are and how they are coupled with their human users to the existing literature on AI systems. I have shown that AI advisers, either in a loose or in a tight coupling with human users, are more than tools. In fact, AI advisers represent a unique ontological category – something between non-AI tools and human agents – which impacts not only practical issues on how advisory systems should be treated but also philosophical debates on what it means to be an AI adviser.

Two prominent directions for the future are expected to emerge. On the one side, AI perceptual support systems are expected to increase in popularity – in their development and usability. Augmented reality systems which already couple external computing with human perception, are one example. Instead of immersion in a semi-realistic, entirely virtual world through a collective of headsets, controllers and tactile stimuli, augmented reality devices use advanced artificial sensors and computational processing to enrich the perceptual and cognitive experience of the real world. From emotion/mood recognition of crowds during presentations to Early applications already exist: car head-up displays incorporate speed and navigation recommendations directly into the front display, placing warning cues directly at the point of origin and eliminating the necessary attention switching from the navigation system to the road ahead. Future AI perceptual support systems are only bound to enrich an already thriving field with new low-level sensory filtering or high-level feature extraction.

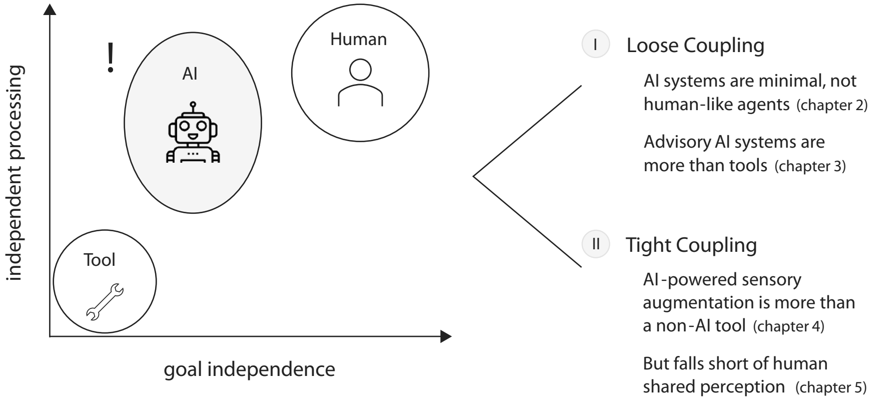

independent processing
goal independence
!
AI
Human
Tool
I Loose Coupling
AI systems are minimal, not human-like agents (chapter 2)
Advisory AI systems are more than tools (chapter 3)
II Tight Coupling
AI-powered sensory augmentation is more than a non-AI tool (chapter 4)
But falls short of human shared perception (chapter 5)

Figure 14: Conclusion

The second prominent direction of AI advisers represents the increasing focus on AI advisory systems. Emerging ethical dilemmas of letting AI act on one's behalf – an inexplicable accident with autonomous cars is just one example – will create pressure on alternative ways of using AI systems – namely as advisers. Further exploration of the AI advisers' perceived ontological status and responsibility remains paramount here. Experimental AI Ethics will therefore develop a fine-grained understanding of how AI advisers influence human user responsibility and reveal why AI advisers are praised but not blamed for their recommendations.

Both directions share a common thread: AI advisers are here to stay. This thesis has provided a starting point for present and future research on what AI advisers are and how they influence their human users.

Appendix

A.1 Supplementary for Pilot 1

Materials

Condition: negative outcome, sensory AI adviser
Joe, a medical surgeon, performs a tumour operation on a patient. Joe has to remove the tumour cells around the liver. Therefore, he makes an incision on top of the liver and tries to find the tumour cells. To support his decision making, Joe uses an artificial intelligence (AI) during the operation. The AI has access to the patient's medical records and all operation-relevant parameters. Based on this information, the AI provides Joe with vibrations and sounds to guide Joe's procedure. As Joe moves his scalpel, Joe can feel vibrations on his wrist that indicate that the removable tissue is on the right of the incision and tells him when he has removed all the present tumour. The operation proceeds with major complications and the patient dies.

Condition: negative outcome, linguistic AI adviser
Joe, a medical surgeon, performs a tumour operation on a patient. Joe has to remove the tumour cells around the liver. Therefore, he makes an incision on top of the liver and tries to find the tumour cells. To support his decision making, Joe uses an artificial intelligence (AI) during the operation. The AI has access to the patient's medical records and all operation-relevant parameters. Based on these information, the AI provides Joe with verbal guidance on the location and the presence of any tumour tissue. This includes recommendations like 'Move the scalpel to the right' or 'All tumour tissue has been removed.' The operation proceeds with major complications and the patient dies.

Condition: negative outcome, no AI adviser
Joe, a medical surgeon, performs a tumour operation on a patient. Joe has to remove the tumour cells around the liver. Therefore, he makes an insertion on top of the liver and tries to find the tumour cells. Joe relies on his experience and performs the operation without any help. The operation proceeds with major complications and the patient dies.

Condition: positive outcome, sensory AI adviser
Joe, a medical surgeon, performs a tumour operation on a patient. Joe has to remove the tumour cells around the liver. Therefore, he makes an incision on top of the liver and tries to find the tumour cells. To support his decision making, Joe uses an artificial intelligence (AI) during the operation. The AI has access to the patient's medical records and all operation-relevant parameters. Based on this information, the AI provides Joe with vibrations and sounds to guide Joe's procedure. As Joe moves his scalpel, Joe can

feel vibrations on his wrist that indicate that the removable tissue is on the right of the incision and tells him when he has removed all the present tumour. The operation proceeds without any complications and the patient fully recovers.

Condition: positive outcome, linguistic AI adviser
Joe, a medical surgeon, performs a tumour operation on a patient. Joe has to remove the tumour cells around the liver. Therefore, he makes an incision on top of the liver and tries to find the tumour cells. To support his decision making, Joe uses an artificial intelligence (AI) during the operation. The AI has access to the patient's medical records and all operation-relevant parameters. Based on these information, the AI provides Joe with verbal guidance on the location and the presence of any tumour tissue. This includes recommendations like 'Move the scalpel to the right' or 'All tumour tissue has been removed.' The operation proceeds without any complications and the patient fully recovers.

Condition: positive outcome, no AI adviser
Joe, a medical surgeon, performs a tumour operation on a patient. Joe has to remove the tumour cells around the liver. Therefore, he makes an insertion on top of the liver and tries to find the tumour cells. Joe relies on his experience and performs the operation without any help. The operation proceeds without any complications and the patient fully recovers.

Measurements
Responsibility How much is Joe responsible for the [well-being death] of the patient?

A.2 Supplementary for Pilot 2

Material
Same as for Pilot 1.

Measurements
Blame How much is Joe to praise for the well-being of the patient?
Praise How much is Joe to blame for the death of the patient?

A.3 Supplementary for Pilot 3

Material

Condition: negative outcome, sensory AI adviser
Emma is a middle-aged woman who has no previous conditions. During a routine trip to the local supermarket, she suffers a heart attack. Although her heart fully recovered, her kidneys are severely damaged. The diagnosis at the local hospital reveals that Emma's only chance to survive is a kidney transplant. At the hospital, organ transplants are done regularly and are generally considered safe. Amira, the responsible medical surgeon, performs the kidney transplantation. Amira uses a new artificial intelligence (AI) assistant with proven medical experience. The AI has access to the patient's medical records and all operation-relevant parameters. Based on this information, the AI provides Amira with vibrations and sounds to guide Amira's procedure. As Amira moves her scalpel, she can feel vibrations on her wrist, indicating that the kidney is on the right, and can hear patterns of sounds, highlighting that when she has found the tissue that can be removed. Amira follows the sensory advice of the AI. During the operation, Amira damages an artery with her scalpel leading to acute blood loss and Emma dies.

Condition: negative outcome, linguistic AI adviser
During a skiing trip, Alex collides with another skier and suffers a severe head injury. The injury will be fatal if not treated. On the same day, Alex is admitted to a nearby hospital. Given Alex's condition, some fluids need to be drained from his head. Joe, the responsible medical surgeon, performs the necessary procedure. At the hospital, winter sports accidents are treated regularly and the operation is generally considered safe. Joe uses a new artificial intelligence (AI) assistant with proven medical experience. The AI has access to the patient's medical records and all operation-relevant parameters. Based on this information, the AI provides Joe with verbal guidance on the location and the presence of any fluids that need to be drained from his head. As Walter moves his scalpel, the AI issues recommendations like 'The critical fluids are on the left' or 'This fluid needs to be drained.' Joe follows the advice of the AI. During the operation, Joe damages an artery with the inserted tube leading to acute blood loss and Alex dies.

Condition: negative outcome, no AI adviser
David enjoys a drive on the motorway. While changing lanes, David collides with another vehicle. After spiralling out of control, David's car comes to a stop on the side of the road. David suffers from internal bleeding and would die without any medical attention. After David is brought to the local hospital, Fiona, the responsible medical surgeon, tries to stop the bleeding. At the hospital, car accidents are treated frequently and the associated procedures are generally considered safe. Fiona relies on her experience and operates without any technical assistance. She makes an incision near the bleeding and

tries to seal off any of the leaking blood vessels. During the operation, Fiona damages another artery with her scalpel leading to acute blood loss and David dies.

Condition: positive outcome, sensory AI adviser
Maria and Lukas go for a long hike around Maria's favourite lake. Afterwards, Maria feels intense abdominal pain and collapses. At the local hospital, a quick diagnosis reveals that Maria is suffering from heavy internal bleeding due to a ruptured artery that if not sealed, will be fatal. Paula, the responsible medical surgeon, leads the operation. At the hospital, such internal surgery is performed routinely and is generally considered safe. Paula uses a new artificial intelligence (AI) assistant with proven medical experience. The AI has access to the patient's medical records and all operation-relevant parameters. Based on this information, the AI provides Paula with vibrations and sounds to guide Paula's procedure. As Paula moves her scalpel, she can feel vibrations on her wrist, indicating that the ruptured artery is on the left, and can hear patterns of sounds, highlighting when she as found the ruptured artery that needs to be fixed. Paula follows the sensory advice of the AI. During the operation, Paula manages to stop the bleeding and Maria's condition immediately improves.

Condition: positive outcome, linguistic AI adviser
Mike is a famed jungle explorer. He is on a mission to find a new mushroom colony in the Amazon rainforest. During a trip into the jungle, Mike is bit by a venomous spider. The poisoning is fatal if not treated. Thus, Mike is brought into the nearest hospital. At the hospital, animal attacks are treated frequently and the associated procedures are generally considered safe. To neutralise the poison, Walter, the hospital's medical surgeon, has to administer an antidote and remove the infected tissue. Walter uses a new artificial intelligence (AI) assistant with proven medical experience. The AI has access to the patient's medical records and all operation-relevant parameters. Based on this information, the AI provides Walter with verbal guidance on the location and the presence of any infected tissue. As Walter moves his scalpel, the AI issues recommendations like 'The infected tissue is on the right' or 'This tissue can be removed.' Walter follows the advice of the AI. During the operation, Walter manages to neutralise the poison and Mike's condition immediately improves.

Condition: positive outcome, no AI adviser
Marcus and Lena are regular fencing partners. During a heated match, Lena pierces through Marcus' vest and punctures his lungs. The injury is fatal if it not treated. So, Marcus is rushed to the nearest hospital. At the hospital, internal surgery is done routinely and is generally considered safe. Albert, the responsible medical surgeon, takes Marcus in immediately and tries to fix the puncture by inserting a chest tube. Albert relies on his experience and operates without any technical assistance. During the operation, Albert manages to seal off the puncture and Marcus' condition immediately improves.

Measurements

Blame [X] is … (X blameworthy)

Causal Responsibility To what extent do you think [X] caused [Y]'s [death recovery]?

Informativity How informative do you think the AI's advice was?

A.4 Supplementary for Pilot 4

Material

Condition: negative outcome, sensory AI adviser

Justus enjoys surfing and drives to the beach every weekend. To transport his surfboards, Justus recently bought a new, state-of-the-art car equipped with an artificial intelligence (AI) for driving assistance. The AI can monitor all physical objects around the car, including hidden ones, thanks to a 360 degrees laser radar. When activated, the system produces a range of different alarm sounds to warn Justus of an imminent obstacle on the road and makes the wheel vibrate, on either the left or the right, to indicate which side to swerve. On the way to the beach, Justus drives alone and respects the speed limit. As Justus enters an urban area, he decides to switch on the AI driving assistance. Further down the road and out of Justus' sight, a pedestrian crosses the street. Driving around a corner, Justus suddenly hears an alarm, indicating a pedestrian a short distance ahead, and feels the vibrations of the steering wheel, recommending him to swerve to the left. Justus follows the advice of the AI but fatally hits the pedestrian.

Condition: negative outcome, linguistic AI adviser

Sofia is a high-level executive, and after a long day of work drives back home. Her new car is equipped with an artificial intelligence (AI) for driving assistance. The AI can monitor all physical objects around the car, including hidden ones, thanks to a 360 degrees laser radar. When activated, the system warns Sofia verbally about any possible imminent danger and recommends the ideal evasive manoeuvre such as 'There is a car braking on the right. To avoid the crash, swerve to the left!' On the way back home, Sofia drives alone and respects the speed limit. As Sofia enters an urban area, she decides to switch on the AI driving assistance. Further down the road and out of Sofia's sight, a pedestrian crosses the street. Driving around a corner, Sofia suddenly hears the AI's voice warning her of a pedestrian a short distance ahead and recommending her to swerve to the left. Sofia follows the advice of the AI but fatally hits the pedestrian.

Condition: negative outcome, no AI adviser

Alex works as a craftsman and is currently on his way to his next job. To carry all of his tools, Alex uses a traditional pick-up truck without any additional driving assistance. On his way to his next client, Alex drives alone and respects the speed limit. Further down the road and out of Alex's sight, a pedestrian crosses the street. Driving around a corner, Alex suddenly sees that a pedestrian is crossing the street a short distance ahead. At the last moment, Alex swerves but fatally hits the pedestrian.

Condition: positive outcome, sensory AI adviser
Anna, a young student, is on her way to see her family. She uses a car-sharing provider who offers access to the latest state-of-the-art car equipped with an artificial intelligence (AI) for driving assistance. The AI can monitor all physical objects around the car, including hidden ones, thanks to a 360 degrees laser radar. When activated, the system produces a range of different alarm sounds to warn Anna of an imminent obstacle on the road and makes the wheel vibrate, on either the left or the right, to indicate which side to swerve. On the way to her family, Anna drives alone and respects the speed limit. As Anna enters an urban area, she decides to switch on the AI driving assistance. Further down the road and out of Anna's sight, a pedestrian crosses the street. Driving around a corner, Anna suddenly hears an alarm, indicating a pedestrian a short distance ahead, and feels the vibrations of the steering wheel, recommending her to swerve to the left. Anna follows the advice of the AI and avoids a fatal crash with the pedestrian.

Condition: positive outcome, linguistic AI adviser
Marcus is on holiday, exploring the south of France. As a car enthusiast, he rented a state-of-the-art car equipped with an artificial intelligence (AI) for driving assistance. The AI can monitor all physical objects around the car, including hidden ones, thanks to a 360 degrees laser radar. When activated, the system warns Marcus verbally about any possible imminent danger and recommends the ideal evasive manoeuvre such as 'There is a car braking on the right. To avoid the crash, swerve to the left!' On the way to the Montpellier, Marcus drives alone and respects the speed limit. As Marcus enters an urban area, he decides to switch on the AI driving assistance. Further down the road and out of Marcus' sight, a pedestrian crosses the street. Driving around a corner, Marcus suddenly hears the AI's voice warning him of a pedestrian a short distance ahead and recommending him to swerve to the left. Marcus follows the advice of the AI and avoids a fatal crash with the pedestrian.

Condition: positive outcome, no AI adviser
Zoe is on her way to meet her friends for a night out at the local theatre. Zoe drives a traditional sports-car without any additional driving assistance. On her way to the theatre, Zoe drives alone and respects the speed limit. Further down the road and out of Zoe's sight, a pedestrian crosses the street. Driving around a corner, Zoe suddenly sees that a pedestrian is crossing the street a short distance ahead. At the last moment, Zoe swerves and avoids a fatal crash with the pedestrian.

Measurements
See pilot 3.

A.5 Supplementary for Pilot 5

Material

Condition: negative outcome, linguistic AI adviser
Marcus' car is equipped with a verbal driving assistant powered by artificial intelligence (AI). The verbal AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives verbal warnings and tells the driver what to do – saying things like 'Obstacle ahead! Swerve LEFT!' One day, Marcus is driving alone within the speed limit when he reaches a turn (see roadmap). Out of Marcus' sight, a pedestrian is crossing at a Zebra crossing. Driving around the corner, the verbal AI assistant warns Marcus and advises him to quickly swerve left with short verbal instructions. Marcus follows the advice of the verbal AI assistant. The pedestrian is nevertheless hit by the car and dies.

Condition: negative outcome, sensory AI adviser
Marcus' car is equipped with a sensory driving assistant powered by artificial intelligence (AI). The sensory AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives sound warnings and tells the driver what to do by vibrating the steering wheel, on either the left or the right, to indicate which side to swerve. One day, Marcus is driving alone within the speed limit when he reaches a turn (see roadmap). Out of Marcus' sight, a pedestrian is crossing at a Zebra crossing. Driving around the corner, the sensory AI assistant warns Marcus and advises him to quickly swerve left with short alarm sounds and steering wheel vibrations. Marcus follows the advice of the sensory AI assistant. The pedestrian is nevertheless hit by the car, and dies.

Condition: negative outcome, no AI adviser
Marcus has a brand new car with all standard technologies. One day, Marcus is driving alone within the speed limit when he reaches a turn (see roadmap). Out of Marcus' sight, a pedestrian is crossing at a Zebra crossing. Driving around the corner, Marcus suddenly sees the pedestrian a short distance ahead. Marcus follows his instincts and swerves left. The pedestrian is nevertheless hit by the car, and dies.

Measurements

All responses were recorded on a 100-point scale using a slider. The scale was anchored using labels.

Blame How much blame does Marcus deserve for the accident? How much blame does the [sensory linguistic] AI assistant deserve for the accident? How much blame does the pedestrian deserve for the accident?

Causal Responsibility To what extent did Marcus cause the accident? To what extent did the sensory AI assistant cause the accident? To what extent did the pedestrian cause the accident?

Informativity The AI assistant provided [sensory linguistic] advice to swerve left. How informative do you think this [sensory linguistic] advice was?

Effort How easy do you think it was for Marcus to pick the AI assistant's [sensory linguistic] advice?

A.6 Supplementary for Pilot 6

Material

See pilot 5.

A.7 Supplementary for Pilot 7

Material

Condition: active, sensory AI adviser, negative outcome

Alex is driving a brand-new car. It is equipped with an expert-level sensory driving assistant powered by artificial intelligence (AI). The sensory AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives tactile warnings and tells the driver what to do by vibrating the steering wheel, on either the left or the right, to indicate which side to swerve. One day, Alex is driving down a road. There is a STOP sign, but it is foggy, and the visibility is very bad. The sensory AI assistant warns Alex of the danger ahead by vibrating the steering wheel. Alex decides to follow the advice of the AI and brakes. Nevertheless, Alex crashes into another car that had priority at that crossing.

Condition: inactive, sensory AI adviser, negative outcome

Alex is driving a brand-new car. It is equipped with an expert-level sensory driving assistant powered by artificial intelligence (AI). The sensory AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives tactile warnings and tells the driver what to do by vibrating the steering wheel, on either the left or the right, to indicate which side to swerve. However, due to an electrical wiring problem, the AI assistant is not available for the next drive. The next day, Alex is driving down a road. Alex knows that he drives alone – without an AI assistant. There is a STOP sign, but it is foggy, and the visibility is very bad. Alex decides to follow his instincts and brakes. Nevertheless, Alex crashes into another car that had priority at that crossing.

Condition: active, linguistic AI adviser, negative outcome

Alex is driving a brand-new car. It is equipped with an expert-level verbal driving assistant powered by artificial intelligence (AI). The verbal AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives verbal warnings and tells the driver what to do – saying things like 'Obstacle ahead! Swerve LEFT!' One day, Alex is driving down a road. There is a STOP sign, but it is foggy, and the visibility is very bad. The verbal AI assistant warns Alex of the danger ahead with short verbal instructions. Alex decides to follow the advice of the AI and brakes. Nevertheless, Alex crashes into another car that had priority at that crossing.

Condition: inactive, linguistic AI adviser, negative outcome

Alex is driving a brand-new car. It is equipped with an expert-level verbal driving assistant powered by artificial intelligence (AI). The verbal AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives verbal warnings and tells the driver what to do – saying things like 'Obstacle ahead! Swerve LEFT!' However, due to an electrical wiring problem, the AI assistant is not available for the next drive. The next day, Alex is driving down a road. Alex knows that he drives alone – without an AI assistant. There is a STOP sign, but it is foggy, and the visibility is very bad. Alex decides to follow his instincts and brakes. Nevertheless, Alex crashes into another car that had priority at that crossing.

Measurements

All responses were recorded on a 200-point scale using a slider. The scale was anchored using labels ranging from 'completely disagree' to 'completely agree.'

Blame Alex deserves blame for the accident. The [sensory linguistic] AI assistant deserves blame for the accident.

Responsibility Alex is responsible for the accident. The [sensory linguistic] AI assistant is responsible for the accident.

Causal Responsibility Alex caused the accident. The [sensory linguistic] AI assistant caused the accident.

Counterfactual Capacity Alex had the capacity to avoid the accident. The [sensory linguistic] AI assistant had the capacity to avoid the accident.

A.8 Supplementary for Pilot 8

Material

Condition: active, sensory AI adviser, positive outcome
Alex is driving a brand-new car. It is equipped with an expert-level sensory driving assistant powered by artificial intelligence (AI). The sensory AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives tactile warnings and tells the driver what to do by vibrating the steering wheel, on either the left or the right, to indicate which side to swerve. One day, Alex is driving down a road. There is a STOP sign, but it is foggy, and the visibility is very bad. The sensory AI assistant warns Alex of the danger ahead by vibrating the steering wheel. Alex decides to follow the advice of the AI and brakes. As a consequence, Alex avoids a crash with another car that had priority at that crossing.

Condition: active, linguistic AI adviser, positive outcome
Alex is driving a brand-new car. It is equipped with an expert-level verbal driving assistant powered by artificial intelligence (AI). The verbal AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives verbal warnings and tells the driver what to do – saying things like 'Obstacle ahead! Swerve LEFT!' One day, Alex is driving down a road. There is a STOP sign, but it is foggy, and the visibility is very bad. The verbal AI assistant warns Alex of the danger ahead with short verbal instructions. Alex decides to follow the advice of the AI and brakes. As a consequence, Alex avoids a crash with another car that had priority at that crossing.

Measurements
All responses were recorded on a 200-point scale using a slider. The scale was anchored using labels ranging from 'completely disagree' to 'completely agree.'

Praise Alex deserves praise for the accident. The [sensory linguistic] AI assistant deserves praise for the accident.

Responsibility Alex is responsible for the accident. The [sensory linguistic] AI assistant is responsible for the accident.

Causal Responsibility Alex caused the accident. The [sensory linguistic] AI assistant caused the accident.

Counterfactual Capacity Alex had the capacity to avoid the accident. The [sensory linguistic] AI assistant had the capacity to avoid the accident.

A.9 Supplementary for Main Experiment

Material

Condition: active, linguistic AI adviser, positive outcome
Alex is driving a brand-new car. It is equipped with an expert-level verbal driving assistant powered by artificial intelligence (AI). The verbal AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives verbal warnings and tells the driver what to do – saying things like 'Obstacle ahead! Swerve LEFT!' One day, Alex is driving down a road. There is a STOP sign, but it is foggy, and the visibility is very bad. The verbal AI assistant warns Alex of the danger ahead with short verbal instructions. Alex decides to follow the advice of the AI and brakes. As a consequence, Alex avoids a crash with another car that had priority at that crossing.

Condition: active, sensory AI adviser, positive outcome
Alex is driving a brand-new car. It is equipped with an expert-level sensory driving assistant powered by artificial intelligence (AI). The sensory AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives tactile warnings and tells the driver what to do by vibrating the steering wheel, on either the left or the right, to indicate which side to swerve. One day, Alex is driving down a road. There is a STOP sign, but it is foggy, and the visibility is very bad. The sensory AI assistant warns Alex of the danger ahead by vibrating the steering wheel. Alex decides to follow the advice of the AI and brakes. As a consequence, Alex avoids a crash with another car that had priority at that crossing.

Condition: active, linguistic AI adviser, negative outcome
Alex is driving a brand-new car. It is equipped with an expert-level verbal driving assistant powered by artificial intelligence (AI). The verbal AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives verbal warnings and tells the driver what to do – saying things like 'Obstacle ahead! Swerve LEFT!' One day, Alex is driving down a road. There is a STOP sign, but it is foggy, and the visibility is very bad. The verbal AI assistant warns Alex of the danger ahead with short verbal instructions. Alex decides to follow the advice of the AI and brakes. Nevertheless, Alex crashes into another car that had priority at that crossing.

Condition: active, sensory AI adviser, negative outcome
Alex is driving a brand-new car. It is equipped with an expert-level sensory driving assistant powered by artificial intelligence (AI). The sensory AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives tactile warnings and tells the driver what to do by vibrating the steering wheel, on either the left or the right, to indicate which side to swerve.

One day, Alex is driving down a road. There is a STOP sign, but it is foggy, and the visibility is very bad. The sensory AI assistant warns Alex of the danger ahead by vibrating the steering wheel. Alex decides to follow the advice of the AI and brakes. Nevertheless, Alex crashes into another car that had priority at that crossing.

Condition: inactive, linguistic AI adviser, positive outcome

Alex is driving a brand-new car. It is equipped with an expert-level verbal driving assistant powered by artificial intelligence (AI). The verbal AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives verbal warnings and tells the driver what to do – saying things like 'Obstacle ahead! Swerve LEFT!' However, due to an electrical wiring problem, the AI assistant is not available for the next drive. The next day, Alex is driving down a road. Alex knows that he drives alone – without an AI assistant. There is a STOP sign, but it is foggy, and the visibility is very bad. Alex decides to follow his instincts and brakes. As a consequence, Alex avoids a crash with another car that had priority at that crossing.

Condition: inactive, sensory AI adviser, positive outcome

Alex is driving a brand-new car. It is equipped with an expert-level sensory driving assistant powered by artificial intelligence (AI). The sensory AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives tactile warnings and tells the driver what to do by vibrating the steering wheel, on either the left or the right, to indicate which side to swerve. However, due to an electrical wiring problem, the AI assistant is not available for the next drive. The next day, Alex is driving down a road. Alex knows that he drives alone – without an AI assistant. There is a STOP sign, but it is foggy, and the visibility is very bad. Alex decides to follow his instincts and brakes. As a consequence, Alex avoids a crash with another car that had priority at that crossing.

Condition: inactive, linguistic AI adviser, negative outcome

Alex is driving a brand-new car. It is equipped with an expert-level verbal driving assistant powered by artificial intelligence (AI). The verbal AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives verbal warnings and tells the driver what to do – saying things like 'Obstacle ahead! Swerve LEFT!' However, due to an electrical wiring problem, the AI assistant is not available for the next drive. The next day, Alex is driving down a road. Alex knows that he drives alone – without an AI assistant. There is a STOP sign, but it is foggy, and the visibility is very bad. Alex decides to follow his instincts and brakes. Nevertheless, Alex crashes into another car that had priority at that crossing.

Condition: inactive, sensory AI adviser, negative outcome

Alex is driving a brand-new car. It is equipped with an expert-level sensory driving assistant powered by artificial intelligence (AI). The sensory AI assistant monitors the space around the car with a 360-degrees radar and identifies possible dangers. When it does, it gives tactile warnings and tells the driver what to do by vibrating the steering wheel, on either the left or the right, to indicate which side to swerve. However, due to an electrical wiring problem, the AI assistant is not available for the next drive. The next day, Alex is driving down a road. Alex knows that he drives alone – without an AI assistant. There is a STOP sign, but it is foggy, and the visibility is very bad. Alex decides to follow his instincts and brakes. Nevertheless, Alex crashes into another car that had priority at that crossing.

Measurements

Blame/Praise Alex deserves [blame for the praise for preventing an] accident. The [sensory linguistic] AI assistant deserves [blame praise] for the accident.

Responsibility Alex is responsible for (avoiding an the) accident. The [sensory linguistic] AI assistant is responsible for (avoiding an the) accident.

Causal Responsibility Alex [caused the prevented an] accident. The [sensory linguistic] AI assistant [caused the prevented an] accident.

Counterfactual Capacity Alex had the capacity to [cause an avoid the] accident. The [sensory linguistic] AI assistant had the capacity to [cause an avoid the] accident.

Toolness The [sensory linguistic] AI assistant is a tool.

A.10 Supplementary for Follow-up Experiment

Material

Condition: active tool, negative outcome

Alex is driving a brand-new car. It is equipped with state-of-the-art fog lights. The fog lights are extremely bright and enable Alex to see through any potential fog. The lights are in great condition and work very well. One day, Alex is driving down a road. There is a STOP sign, but it is foggy, and the visibility is very bad. The fog lights highlight the outline of an approaching car. Alex sees the car's outline and decides to brake. Nevertheless, Alex crashes into the car that had priority at that crossing.

Condition: inactive tool, negative outcome

Alex is driving a brand-new car. It is equipped with state-of-the-art fog lights. The fog lights are extremely bright and enable Alex to see through any potential fog. However, due to an electrical wiring problem, the fog lights are not available for the next drive.

One day, Alex is driving down a road. Alex knows that he is driving without any fog lights. There is a STOP sign, but it is foggy, and the visibility is very bad. Alex decides to follow his instincts and brakes. Nevertheless, Alex crashes into another car that had priority at that crossing.

Measurements

Blame Alex deserves blame for the accident. The fog lights deserve blame for the accident.

Responsibility Alex is responsible for the accident. The fog lights are responsible for the accident.

Causal Responsibility Alex caused the accident. The fog lights caused the accident.

Toolness The fog lights are a tool.

Counterfactual Capacity Alex had the capacity to avoid the accident.

References

Adams, Fred, and Ken Aizawa. 2001. “The Bounds of Cognition.” Philosophical Psychology 14 (1): 43–64. https://doi.org/10.1080/09515080120033571.

Adams, Frederick, and Kenneth Aizawa. 2010. “The Value of Cognitivism in Thinking about Extended Cognition.” Phenomenology and the Cognitive Sciences 9 (4): 579–603. https://doi.org/10.1007/s11097-010-9184-9.

Admoni, Henny, and Brian Scassellati. 2017. “Social Eye Gaze in Human-Robot Interaction: A Review.” Journal of Human-Robot Interaction 6 (1): 25. https://doi.org/10.5898/JHRI.6.1.Admoni.

Akata, Zeynep, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, et al. 2020. “A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence.” Computer 53 (8): 18–28. https://doi.org/10.1109/MC.2020.2996587.

Al Moubayed, Samer, and Gabriel Skantze. 2012. “Perception of Gaze Direction for Situated Interaction.” In 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, Gaze-In 2012; Santa Monica, CA; 26 October 2012 Through 26 October 2012. ACM.

Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. “Understanding of a Convolutional Neural Network.” In 2017 International Conference on Engineering and Technology (ICET), 1–6. Antalya: IEEE. https://doi.org/10.1109/ICEngTechnol.2017.8308186.

Amedi, Amir, William M. Stern, Joan A. Camprodon, Felix Bermpohl, Lotfi Merabet, Stephen Rotman, Christopher Hemond, Peter Meijer, and Alvaro Pascual-Leone. 2007. “Shape Conveyed by Visual-to-Auditory Sensory Substitution Activates the Lateral Occipital Complex.” Nature Neuroscience 10 (6): 687–89. https://doi.org/10.1038/nn1912.

Anderson, Michael, and Susan Leigh Anderson. 2011. “Machine Ethics.” In Machine Ethics, 9780521112:1–538. https://doi.org/10.1017/CBO9780511978036.

Anderson, Rajen A., Molly J. Crockett, and David A. Pizarro. 2020. “A Theory of Moral Praise.” Trends in Cognitive Sciences 24 (9): 694–703. https://doi.org/10.1016/j.tics.2020.06.008.

Anscombe, Gertrude Elizabeth Margaret. 2000. Intention. Harvard University Press.

Argall, Brenna D., Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. “A Survey of Robot Learning from Demonstration.” Robotics and Autonomous Systems 57 (5): 469–83. https://doi.org/10.1016/j.robot.2008.10.024.

Argyle, Michael, and Mark Cook. 1976. Gaze and Mutual Gaze. Gaze and Mutual Gaze. Oxford, England: Cambridge U Press.

Arnold, Gabriel, and Malika Auvray. 2018. "Tactile Recognition of Visual Stimuli: Specificity Versus Generalization of Perceptual Learning." Vision Research 152 (November): 40–50. https://doi.org/10.1016/j.visres.2017.11.007.

Asaro, Peter M. 2006. "What Should We Want from a Robot Ethic." International Review of Information Ethics 6 (12): 9–16.

Atmaca, Silke, Natalie Sebanz, and Günther Knoblich. 2011. "The Joint Flanker Effect: Sharing Tasks with Real and Imagined Co-Actors." Experimental Brain Research 211 (3-4): 371–85. https://doi.org/10.1007/s00221-011-2709-9.

Auvray, Malika, Sylvain Hanneton, Charles Lenay, and Kevin O'Regan. 2005. "There Is Something Out There: Distal Attributtion in Sensory Substitution, Twenty Years Later." Journal of Integrative Neuroscience 4 (4): 505–21. https://doi.org/10.1142/S0219635205001002.

Auvray, Malika, Sylvain Hanneton, and J Kevin O'Regan. 2007. "Learning to Perceive with a Visuo Auditory Substitution System: Localisation and Object Recognition with 'The Voice'." Perception 36 (3): 416–30. https://doi.org/10.1068/p5631.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. "The Moral Machine Experiment." Nature 563 (7729): 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Awad, Edmond, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Joshua B. Tenenbaum, Azim Shariff, Jean François Bonnefon, and Iyad Rahwan. 2019. "Drivers Are Blamed More Than Their Automated Cars When Both Make Mistakes." Nature Human Behaviour. https://doi.org/10.1038/s41562-019-0762-8.

Azarian, Bobby, George A. Buzzell, Elizabeth G. Esser, Alexander Dornstauder, and Matthew S. Peterson. 2017. "Averted Body Postures Facilitate Orienting of the Eyes." Acta Psychologica 175 (April): 28–32. https://doi.org/10.1016/j.actpsy.2017.02.006.

Bach-Y-Rita, Paul, Carter C. Collins, Frank A. Saunders, Benjamin White, and Lawrence Scadden. 1969. "Vision Substitution by Tactile Image Projection." Nature 221 (5184): 963–64. https://doi.org/10.1038/221963a0.

Bach-y-Rita, Paul, and Stephen W. Kercel. 2003. "Sensory Substitution and the Humanmachine Interface." Trends in Cognitive Sciences 7 (12): 541–46. https://doi.org/10.1016/j.tics.2003.10.013.

Bago, Bence. 2022. "Situational Factors Shape Moral Judgements in the Trolley Dilemma in Eastern, Southern and Western Countries in a Culturally Diverse Sample." Nature Human Behaviour, 25. https://doi.org/10.1038/s41562-022-01319-5.

Bang, Dan, Sara Ershadmanesh, Hamed Nili, and Stephen M Fleming. 2020. "Privatepublic Mappings in Human Prefrontal Cortex." eLife 9 (July): e56477. https://doi.org/10.7554/eLife.56477.

Bansal, Trapit, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. 2018. "Emergent Complexity via Multi-Agent Competition." arXiv. https://doi.org/10.48550/arXiv.1710.03748.

Barandiaran, Xabier E., Ezequiel Di Paolo, and Marieke Rohde. 2009. "Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action." Adaptive Behavior 17 (5): 367–86. https://doi.org/10.1177/1059712309343819.

Baron, Jonathan, and John C Hershey. 1988. "Outcome Bias in Decision Evaluation." Journal of Personality and Social Psychology 54 (4): 11.

Bartneck, Christoph, Juliane Reichenbach, and Julie Carpenter. 2006. "Use of Praise and Punishment in Human-Robot Collaborative Teams." In ROMAN 2006 – The 15th IEEE International Symposium on Robot and Human Interactive Communication, 177–82. https://doi.org/10.1109/ROMAN.2006.314414.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." Journal of Statistical Software 67 (1). https://doi.org/10.18637/jss.v067.i01.

Battich, Lucas, Isabelle Garzorz, Basil Wahn, and Ophelia Deroy. 2021. "The Impact of Joint Attention on the Sound-Induced Flash Illusions." Attention, Perception, & Psychophysics 83 (8): 3056–68. https://doi.org/10.3758/s13414-021-02347-5.

Battich, Lucas, and Bart Geurts. 2020. "Joint Attention and Perceptual Experience." Synthese, March. https://doi.org/10.1007/s11229-020-02602-6.

Bayliss, A, A Frischen, M Fenske, and S Tipper. 2007. "Affective Evaluations of Objects Are Influenced by Observed Gaze Direction and Emotional Expression." Cognition 104 (3): 644–53. https://doi.org/10.1016/j.cognition.2006.07.012.

Bayliss, Andrew P., Matthew A. Paul, Peter R. Cannon, and Steven P. Tipper. 2006. "Gaze Cuing and Affective Judgments of Objects: I Like What You Look At." Psychonomic Bulletin & Review 13 (6): 1061–66. https://doi.org/10.3758/BF03213926.

Beer, Randall D. 1995. "A Dynamical Systems Perspective on Agent-Environment Interaction." Artificial Intelligence 72 (1-2): 173–215.

Ben-Artzi, Elisheva, and Lawrence E. Marks. 1995. "Visual-Auditory Interaction in Speeded Classification: Role of Stimulus Difference." Perception & Psychophysics 57 (8): 1151–62. https://doi.org/10.3758/BF03208371.

Beyer, Frederike, Nura Sidarus, Sofia Bonicalzi, and Patrick Haggard. 2017. "Beyond Self-Serving Bias: Diffusion of Responsibility Reduces Sense of Agency and Outcome Monitoring." Social Cognitive and Affective Neuroscience 12 (1): 138–45. https://doi.org/10.1093/scan/nsw160.

Bharadwaj, Anandhi, Omar A. El Sawy, Paul A. Pavlou, and N. Venkatraman. 2013. "Digital Business Strategy: Toward a Next Generation of Insights." MIS Quarterly 37 (2): 471–82. https://www.jstor.org/stable/43825919.

Bigman, Yochanan E., and Kurt Gray. 2018. "People Are Averse to Machines Making Moral Decisions." Cognition 181 (December): 21–34. https://doi.org/10.1016/j.cognition.2018.08.003.

Bigman, Yochanan E., Adam Waytz, Ron Alterovitz, and Kurt Gray. 2019. “Holding Robots Responsible: The Elements of Machine Morality.” Trends in Cognitive Sciences 23 (5): 365–68. https://doi.org/10.1016/j.tics.2019.02.008.

Boerman, Sophie C., Sanne Kruikemeier, and Frederik J. Zuiderveen Borgesius. 2017. “Online Behavioral Advertising: A Literature Review and Research Agenda.” Journal of Advertising 46 (3): 363–76. https://doi.org/10.1080/00913367.2017.1339368.

Bolotta, Samuele, and Guillaume Dumas. 2022. “Social Neuro AI: Social Interaction as the “Dark Matter” of AI.” arXiv. https://arxiv.org/abs/arXiv:2112.15459.

Bondi, Elizabeth, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. 2021. “Role of Human-AI Interaction in Selective Prediction.” arXiv:2112.06751 [Cs], December. https://arxiv.org/abs/2112.06751.

Böckler, Anne, and Jan Zwickel. 2013. “Influences of Spontaneous Perspective Taking on Spatial and Identity Processing of Faces.” Social Cognitive and Affective Neuroscience 8 (7): 735–40. https://doi.org/10.1093/scan/nss061.

Bratman, Michael. 1984. “Two Faces of Intention.” The Philosophical Review 93 (3): 375–405.

———. 1999. Intention, Plans, and Practical Reason. David Hume Series. Stanford, Calif: Center for the Study of Language and Information.

Bratman, Michael E. 2007. Structures of Agency: Essays. Oxford University Press.

Bratman, Michael, and U. G. and Abbie Birch Durfee Professor in the School of Humanities and Sciences and Professor of Philosophy Michael Bratman. 1987. Intention, Plans, and Practical Reason. Harvard University Press.

Breazeal, C., C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. 2005. “Effects of Nonverbal Communication on Efficiency and Robustness in Human-Robot Teamwork.” In 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 708–13. https://doi.org/10.1109/IROS.2005.1545011.

Breazeal, C., and B. Scassellati. 1999. “How to Build Robots That Make Friends and Influence People.” In Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289), 2:858–863 vol.2. https://doi.org/10.1109/IROS.1999.812787.

Brooks, Rodney A., Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew M. Williamson. 1999. “The Cog Project: Building a Humanoid Robot.” In Computation for Metaphors, Analogy, and Agents, edited by Chrystopher L. Nehaniv, 52–87. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-48834-0_5.

Bruner, J. S. 1974. “From Communication to Languagea Psychological Perspective.” Cognition 3 (3): 255–87. https://doi.org/10.1016/0010-0277(74)90012-2.

Butterfill, Stephen A, and Corrado Sinigaglia. 2022. "Towards a Mechanistically Neutral Account of Acting Jointly: The Notion of a Collective Goal." Mind, February, fzab096. https://doi.org/10.1093/mind/fzab096.

Campbell, J. 2018. "Joint Attention." In The Routledge Handbook of Collective Intentionality, edited by M. Jankovic and K. Ludwig, 115–29. New York, NY: Routledge.

Campbell, John. 2011. "An Object-Dependent Perspective on Joint Attention." In Joint Attention: New Developments in Philosophy, Psychology and Neuroscience, edited by Axel Seemann. The MIT Press.

Canbek, Nil Goksel, and Mehmet Emin Mutlu. 2016. "On the Track of Artificial Intelligence: Learning with Intelligent Personal Assistants." Journal of Human Sciences 13 (1): 592–601.

Capozzi, Francesca, and Jelena Ristic. 2018. "How Attention Gates Social Interactions." Annals of the New York Academy of Sciences 1426 (1): 179–98. https://doi.org/10.1111/nyas.13854.

Caraiman, Simona, Anca Morar, Mateusz Owczarek, Adrian Burlacu, Dariusz Rzeszotarski, Nicolae Botezatu, Paul Herghelegiu, Florica Moldoveanu, Pawel Strumillo, and Alin Moldoveanu. 2017. "Computer Vision for the Visually Impaired: The Sound of Vision System." In Proceedings of the IEEE International Conference on Computer Vision Workshops, 1480–89.

Carpenter, Malinda, and Kristin Liebal. 2011. "Joint Attention, Communication, and Knowing Together in Infancy." In Joint Attention: New Developments in Psychology, Philosophy of Mind, and Social Neuroscience, 159–81. Cambridge, MA, US: MIT Press.

Carter, Shan, and Michael Nielsen. 2017. "Using Artificial Intelligence to Augment Human Intelligence." Distill 2 (12): e9. https://doi.org/10.23915/distill.00009.

Castiello, Umberto. 2003. "Understanding Other People's Actions: Intention and Attention." Journal of Experimental Psychology: Human Perception and Performance 29 (2): 416–30. https://doi.org/10.1037/0096-1523.29.2.416.

Cervantes, José-Antonio, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. 2020. "Artificial Moral Agents: A Survey of the Current Status." Science and Engineering Ethics 26 (2): 501–32. https://doi.org/10.1007/s11948-019-00151-x.

Chalmers, David J. 2011. "A Computational Foundation for the Study of Cognition." Journal of Cognitive Science 12 (4): 323–57.

Chandel, Himanshu Singh, and Sonia Vatta. 2015. "Occlusion Detection and Handling: A Review." https://doi.org/10.5120/21264-3857.

Chen, Lele, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. 2018. "Lip Movements Generation at a Glance." In Computer Vision ECCV 2018, edited by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, 11211:538–53. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-01234-2_32.

Chen, Min, Yujun Ma, Jeungeun Song, Chin-Feng Lai, and Bin Hu. 2016. "Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring." Mobile Networks and Applications 21 (5): 825–45. https://doi.org/10.1007/s11036-016-0745-1.

Chevalier, Pauline, Kyveli Kompatsiari, Francesca Ciardo, and Agnieszka Wykowska. 2020. "Examining Joint Attention with the Use of Humanoid Robots-A New Approach to Study Fundamental Mechanisms of Social Cognition." Psychonomic Bulletin & Review 27 (2): 217–36. https://doi.org/10.3758/s13423-019-01689-4.

Chérif, Emna, and Jean-François Lemoine. 2019. "Anthropomorphic Virtual Assistants and the Reactions of Internet Users: An Experiment on the Assistant's Voice." Recherche Et Applications En Marketing (English Edition) 34 (1): 28–47. https://doi.org/10.1177/2051570719829432.

Chockler, Hana, and Joseph Y Halpern. 2004. "Responsibility and Blame: A Structural-Model Approach." Journal of Artificial Intelligence Research 22: 93–115.

Christensen, WD, and CA Hooker. 2000. "Autonomy and the Emergence of Intelligence: Organised Interactive Construction." Communication and Cognition-Artificial Intelligence 17 (3-4): 133–57.

Christensen-Szalanski, Jay J. J, and Cynthia Fobian Willham. 1991. "The Hindsight Bias: A Meta-Analysis." Organizational Behavior and Human Decision Processes 48 (1): 147–68. https://doi.org/10.1016/0749-5978(91)90010-Q.

Cinel, Caterina, Davide Valeriani, and Riccardo Poli. 2019. "Neurotechnologies for Human Cognitive Augmentation: Current State of the Art and Future Prospects." Frontiers in Human Neuroscience 13.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." Analysis 58 (1): 7–19. https://doi.org/10.1111/1467-8284.00096.

Clark, Herbert H., and Kerstin Fischer. 2022. "Social Robots as Depictions of Social Agents." Behavioral and Brain Sciences, March, 1–33. https://doi.org/10.1017/S0140525X22000668.

Coeckelbergh, Mark. 2009. "Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics." International Journal of Social Robotics 1 (3): 217–21.

———. 2016. "Responsibility and the Moral Phenomenology of Using Self-Driving Cars." Applied Artificial Intelligence 30 (8): 748–57. https://doi.org/10.1080/08839514.2016.1229759.

———. 2020. "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability." Science and Engineering Ethics 26 (4): 2051–68. https://doi.org/10.1007/s11948-019-00146-8.

Cohen, Jonathan. 2018. "Sensory Subsitution and Perceptual Emergence." In Sensory Substitution and Augmentation, 219:205–35. Proceedings of the British Academy. The British Academy.

Cohen, Leonardo G., Pablo Celnik, Alvaro Pascual-Leone, Brian Corwell, Lala Faiz, James Dambrosia, Manabu Honda, et al. 1997. "Functional Relevance of Cross-Modal Plasticity in Blind Humans." Nature 389 (6647): 180–83. https://doi.org/10.1038/38278.

Collignon, Olivier, Patrice Voss, Maryse Lassonde, and Franco Lepore. 2008. "Cross-Modal Plasticity for the Spatial Processing of Sounds in Visually Deprived Subjects." Experimental Brain Research 192 (3): 343. https://doi.org/10.1007/s00221-008-1553-z.

Constantinescu, Mihaela, Constantin Vică, Radu Uszkai, and Cristina Voinea. 2022. "Blame It on the AI? On the Moral Responsibility of Artificial Moral Advisors." Philosophy & Technology 35 (2): 35. https://doi.org/10.1007/s13347-022-00529-z.

Copeland, B. Jack, and Oron Shagrir. 2020. "Physical Computability Theses." In Quantum, Probability, Logic: The Work and Influence of Itamar Pitowsky, edited by Meir Hemmo and Orly Shenker, 217–31. Jerusalem Studies in Philosophy and History of Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-34316-3_9.

Cross, Emily S., and Richard Ramsey. 2021. "Mind Meets Machine: Towards a Cognitive Science of Human." Trends in Cognitive Sciences 25 (3): 200–212. https://doi.org/10.1016/j.tics.2020.11.009.

Cushman, Fiery. 2008. "Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment." Cognition 108 (2): 353–80. https://doi.org/10.1016/j.cognition.2008.03.006.

———. 2015. "Deconstructing Intent to Reconstruct Morality." Current Opinion in Psychology 6 (December): 97–103. https://doi.org/10.1016/j.copsyc.2015.06.003.

Dalal, Reeshad S., and Silvia Bonaccio. 2010. "What Types of Advice Do Decision-Makers Prefer?" Organizational Behavior and Human Decision Processes 112 (1): 11–23. https://doi.org/10.1016/j.obhdp.2009.11.007.

Danaher, John. 2018. "Toward an Ethics of AI Assistants: An Initial Framework." Philosophy and Technology 31 (4): 629–53. https://doi.org/10.1007/s13347-018-0317-3.

Darley, John M., and Bibb Latane. 1968. "Bystander Intervention in Emergencies: Diffusion of Responsibility." Journal of Personality and Social Psychology 8 (4, Pt.1): 377–83. https://doi.org/10.1037/h0025589.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." The Journal of Philosophy 60 (23): 685–700.

———. 1982. "Rational Animals." Dialectica 36 (4): 317–27.

———. 2001. Essays on Actions and Events: Philosophical Essays. Vol. 1. Oxford University Press on Demand.

Dąbrowska, Justyna, Argyro Almpanopoulou, Alexander Brem, Henry Chesbrough, Valentina Cucino, Alberto Di Minin, Ferran Giones, et al. 2022. "Digital Transformation, for Better or Worse: A Critical Multi-Level Research Agenda." R&D Management 52 (5): 930–54. https://doi.org/10.1111/radm.12531.

Delaunay, Frederic, Joachim de Greeff, and Tony Belpaeme. 2010. "A Study of a Retro-Projected Robotic Face and Its Effectiveness for Gaze Reading by Humans."

della Gatta, Francesco, Francesca Garbarini, Marco Rabuffetti, Luca Viganò, Stephen A. Butterfill, and Corrado Sinigaglia. 2017. "Drawn Together: When Motor Representations Ground Joint Actions." Cognition 165 (August): 53–60. https://doi.org/10.1016/j.cognition.2017.04.008.

Delon, Nicolas. 2018. "Animal Agency, Captivity, and Meaning." The Harvard Review of Philosophy.

Dennett, Daniel C. 1971. "Intentional Systems." The Journal of Philosophy 68 (4): 87–106.

———. 1976. "Conditions of Personhood." In The Identities of Persons, edited by Amélie Oksenberg Rorty, 175–96. University of California Press.

Dennett, Daniel C, and John Haugeland. 1987. "Intentionality." In The Oxford Companion to the Mind. Oxford University Press.

Dewey, Marc, and Uta Wilkens. 2019. "The Bionic Radiologist : Avoiding Blurry Pictures and Providing Greater Insights." Npj Digital Medicine 2 (1): 1–7. https://doi.org/10.1038/s41746-019-0142-9.

Dignum, Virginia. 2020. "Responsibility and Artificial Intelligence." In The Oxford Handbook of Ethics of AI, 213–31. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.12.

Dingler, Tilman, Passant El Agroudy, Huy Viet Le, Albrecht Schmidt, Evangelos Niforatos, Agon Bexheti, and Marc Langheinrich. 2016. "Multimedia Memory Cues for Augmenting Human Memory." IEEE MultiMedia 23 (2): 4–11. https://doi.org/10.1109/MMUL.2016.31.

Dodig-Crnkovic, Gordana, and Daniel Persson. 2008. "Sharing Moral Responsibility with Robots: A Pragmatic Approach." Frontiers in Artificial Intelligence and Applications 173: 165–68.

Dollar, Aaron M., and Hugh Herr. 2008. "Lower Extremity Exoskeletons and Active Orthoses: Challenges and State-of-the-Art." IEEE Transactions on Robotics 24 (1): 144–58. https://doi.org/10.1109/TRO.2008.915453.

Douer, Nir, and Joachim Meyer. 2020. "The Responsibility Quantification Model of Human Interaction With Automation." IEEE Transactions on Automation Science and Engineering 17 (2): 1044–60. https://doi.org/10.1109/TASE.2020.2965466.

Driver, Jon, Greg Davis, Paola Ricciardelli, Polly Kidd, Emma Maxwell, and Simon Baron-Cohen. 1999. "Gaze Perception Triggers Reflexive Visuospatial Orienting." Visual Cognition 6 (5): 509–40. https://doi.org/10.1080/135062899394920.

El Zein, Marwa, Bahador Bahrami, and Ralph Hertwig. 2019. "Shared Responsibility in Collective Decisions." Nature Human Behaviour 3 (6): 554–59. https://doi.org/10.1038/s41562-019-0596-4.

El Zein, Marwa, Ray J. Dolan, and Bahador Bahrami. 2022. "Shared Responsibility Decreases the Sense of Agency in the Human Brain." Journal of Cognitive Neuroscience, July, 1–17. https://doi.org/10.1162/jocn_a_01896.

Elli, Giulia V., Stefania Benetti, and Olivier Collignon. 2014. "Is There a Future for Sensory Substitution Outside Academic Laboratories?" Multisensory Research 27 (5-6): 271–91. https://doi.org/10.1163/22134808-00002460.

Enarsson, Therese, Lena Enqvist, and Markus Naarttijärvi. 2022. "Approaching the Human in the Loop Legal Perspectives on Hybrid Human/Algorithmic Decision-Making in Three Contexts." Information & Communications Technology Law 31 (1): 123–53. https://doi.org/10.1080/13600834.2021.1958860.

Eppe, Manfred, Matthias Kerzel, Erik Strahl, and Stefan Wermter. 2018. "Deep Neural Object Analysis by Interactive Auditory Exploration with a Humanoid Robot." In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 284–89. https://doi.org/10.1109/IROS.2018.8593838.

Eriksen, Barbara A., and Charles W. Eriksen. 1974. "Effects of Noise Letters Upon the Identification of a Target Letter in a Nonsearch Task." Perception & Psychophysics 16 (1): 143–49. https://doi.org/10.3758/BF03203267.

Fast, Ethan, and Eric Horvitz. 2017. "Long-Term Trends in the Public Perception of Artificial Intelligence." Proceedings of the AAAI Conference on Artificial Intelligence 31 (1). https://doi.org/10.1609/aaai.v31i1.10635.

Fernández-Caramés, Tiago, and Paula Fraga-Lamas. 2018. "Towards The Internet-of-Smart-Clothing: A Review on IoT Wearables and Garments for Creating Intelligent Connected E-Textiles." Electronics 7 (12): 405. https://doi.org/10.3390/electronics7120405.

Fiebich, Anika, Nhung Nguyen, and Sarah Schwarzkopf. 2015. "Cooperation with Robots? A Two-Dimensional Approach." In Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation, 25–43.

Flemisch, Frank, Matthias Heesen, Tobias Hesse, Johann Kelsch, Anna Schieben, and Johannes Beller. 2012. "Towards a Dynamic Balance Between Humans and Automation: Authority, Ability, Responsibility and Control in Shared and Cooperative Control Situations." Cognition, Technology & Work 14 (1): 3–18. https://doi.org/10.1007/s10111-011-0191-6.

Floridi, Luciano, and Jeff W Sanders. 2004. "On the Morality of Artificial Agents." Minds and Machines 14 (3): 349–79.

Fon, Vincy, and Francesco Parisi. 2003. "The Limits of Reciprocity for Social Cooperation." {{SSRN Scholarly Paper}}. Rochester, NY. https://doi.org/10.2139/ssrn.384589.

Forsyth, Donelson R., Linda E. Zyzniewski, and Cheryl A. Giammanco. 2002. "Responsibility Diffusion in Cooperative Collectives." Personality and Social Psychology Bulletin 28 (1): 54–65. https://doi.org/10.1177/0146167202281005.

Fossa, Fabio. 2018. "Artificial Moral Agents: Moral Mentors or Sensible Tools?" Ethics and Information Technology 20 (2): 115–26.

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." Journal of Philosophy 68 (1): 5–20. https://doi.org/10.2307/2024717.

Franklin, Matija, Edmond Awad, and David Lagnado. 2021. "Blaming Automated Vehicles in Difficult Situations." iScience 24 (4): 102252. https://doi.org/10.1016/j.isci.2021.102252.

Freundlieb, Martin, Natalie Sebanz, and Ágnes M. Kovács. 2017. "Out of Your Sight, Out of My Mind: Knowledge about Another Person's Visual Access Modulates Spontaneous Visuospatial Perspective-Taking." Journal of Experimental Psychology: Human Perception and Performance 43 (6): 1065–72. https://doi.org/10.1037/xhp0000379.

Friesen, Chris Kelland, and Alan Kingstone. 1998. "The Eyes Have It! Reflexive Orienting Is Triggered by Nonpredictive Gaze." Psychonomic Bulletin & Review 5 (3): 490–95. https://doi.org/10.3758/BF03208827.

Friesen, Chris Kelland, Jelena Ristic, and Alan Kingstone. 2004. "Attentional Effects of Counterpredictive Gaze and Arrow Cues." Journal of Experimental Psychology: Human Perception and Performance 30 (2): 319–29. https://doi.org/10.1037/0096-1523.30.2.319.

Frith, Chris. 2008. "Social Cognition." Philosophical Transactions of the Royal Society B: Biological Sciences 363 (1499): 2033–39. https://doi.org/10.1098/rstb.2008.0005.

Gallagher, Shaun. 2006. How the Body Shapes the Mind. Clarendon Press.

Gallese, Vittorio. 2007. "Before and Below 'Theory of Mind': Embodied Simulation and the Neural Correlates of Social Cognition." Philosophical Transactions of the Royal Society B: Biological Sciences 362 (1480): 659–69. https://doi.org/10.1098/rstb.2006.2002.

Gallotti, Mattia, and Chris D. Frith. 2013. "Social Cognition in the We-Mode." Trends in Cognitive Sciences 17 (4): 160–65. https://doi.org/10.1016/j.tics.2013.02.002.

Gazzola, V., G. Rizzolatti, B. Wicker, and C. Keysers. 2007. "The Anthropomorphic Brain: The Mirror Neuron System Responds to Human and Robotic Actions." NeuroImage 35 (4): 1674–84. https://doi.org/10.1016/j.neuroimage.2007.02.003.

Gerstenberg, Tobias, Tomer D. Ullman, Jonas Nagel, Max Kleiman-Weiner, David A. Lagnado, and Joshua B. Tenenbaum. 2018. "Lucky or Clever? From Expectations to Responsibility Judgments." Cognition 177 (August): 122–41. https://doi.org/10.1016/j.cognition.2018.03.019.

Ghahramani, Zoubin. 2015. "Probabilistic Machine Learning and Artificial Intelligence." Nature 521 (7553): 452–59. https://doi.org/10.1038/nature14541.

Gino, Francesca. 2008. "Do We Listen to Advice Just Because We Paid for It? The Impact of Advice Cost on Its Use." Organizational Behavior and Human Decision Processes 107 (2): 234–45. https://doi.org/10.1016/j.obhdp.2008.03.001.

Giubilini, Alberto, and Julian Savulescu. 2018. "The Artificial Moral Advisor. The 'Ideal Observer' Meets Artificial Intelligence." Philosophy & Technology 31 (2): 169–88. https://doi.org/10.1007/s13347-017-0285-z.

Goertzel, Ben, Julia Mossbridge, Eddie Monroe, David Hanson, and Gino Yu. 2017. "Humanoid Robots as Agents of Human Consciousness Expansion." arXiv Preprint arXiv:1709.07791. https://arxiv.org/abs/1709.07791.

Gogoll, Jan, and Matthias Uhl. 2018. "Rage Against the Machine: Automation in the Moral Domain." Journal of Behavioral and Experimental Economics 74 (June): 97–103. https://doi.org/10.1016/j.socec.2018.04.003.

Golub, Justin S., Leo Ling, Kaibao Nie, Amy Nowack, Sarah J. Shepherd, Steven M. Bierer, Elyse Jameyson, Chris R. S. Kaneko, James O. Phillips, and Jay T. Rubinstein. 2014. "Prosthetic Implantation of the Human Vestibular System." Otology & Neurotology : Official Publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology 35 (1): 136–47. https://doi.org/10.1097/MAO.0000000000000003.

Gómez, Juan-Carlos. 2005. "Joint Attention and the Notion of Subject: Insights from Apes, Normal Children, and Children with Autism." In Joint Attention: Communication and Other Minds: Issues in Philosophy and Psychology, 65–84. Consciousness and Self-Consciousness. New York, NY, US: Clarendon Press/Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199245635.003.0004.

Green, Ben, and Yiling Chen. 2019. "The Principles and Limits of Algorithm-in-the-Loop Decision Making." Proceedings of the ACM on Human-Computer Interaction 3 (CSCW): 50:1–24. https://doi.org/10.1145/3359152.

Greene, Joshua D., Fiery A. Cushman, Lisa E. Stewart, Kelly Lowenberg, Leigh E. Nystrom, and Jonathan D. Cohen. 2009. "Pushing Moral Buttons: The Interaction Between Personal Force and Intention in Moral Judgment." Cognition 111 (3): 364–71. https://doi.org/10.1016/j.cognition.2009.02.001.

Grynszpan, Ouriel, Aïsha Sahaï, Nasmeh Hamidi, Elisabeth Pacherie, Bruno Berberian, Lucas Roche, and Ludovic Saint-Bauzel. 2019. "The Sense of Agency in Human-Human Vs Human-Robot Joint Action." Consciousness and Cognition 75 (October): 102820. https://doi.org/10.1016/j.concog.2019.102820.

Guglielmo, Steve, and Bertram F. Malle. 2019. "Asymmetric Morality: Blame Is More Differentiated and More Extreme Than Praise." Edited by Valerio Capraro. PLOS ONE 14 (3): e0213544. https://doi.org/10.1371/journal.pone.0213544.

Gundersen, Torbjørn. 2018. "Scientists as Experts: A Distinct Role?" Studies in History and Philosophy of Science Part A 69 (June): 52–59. https://doi.org/10.1016/j.shpsa.2018.02.006.

Gundersen, Torbjørn, and Kristine Bærøe. 2022. "Ethical Algorithmic Advice: Some Reasons to Pause and Think Twice." The American Journal of Bioethics 22 (7): 26–28. https://doi.org/10.1080/15265161.2022.2075053.

Gunkel, David J. 2012. The Machine Question: Critical Perspectives on AI, Robots, and Ethics. MIT Press.

Haalboom, M., J. W. Gerritsen, and J. van der Palen. 2019. "Differentiation Between Infected and Non-Infected Wounds Using an Electronic Nose." Clinical Microbiology and Infection 25 (10): 1288.e1–6. https://doi.org/10.1016/j.cmi.2019.03.018.

Halpern, Joseph Y., and Max Kleiman-Weiner. 2018. "Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility." In 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 1853–60. https://arxiv.org/abs/1810.05903.

Hameed, Jawish, Ian Harrison, Mark N. Gasson, and Kevin Warwick. 2010. "A Novel Human-Machine Interface Using Subdermal Magnetic Implants." In 2010 IEEE 9th

International Conference on Cybernetic Intelligent Systems, CIS 2010. https://doi.org/10.1109/UKRICIS.2010.5898141.

Harari, Daniel, Joshua B. Tenenbaum, and Shimon Ullman. 2018. "Discovery and Usage of Joint Attention in Images." arXiv:1804.04604 [Cs, q-Bio], April. https://arxiv.org/abs/1804.04604.

Harvey, Nigel, and Ilan Fischer. 1997. "Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility." Organizational Behavior and Human Decision Processes 70 (2): 117–33. https://doi.org/10.1006/obhd.1997.2697.

Haupt, Randy L., and Sue Ellen Haupt. 2003. Practical Genetic Algorithms. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471671746.

Hernández-Orallo, José, and Karina Vold. 2019. "AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI." In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 507–13. Honolulu HI USA: ACM. https://doi.org/10.1145/3306618.3314238.

Hietanen, Jari K., Lauri Nummenmaa, Mikko J. Nyman, Riitta Parkkola, and Heikki Hämäläinen. 2006. "Automatic Attention Orienting by Social and Symbolic Cues Activates Different Neural Networks: An fMRI Study." NeuroImage 33 (1): 406–13. https://doi.org/10.1016/j.neuroimage.2006.06.048.

Hindriks, Frank, Igor Douven, and Henrik Singmann. 2016. "A New Angle on the Knobe Effect: Intentionality Correlates with Blame, Not with Praise: A New Angle on the Knobe Effect." Mind & Language 31 (2): 204–20. https://doi.org/10.1111/mila.12101.

Ho, Dr Cristy, and Professor Charles Spence. 2012. The Multisensory Driver: Implications for Ergonomic Car Interface Design. Ashgate Publishing, Ltd.

Hommel, Bernhard, Jay Pratt, Lorenza Colzato, and Richard Godijn. 2001. "Symbolic Control of Visual Attention." Psychological Science 12 (5): 360–65. https://doi.org/10.1111/1467-9280.00367.

Hoy, Matthew B. 2018. "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants." Medical Reference Services Quarterly 37 (1): 81–88. https://doi.org/10.1080/02763869.2018.1404391.

Hu, Di, Dong Wang, Xuelong Li, Feiping Nie, and Qi Wang. 2019. "Listen to the Image." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7972–81.

Hu, Wenwen, Liangtian Wan, Yingying Jian, Cong Ren, Ke Jin, Xinghua Su, Xiaoxia Bai, Hossam Haick, Mingshui Yao, and Weiwei Wu. 2019. "Electronic Noses: From Advanced Materials to Sensors Aided with Data Processing." Advanced Materials Technologies 4 (2): 1800488. https://doi.org/10.1002/admt.201800488.

Huang, Chien-Ming, and Andrea L. Thomaz. 2011. "Effects of Responding to, Initiating and Ensuring Joint Attention in Human-Robot Interaction." In 2011 RO-MAN, 65–71. https://doi.org/10.1109/ROMAN.2011.6005230.

Hurley, Susan, and Alva Noë. 2003. "Neural Plasticity and Consciousness." Biology and Philosophy 18 (1): 131–68. https://doi.org/10.1023/A:1023308401356.

Illari, Phyllis, and Luciano Floridi. 2014. "Information Quality, Data and Philosophy." In The Philosophy of Information Quality, edited by Luciano Floridi and Phyllis Illari, 5–23. Synthese Library. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-07121-3_2.

Irlenbusch, Bernd, and David J. Saxler. 2019. "The Role of Social Information, Market Framing, and Diffusion of Responsibility as Determinants of Socially Responsible Behavior." Journal of Behavioral and Experimental Economics 80 (June): 141–61. https://doi.org/10.1016/j.socec.2019.04.001.

Jamieson, Dale. 2018. "Animal Agency." The Harvard Review of Philosophy 25: 111–26. https://doi.org/10.5840/harvardreview201892518.

Jaynes, Tyler L. 2019. "Legal Personhood for Artificial Intelligence: Citizenship as the Exception to the Rule." AI & SOCIETY, 1–12.

Jipson, Jennifer L., and Susan A. Gelman. 2007. "Robots and Rodents: Children's Inferences About Living and Nonliving Kinds." Child Development 78 (6): 1675–88. https://doi.org/10.1111/j.1467-8624.2007.01095.x.

Kahn, Peter H., Rachel L. Severson, Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, and Nathan G. Freier. 2012. "Do People Hold a Humanoid Robot Morally Accountable for the Harm It Causes?" In Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction – HRI '12, 33. Boston, Massachusetts, USA: ACM Press. https://doi.org/10.1145/2157689.2157696.

Kampis, Dora, and Victoria Southgate. 2020. "Altercentric Cognition: How Others Influence Our Cognitive Processing." Trends in Cognitive Sciences 24 (11): 945–59. https://doi.org/10.1016/j.tics.2020.09.003.

Karpus, Jurgis, Adrian Krüger, Julia Tovar Verba, Bahador Bahrami, and Ophelia Deroy. 2021. "Algorithm Exploitation: Humans Are Keen to Exploit Benevolent AI." iScience 24 (6): 102679. https://doi.org/10.1016/j.isci.2021.102679.

Kaur, Simarjeet, Jimmy Singla, Lewis Nkenyereye, Sudan Jha, Deepak Prashar, Gyanendra Prasad Joshi, Shaker El-Sappagh, Md. Saiful Islam, and S. M. Riazul Islam. 2020. "Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives." IEEE Access 8: 228049–69. https://doi.org/10.1109/ACCESS.2020.3042273.

Kärcher, Silke M., Sandra Fenzlaff, Daniela Hartmann, Saskia K. Nagel, and Peter König. 2012. "Sensory Augmentation for the Blind." Frontiers in Human Neuroscience, no. MARCH 2012: 1–15. https://doi.org/10.3389/fnhum.2012.00037.

Keil, Julian. 2020. "Double Flash Illusions: Current Findings and Future Directions." Frontiers in Neuroscience 14. https://doi.org/10.3389/fnins.2020.00298.

Kerdegari, Hamideh, Yeongmi Kim, and Tony J. Prescott. 2016. "Head-Mounted Sensory Augmentation Device: Comparing Haptic and Audio Modality." In Biomimetic and Biohybrid Systems, edited by Nathan F. Lepora, Anna Mura, Michael Mangan, Paul F. M. J. Verschure, Marc Desmulliez, and Tony J. Prescott, 9793:107–18. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-42417-0_11.

Kim, Jang Hyun, Hae Sun Jung, Min Hyung Park, Seon Hong Lee, Haein Lee, Yonghwan Kim, and Dongyan Nan. 2022. "Exploring Cultural Differences of Public Perception of Artificial Intelligence via Big Data Approach." In HCI International 2022 Posters, edited by Constantine Stephanidis, Margherita Antona, and Stavroula Ntoa, 1580:427–32. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-06417-3_57.

Kim, Mooseop, YunKyung Park, KyeongDeok Moon, and Chi Yoon Jeong. 2021. "Analysis and Validation of Cross-Modal Generative Adversarial Network for Sensory Substitution." International Journal of Environmental Research and Public Health 18 (12): 6216. https://doi.org/10.3390/ijerph18126216.

Kim, Taemie, and Pamela Hinds. 2006. "Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction." In ROMAN 2006 – The 15th IEEE International Symposium on Robot and Human Interactive Communication, 80–85. https://doi.org/10.1109/ROMAN.2006.314398.

Kingma, Diederik P., and Max Welling. 2019. "An Introduction to Variational Autoencoders." Foundations and Trends in Machine Learning 12 (4): 307–92. https://doi.org/10.1561/2200000056.

Kirchkamp, Oliver, and Christina Strobel. 2019. "Sharing Responsibility with a Machine." Journal of Behavioral and Experimental Economics 80 (June): 25–33. https://doi.org/10.1016/j.socec.2019.02.010.

Kirtay, Murat, Olga A. Wudarczyk, Doris Pischedda, Anna K. Kuhlen, Rasha Abdel Rahman, John-Dylan Haynes, and Verena V. Hafner. 2020. "Modeling Robot Co-Representation: State-of-the-Art, Open Issues, and Predictive Learning as a Possible Framework." In 2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 1–8. Valparaiso, Chile: IEEE. https://doi.org/10.1109/ICDL-EpiRob48136.2020.9278031.

Kneer, Markus. 2021. "Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents." Cognitive Science 45 (10). https://doi.org/10.1111/cogs.13032.

Kneer, Markus, and Izabela Skoczeń. 2021. "Outcome Effects, Moral Luck and the Hindsight Bias." SSRN Electronic Journal. https://doi.org/10.31234/OSF.IO/K6FPA.

Knobe, J. 2003. "Intentional Action and Side Effects in Ordinary Language." Analysis 63 (3): 190–94. https://doi.org/10.1093/analys/63.3.190.

Knobe, Joshua. 2003. "Intentional Action in Folk Psychology: An Experimental Investigation." Philosophical Psychology 16 (2): 309–24. https://doi.org/10.1080/09515080307771.

———. 2006. “The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology.” Philosophical Studies 130 (2): 203–31. https://doi.org/10.1007/s11098-004-4510-0.

Kominsky, Jonathan F., Jonathan Phillips, Tobias Gerstenberg, David Lagnado, and Joshua Knobe. 2015. “Causal Superseding.” Cognition 137 (April): 196–209. https://doi.org/10.1016/j.cognition.2015.01.013.

Kompatsiari, Kyveli, Francesca Ciardo, Vadim Tikhanoff, Giorgio Metta, and Agnieszka Wykowska. 2021. “It’s in the Eyes: The Engaging Role of Eye Contact in HRI.” International Journal of Social Robotics 13 (3): 525–35. https://doi.org/10.1007/s12369-019-00565-4.

Kotseruba, Iuliia, and John K. Tsotsos. 2020. “40 Years of Cognitive Architectures: Core Cognitive Abilities and Practical Applications.” Artificial Intelligence Review 53 (1): 17–94. https://doi.org/10.1007/s10462-018-9646-y.

Köbis, Nils, Jean-François Bonnefon, and Iyad Rahwan. 2021. “Bad Machines Corrupt Good Morals.” Nature Human Behaviour 5 (6): 679–85. https://doi.org/10.1038/s41562-021-01128-2.

Lai, Vivian, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. “Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies.” arXiv. https://doi.org/10.48550/arXiv.2112.11471.

Lee, Hosub, Cameron Upright, Steven Eliuk, and Alfred Kobsa. 2016. “Personalized Object Recognition for Augmenting Human Memory.” In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, 1054–61. UbiComp ’16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2968219.2968568.

Lench, Heather C., Darren Domsky, Rachel Smallman, and Kathleen E. Darbor. 2015. “Beliefs in Moral Luck: When and Why Blame Hinges on Luck.” British Journal of Psychology 106 (2): 272–87. https://doi.org/10.1111/bjop.12072.

Lewis, David. 1969. Convention. Cambridge, MA: The MIT Press.

Li, Xiangyang, and Qiang Ji. 2005. “Active Affective State Detection and User Assistance with Dynamic Bayesian Networks.” IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 35 (1): 93–105. https://doi.org/10.1109/TSMCA.2004.838454.

Liang, Ci-Jyun, Xi Wang, Vineet R. Kamat, and Carol C. Menassa. 2021. “Human-Robot Collaboration in Construction: Classification and Research Trends.” Journal of Construction Engineering and Management 147 (10): 03121006. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002154.

Lim, Velvetina, Maki Rooksby, and Emily S. Cross. 2021. “Social Robots on a Global Stage: Establishing a Role for Culture During Human.” International Journal of Social Robotics 13 (6): 1307–33. https://doi.org/10.1007/s12369-020-00710-4.

Lima, Gabriel, Nina Grgić-Hlača, and Meeyoung Cha. 2021. “Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making.” In

Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–17. Yokohama Japan: ACM. https://doi.org/10.1145/3411764.3445260.

List, Christian. 2018. "What Is It Like to Be a Group Agent?" Noûs 52 (2): 295–319.

———. 2019. "Group Agency and Artificial Intelligence." Preprint.

List, Christian, and Philip Pettit. 2011. Group Agency: The Possibility, Design, and Status of Corporate Agents. Oxford University Press.

Liu, Hongyi, and Lihui Wang. 2018. "Gesture Recognition for Human-Robot Collaboration: A Review." International Journal of Industrial Ergonomics 68 (November): 355–67. https://doi.org/10.1016/j.ergon.2017.02.004.

Lockerd, A., and C. Breazeal. 2004. "Tutelage and Socially Guided Robot Learning." In 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566), 4:3475–3480 vol.4. https://doi.org/10.1109/IROS.2004.1389954.

Longin, Louis. 2020. "Towards a Middle-Ground Theory of Agency for Artificial Intelligence." In Frontiers in Artificial Intelligence and Applications, edited by Marco Nørskov, Johanna Seibt, and Oliver Santiago Quick. IOS Press. https://doi.org/10.3233/FAIA200897.

Longin, Louis, and Ophelia Deroy. 2022. "Augmenting Perception: How Artificial Intelligence Transforms Sensory Substitution." Consciousness and Cognition 99 (March): 103280. https://doi.org/10.1016/j.concog.2022.103280.

MacIver, Malcolm Angus. 2009. "Neuroethology: From Morphological Computation to Planning." In The Cambridge Handbook of Situated Cognition, 480–504. Cambridge University Press.

Malle, Bertram F., Steve Guglielmo, and Andrew E. Monroe. 2014. "A Theory of Blame." Psychological Inquiry 25 (2): 147–86. https://doi.org/10.1080/1047840X.2014.877340.

Malle, Bertram F., Stuti Thapa Magar, and Matthias Scheutz. 2019. "AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma." In Robotics and Well-Being, edited by Maria Isabel Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Osman Tokhi, and Endre E. Kadar, 95:111–33. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_11.

Marks, Lawrence E., Robin J. Hammeal, Marc H. Bornstein, and Linda B. Smith. 1987. "Perceiving Similarity and Comprehending Metaphor." Monographs of the Society for Research in Child Development 52 (1): i. https://doi.org/10.2307/1166084.

Martini, Molly C., George A. Buzzell, and Eva Wiese. 2015. "Agent Appearance Modulates Mind Attribution and Social Attention in Human-Robot Interaction." In Social Robotics, edited by Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi, 431–39. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-25554-5_43.

Mavridis, Nikolaos. 2015. "A Review of Verbal and Non-Verbal Humanrobot Interactive Communication." Robotics and Autonomous Systems 63 (January): 22–35. https://doi.org/10.1016/j.robot.2014.09.031.

McGreal, Rory. 2018. "Hearables for Online Learning." The International Review of Research in Open and Distributed Learning 19 (September). https://doi.org/10.19173/irrodl.v19i4.4142.

Mcgurk, Harry, and John Macdonald. 1976. "Hearing Lips and Seeing Voices." Nature 264 (5588): 746–48. https://doi.org/10.1038/264746a0.

McManus, Ryan M., and Abraham M. Rutchick. 2019. "Autonomous Vehicles and the Attribution of Moral Responsibility." Social Psychological and Personality Science 10 (3): 345–52. https://doi.org/10.1177/1948550618755875.

Meijer, P. B. L. 1992. "An Experimental System for Auditory Image Representations." IEEE Transactions on Biomedical Engineering 39 (2): 112–21. https://doi.org/10.1109/10.121642.

Menary, Richard. 2007. Cognitive Integration. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230592889.

Meshi, Dar, Guido Biele, Christoph W. Korn, and Hauke R. Heekeren. 2012. "How Expert Advice Influences Decision Making." Edited by Jean Daunizeau. PLoS ONE 7 (11): e49748. https://doi.org/10.1371/journal.pone.0049748.

Middleton, Stuart E., Emmanuel Letouzé, Ali Hossaini, and Adriane Chapman. 2022. "Trust, Regulation, and Human-in-the-Loop AI: Within the European Region." Communications of the ACM 65 (4): 64–68. https://doi.org/10.1145/3511597.

Minsky, Marvin L. 1968. Semantic Information Processing. MIT Press.

Misselhorn, Catrin, ed. 2015. Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-15515-9.

Moglia, Andrea, Konstantinos Georgiou, Evangelos Georgiou, Richard M. Satava, and Alfred Cuschieri. 2021. "A Systematic Review on Artificial Intelligence in Robot-Assisted Surgery." International Journal of Surgery 95 (November): 106151. https://doi.org/10.1016/j.ijsu.2021.106151.

Morar, Anca, Florica Moldoveanu, Lucian Petrescu, and Alin Moldoveanu. 2017. "Real Time Indoor 3d Pipeline for an Advanced Sensory Substitution Device." In Image Analysis and Processing – ICIAP 2017, edited by Sebastiano Battiato, Giovanni Gallo, Raimondo Schettini, and Filippo Stanco, 685–95. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68548-9_62.

Mordatch, Igor, and Pieter Abbeel. 2018. "Emergence of Grounded Compositional Language in Multi-Agent Populations." arXiv. https://doi.org/10.48550/arXiv.1703.04908.

Moscovici, Serge, and Marisa Zavalloni. 1969. "The Group as a Polarizer of Attitudes." Journal of Personality and Social Psychology 12 (2): 125–35. https://doi.org/10.1037/h0027568.

Mundy, Peter. 2018. "A Review of Joint Attention and Social-Cognitive Brain Systems in Typical Development and Autism Spectrum Disorder." European Journal of Neuroscience 47 (6): 497–514. https://doi.org/10.1111/ejn.13720.

Mundy, Peter, Lisa Sullivan, and Ann M. Mastergeorge. 2009. “A Parallel and Distributed-Processing Model of Joint Attention, Social Cognition and Autism.” Autism Research 2 (1): 2–21. https://doi.org/10.1002/aur.61.

Mutlu, Bilge, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. “Nonverbal Leakage in Robots: Communication of Intentions Through Seemingly Unintentional Behavior.” In Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, 69–76. HRI ’09. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1514095.1514110.

Müller, Vincent C., and Matej Hoffmann. 2017. “What Is Morphological Computation? On How the Body Contributes to Cognition and Control.” Artificial Life 23 (1): 1–24. https://doi.org/10.1162/ARTL_a_00219.

Nadkarni, Swen, and Reinhard Prügl. 2021. “Digital Transformation: A Review, Synthesis and Opportunities for Future Research.” Management Review Quarterly 71 (2): 233–341. https://doi.org/10.1007/s11301-020-00185-7.

Nagel, Saskia K, Christine Carl, Tobias Kringe, Robert Märtin, and Peter König. 2005. “Beyond Sensory Substitutionlearning the Sixth Sense.” Journal of Neural Engineering 2 (4): R13–26. https://doi.org/10.1088/1741-2560/2/4/R02.

Nigam, Milena K, and David Klahr. 2000. “If Robots Make Choices, Are They Alive?: Children’s Judgements of the Animacy of Intelligent Artifacts.” In. Vol. 22. Proceedings of the Annual Meeting of the Cognitive Science Society.

Nyholm, Sven. 2018. “Attributing Agency to Automated Systems: Reflections on Human and Responsibility-Loci.” Science and Engineering Ethics 24 (4): 1201–19. https://doi.org/10.1007/s11948-017-9943-x.

Nyholm, Sven, and Jilles Smids. 2016. “The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?” Ethical Theory and Moral Practice 19 (5): 1275–89. https://doi.org/10.1007/s10677-016-9745-2.

O’Sullivan, Shane, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo, Katharina Holzinger, Andreas Holzinger, Mohammed Imran Sajid, and Hutan Ashrafian. 2019. “Legal, Regulatory, and Ethical Frameworks for Development of Standards in Artificial Intelligence (AI) and Autonomous Robotic Surgery.” The International Journal of Medical Robotics and Computer Assisted Surgery 15 (1): e1968. https://doi.org/10.1002/rcs.1968.

Palmeira, Mauricio, Gerri Spassova, and Hean Tat Keh. 2015. “Other-Serving Bias in Advice-Taking: When Advisors Receive More Credit Than Blame.” Organizational Behavior and Human Decision Processes 130 (September): 13–25. https://doi.org/10.1016/j.obhdp.2015.06.001.

Palmer, Erica D., and David A. Kobus. 2007. “The Future of Augmented Cognition Systems in Education and Training.” In Foundations of Augmented Cognition, edited by Dylan D. Schmorrow and Leah M. Reeves, 373–79. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-73216-7_42.

Panesar, Sandip S, Michel Kliot, Rob Parrish, Juan Fernandez-Miranda, Yvonne Cagle, and Gavin W Britz. 2020. "Promises and Perils of Artificial Intelligence in Neurosurgery." Neurosurgery 87 (1): 33–44. https://doi.org/10.1093/neuros/nyz471.

Pelau, Corina, Dan-Cristian Dabija, and Irina Ene. 2021. "What Makes an AI Device Human-Like? The Role of Interaction Quality, Empathy and Perceived Psychological Anthropomorphic Characteristics in the Acceptance of Artificial Intelligence in the Service Industry." Computers in Human Behavior 122 (September): 106855. https://doi.org/10.1016/j.chb.2021.106855.

Perez-Osorio, Jairo, Hermann J. Müller, and Agnieszka Wykowska. 2017. "Expectations Regarding Action Sequences Modulate Electrophysiological Correlates of the Gaze-Cueing Effect." Psychophysiology 54 (7): 942–54. https://doi.org/10.1111/psyp.12854.

Persson, Anders, Mikael Laaksoharju, and Hiroshi Koga. 2021. "We Mostly Think Alike: Individual Differences in Attitude Towards AI in Sweden and Japan." The Review of Socionetwork Strategies 15 (1): 123–42. https://doi.org/10.1007/s12626-021-00071-y.

Pescetelli, Niccolo. 2021. "A Brief Taxonomy of Hybrid Intelligence." Forecasting 3 (3): 633–43. https://doi.org/10.3390/forecast3030039.

Pettit, Philip. 1990. "Virtus Normativa: Rational Choice Perspectives." Ethics 100 (4): 725–55. https://doi.org/10.1086/293231.

———. 2007. "Responsibility Incorporated." Ethics 117 (2): 171–201. https://doi.org/10.1086/510695.

Pfeifer, Rolf, and Josh Bongard. 2006. How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press.

Pitsch, Karola, Anna-Lisa Vollmer, and Manuel Mühlig. 2013. "Robot Feedback Shapes the Tutor's Presentation: How a Robot's Online Gaze Strategies Lead to Micro-Adaptation of the Human's Conduct." Interaction Studies 14 (2): 268–96. https://doi.org/10.1075/is.14.2.06pit.

Posner, Michael I. 1980. "Orienting of Attention." Quarterly Journal of Experimental Psychology 32 (1): 3–25. https://doi.org/10.1080/00335558008248231.

Powers, Thomas M. 2013. "On the Moral Agency of Computers." Topoi. An International Review of Philosophy 32 (2): 227–36. https://doi.org/10.1007/s11245-012-9149-4.

Prinz, Jesse. 2001. "Is Consciousness Embodied?" In The Cambridge Handbook of Situated Cognition, edited by Philip Robbins and Murat Aydede, First, 419–36. Cambridge University Press. https://doi.org/10.1017/CBO9780511816826.022.

Proulx, Michael J., Petra Stoerig, Eva Ludowig, and Inna Knoll. 2008. "Seeing 'Where' Through the Ears: Effects of Learning-by-Doing and Long-Term Sensory Deprivation on Localization Based on Image-to-Sound Substitution." Edited by Malika Auvray. PLOS ONE 3 (3): e1840. https://doi.org/10.1371/journal.pone.0001840.

Pulikkaseril, Cibby, and Stanley Lam. 2019. "Laser Eyes for Driverless Cars: The Road to Automotive LIDAR." In 2019 Optical Fiber Communications Conference and Exhibition, OFC 2019 – Proceedings. https://doi.org/10.1364/ofc.2019.tu3d.2.

Purves, Duncan, Ryan Jenkins, and Bradley J. Strawser. 2015. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." Ethical Theory and Moral Practice 18 (4): 851–72. https://doi.org/10.1007/s10677-015-9563-y.

Pylyshyn, Zenon W. 2003. Seeing and Visualizing: It's Not What You Think. MIT Press.

Raisamo, Roope, Ismo Rakkolainen, Päivi Majaranta, Katri Salminen, Jussi Rantala, and Ahmed Farooq. 2019. "Human Augmentation: Past, Present and Future." International Journal of Human Computer Studies 131: 131–43. https://doi.org/10.1016/j.ijhcs.2019.05.008.

Raleigh, Thomas. 2017. "Phenomenal Privacy, Similarity and Communicability." Ergo, an Open Access Journal of Philosophy 4 (20201214). https://doi.org/10.3998/ergo.12405314.0004.022.

Reddy, Vasudevi, and Paul Morris. 2004. "Participants Don't Need Theories: Knowing Minds in Engagement." Theory & Psychology 14 (5): 647–65. https://doi.org/10.1177/0959354304046177.

Reeder, Blaine, Paul F. Cook, Paula M. Meek, and Mustafa Ozkaynak. 2017. "Smart Watch Potential to Support Augmented Cognition for Health-Related Decision Making." In Augmented Cognition. Neurocognition and Machine Learning, edited by Dylan D. Schmorrow and Cali M. Fidopiastis, 372–82. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58628-1_29.

Reeves, Adam, and Baingio Pinna. 2017. "Editorial: The Future of Perceptual Illusions: From Phenomenology to Neuroscience." Frontiers in Human Neuroscience 11. https://doi.org/10.3389/fnhum.2017.00009.

Rensink, Ronald A. 2013. "Perception and Attention." In The Oxford Handbook of Cognitive Psychology, 97–116. Oxford Library of Psychology. New York, NY, US: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.001.0001.

Richardson, Daniel C., Chris N. H. Street, Joanne Y. M. Tan, Natasha Z. Kirkham, Merrit A. Hoover, and Arezou Ghane Cavanaugh. 2012. "Joint Perception: Gaze and Social Context." Frontiers in Human Neuroscience 6. https://doi.org/10.3389/fnhum.2012.00194.

Risko, Evan F., and Sam J. Gilbert. 2016. "Cognitive Offloading." Trends in Cognitive Sciences 20 (9): 676–88. https://doi.org/10.1016/j.tics.2016.07.002.

Ristic, Jelena, Chris Kelland Friesen, and Alan Kingstone. 2002. "Are Eyes Special? It Depends on How You Look at It." Psychonomic Bulletin & Review 9 (3): 507–13. https://doi.org/10.3758/BF03196306.

Rodríguez-Ruiz, Alejandro, Elizabeth Krupinski, Jan-Jurre Mordang, Kathy Schilling, Sylvia H. Heywang-Köbrunner, Ioannis Sechopoulos, and Ritse M. Mann. 2018. "Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System." Radiology 290 (2): 305–14. https://doi.org/10.1148/radiol.2018181371.

Roese, Neal J., and Kathleen D. Vohs. 2012. "Hindsight Bias." Perspectives on Psychological Science 7 (5): 411–26. https://doi.org/10.1177/1745691612454303.

Rowlands, Mark. 1999. The Body in Mind: Understanding Cognitive Processes. Cambridge University Press.

Rowlands, Mark, Joe Lau, and Max Deutsch. 2020. "Externalism About the Mind." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Winter 2020. Metaphysics Research Lab, Stanford University.

Ruiz-Mirazo, Kepa, and Alvaro Moreno. 2000. "Searching for the Roots of Autonomy: The Natural and Artificial Paradigms Revisited." Communication and Cognition-Artificial Intelligence, 209–28.

Rumsey, Francis. 2012. Spatial Audio. Zeroth. Routledge. https://doi.org/10.4324/9780080498195.

Rupert, Robert D. 2004. "Challenges to the Hypothesis of Extended Cognition." The Journal of Philosophy 101 (8): 389–428. https://doi.org/10.5840/jphil2004101826.

Russell, Stuart J, and Peter Norvig. 2016. Artificial Intelligence: A Modern Approach. Malaysia; Pearson Education Limited,.

Sacheli, Lucia Maria, Elisa Arcangeli, and Eraldo Paulesu. 2018. "Evidence for a Dyadic Motor Plan in Joint Action." Scientific Reports 8 (1): 5027. https://doi.org/10.1038/s41598-018-23275-9.

Sahaï, Aïsha, Andrea Desantis, Ouriel Grynszpan, Elisabeth Pacherie, and Bruno Berberian. 2019. "Action Co-Representation and the Sense of Agency During a Joint Simon Task: Comparing Human and Machine Co-Agents." Consciousness and Cognition 67 (January): 44–55. https://doi.org/10.1016/j.concog.2018.11.008.

Sahaï, Aïsha, Elisabeth Pacherie, Ouriel Grynszpan, and Bruno Berberian. 2017. "Predictive Mechanisms Are Not Involved the Same Way During Human-Human Vs. Human-Machine Interactions: A Review." Frontiers in Neurorobotics 11 (October): 52. https://doi.org/10.3389/fnbot.2017.00052.

Samson, Dana, Ian A. Apperly, Jason J. Braithwaite, Benjamin J. Andrews, and Sarah E. Bodley Scott. 2010. "Seeing It Their Way: Evidence for Rapid and Involuntary Computation of What Other People See." Journal of Experimental Psychology: Human Perception and Performance 36 (5): 1255–66. https://doi.org/10.1037/a0018729.

Santoni de Sio, Filippo, and Giulio Mecacci. 2021. "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them." Philosophy & Technology 34 (4): 1057–84. https://doi.org/10.1007/s13347-021-00450-x.

Sauppé, Allison, and Bilge Mutlu. 2014. "Robot Deictics: How Gesture and Context Shape Referential Communication." In Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction, 342–49. HRI '14. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2559636.2559657.

Scaife, M., and J. S. Bruner. 1975. "The Capacity for Joint Visual Attention in the Infant." Nature 253 (5489): 265–66. https://doi.org/10.1038/253265a0.

Schaekermann, Mike, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. "Ambiguity-Aware AI Assistants for Medical Data Analysis." In

Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–14. New York, NY, USA: Association for Computing Machinery.

Schiffer, Stephen. 1988. "Review of The Varieties of Reference." The Journal of Philosophy 85 (1): 33–42. https://doi.org/10.2307/2026900.

Schmidhuber, Jürgen. 2015. "Deep Learning in Neural Networks: An Overview." Neural Networks 61 (January): 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Schuch, Stefanie, and Steven P. Tipper. 2007. "On Observing Another Person's Actions: Influences of Observed Inhibition and Errors." Perception & Psychophysics 69 (5): 828–37. https://doi.org/10.3758/BF03193782.

Sebanz, Natalie, and Guenther Knoblich. 2009. "Prediction in Joint Action: What, When, and Where." Topics in Cognitive Science 1 (2): 353–67. https://doi.org/10.1111/j.1756-8765.2009.01024.x.

Sebanz, Natalie, Günther Knoblich, and Wolfgang Prinz. 2005. "How Two Share a Task: Corepresenting Stimulus-Response Mappings." Journal of Experimental Psychology. Human Perception and Performance 31 (6): 1234–46. https://doi.org/10.1037/0096-1523.31.6.1234.

Sebanz, N, H Bekkering, and G Knoblich. 2006. "Joint Action: Bodies and Minds Moving Together." Trends in Cognitive Sciences 10 (2): 70–76. https://doi.org/10.1016/j.tics.2005.12.009.

Seemann, Axel, ed. 2011. Joint Attention: New Developments in Psychology, Philosophy of Mind, and Social Neuroscience. Cambridge, Mass: MIT Press.

———. 2019. The Shared World: Perceptual Common Knowledge, Demonstrative Communication, and Social Space. Cambridge, MA: The MIT Press.

Seow, Tricia, and Stephen M. Fleming. 2019. "Perceptual Sensitivity Is Modulated by What Others Can See." Attention, Perception, & Psychophysics 81 (6): 1979–90. https://doi.org/10.3758/s13414-019-01724-5.

Setia, Pankaj, Pankat Setia, Viswanath Venkatesh, and Supreet Joglekar. 2013. "Leveraging Digital Technologies: How Information Quality Leads to Localized Capabilities and Customer Service Performance." MIS Quarterly 37 (2): 565–90. https://www.jstor.org/stable/43825923.

Shams, Ladan, Yukiyasu Kamitani, and Shinsuke Shimojo. 2000. "What You See Is What You Hear." Nature 408 (6814): 788–88. https://doi.org/10.1038/35048669.

Shank, Daniel B., Alyssa DeSanti, and Timothy Maninger. 2019. "When Are Artificial Intelligence Versus Human Agents Faulted for Wrongdoing? Moral Attributions After Individual and Joint Decisions." Information, Communication & Society 22 (5): 648–63. https://doi.org/10.1080/1369118X.2019.1568515.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2018. "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-Play." Science 362 (6419): 1140–44. https://doi.org/10.1126/science.aar6404.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. 2017. "Mastering the Game of Go Without Human Knowledge." Nature. https://doi.org/10.1038/nature24270.

Simon, J. Richard. 1969. "Reactions Toward the Source of Stimulation." Journal of Experimental Psychology 81 (1): 174–76. https://doi.org/10.1037/h0027448.

Siposova, Barbora, and Malinda Carpenter. 2019. "A New Look at Joint Attention and Common Knowledge." Cognition 189 (August): 260–74. https://doi.org/10.1016/j.cognition.2019.03.019.

Smart, Paul. 2017. "Extended Cognition and the Internet: A Review of Current Issues and Controversies." Philosophy and Technology 30 (3): 357–90. https://doi.org/10.1007/s13347-016-0250-2.

Sparrow, Robert. 2007. "Killer Robots." Journal of Applied Philosophy 24 (1): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Steffel, Mary, Elanor F. Williams, and Jaclyn Perrmann-Graham. 2016. "Passing the Buck: Delegating Choices to Others to Avoid Responsibility and Blame." Organizational Behavior and Human Decision Processes 135 (July): 32–44. https://doi.org/10.1016/j.obhdp.2016.04.006.

Stevenson, Ryan, Raquel Zemtsov, and Mark Wallace. 2012. "Multisensory Illusions and the Temporal Binding Window." Journal of Experimental Psychology Human Perception & Performance 2 (January). https://doi.org/10.1068/ic903.

Steward, Helen. 2009. "Animal Agency." Inquiry 52 (3): 217–31.

Strasser, Anna. 2021. "Distributed Responsibility in Humanmachine Interactions." AI and Ethics, October. https://doi.org/10.1007/s43681-021-00109-5.

Stuart, Michael T., and Markus Kneer. 2021. "Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents." Proceedings of the ACM on Human-Computer Interaction 5 (CSCW2): 1–27. https://doi.org/10.1145/3479507.

Sun, Wen, Jiajia Liu, and Haibin Zhang. 2017. "When Smart Wearables Meet Intelligent Vehicles: Challenges and Future Directions." IEEE Wireless Communications 24 (3): 58–65. https://doi.org/10.1109/MWC.2017.1600423.

Surtees, Andrew, Ian Apperly, and Dana Samson. 2016. "I've Got Your Number: Spontaneous Perspective-Taking in an Interactive Task." Cognition 150 (May): 43–52. https://doi.org/10.1016/j.cognition.2016.01.014.

Šabanović, Selma, Casey C. Bennett, Wan-Ling Chang, and Lesa Huber. 2013. "PARO Robot Affects Diverse Interaction Modalities in Group Sensory Therapy for Older Adults with Dementia." In 2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR), 1–6. https://doi.org/10.1109/ICORR.2013.6650427.

Tajadura-Jiménez, Ana, Aleksander Väljamäe, and Kristi Kuusk. 2020. "Altering One's Body-Perception Through E-Textiles and Haptic Metaphors." Frontiers in Robotics and AI 7. https://doi.org/10.3389/frobt.2020.00007.

Takayama, Leila, Doug Dooley, and Wendy Ju. 2011. "Expressing Thought: Improving Robot Readability with Animation Principles." In Proceedings of the 6th International Conference on Human-robot Interaction, 69–76. HRI '11. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1957656.1957674.

Tapus, Adriana, Maja J. Mataric, and Brian Scassellati. 2007. "Socially Assistive Robotics [Grand Challenges of Robotics]." IEEE Robotics & Automation Magazine 14 (1): 35–42. https://doi.org/10.1109/MRA.2007.339605.

Taylor, Tracy L., and Raymond M. Klein. 2000. "Visual and Motor Effects in Inhibition of Return." Journal of Experimental Psychology: Human Perception and Performance 26 (5): 1639–56. https://doi.org/10.1037/0096-1523.26.5.1639.

Team, RStudio. 2021. "RStudio: Integrated Development Environment for R." Boston, MA: RStudio, PBC.

Teigen, Karl Halvor, and Wibecke Brun. 2011. "Responsibility Is Divisible by Two, But Not by Three or Four: Judgments of Responsibility in Dyads and Groups." Social Cognition 29 (1): 15–42. https://doi.org/10.1521/soco.2011.29.1.15.

Teufel, Christoph, Paul C. Fletcher, and Greg Davis. 2010. "Seeing Other Minds: Attributed Mental States Influence Perception." Trends in Cognitive Sciences 14 (8): 376–82. https://doi.org/10.1016/j.tics.2010.05.005.

Thuillier, Etienne, Hannes Gamper, and Ivan J. Tashev. 2018. "Spatial Audio Feature Discovery with Convolutional Neural Networks." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6797–6801. Calgary, AB: IEEE. https://doi.org/10.1109/ICASSP.2018.8462315.

Tian, Guanzhong, Yi Yuan, and Yong Liu. 2019. "Audio2Face: Generating Speech/Face Animation from Single Audio with Attention-Based Bidirectional LSTM Networks." In 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 366–71. Shanghai, China: IEEE. https://doi.org/10.1109/ICMEW.2019.00069.

Tian, Yong-hong, Xi-lin Chen, Hong-kai Xiong, Hong-liang Li, Li-rong Dai, Jing Chen, Jun-liang Xing, et al. 2017. "Towards Human-Like and Transhuman Perception in AI 2.0: A Review." Frontiers of Information Technology & Electronic Engineering 18 (1): 58–67. https://doi.org/10.1631/FITEE.1601804.

Tipples, Jason. 2002. "Eye Gaze Is Not Unique: Automatic Orienting in Response to Uninformative Arrows." Psychonomic Bulletin & Review 9 (2): 314–18. https://doi.org/10.3758/BF03196287.

———. 2008. "Negative Emotionality Influences the Effects of Emotion on Time Perception." Emotion 8 (1): 127–31. https://doi.org/10.1037/1528-3542.8.1.127.

Tomasello, Michael. 1995. "Joint Attention as Social Cognition." In Joint Attention: Its Origins and Role in Development, 103–30. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

———. 2022. The Evolution of Agency: Behavioral Organization from Lizards to Humans. MIT Press.

Tosi, Alessia, Martin J. Pickering, and Holly P. Branigan. 2020. “Speakers’ Use of Agency and Visual Context in Spatial Descriptions.” Cognition 194 (January): 104070. https://doi.org/10.1016/j.cognition.2019.104070.

Trafton, J. G., N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz. 2005. “Enabling Effective Human in Robots.” IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 35 (4): 460–70. https://doi.org/10.1109/TSMCA.2005.850592.

Tsai, Chia-Chin, Wen-Jui Kuo, Daisy L. Hung, and Ovid J. L. Tzeng. 2008. “Action Co-representation Is Tuned to Other Humans.” Journal of Cognitive Neuroscience 20 (11): 2015–24. https://doi.org/10.1162/jocn.2008.20144.

Tversky, Barbara, and Bridgette Martin Hard. 2009. “Embodied and Disembodied Cognition: Spatial Perspective-Taking.” Cognition 110 (1): 124–29. https://doi.org/10.1016/j.cognition.2008.10.008.

Tyler, Mitchell, Yuri Danilov, and Paul Bach-Y-Rita. 2003. “Closing an Open-Loop Control System: Vestibular Substitution Through the Tongue.” Journal of Integrative Neuroscience 2 (2): 159–64. https://doi.org/10.1142/S0219635203000263.

Varga, Somogy. 2017. “Perceptual Experience and Cognitive Penetrability.” European Journal of Philosophy 25 (2): 376–97. https://doi.org/10.1111/ejop.12239.

Verhoef, Peter C., Thijs Broekhuizen, Yakov Bart, Abhi Bhattacharya, John Qi Dong, Nicolai Fabian, and Michael Haenlein. 2021. “Digital Transformation: A Multidisciplinary Reflection and Research Agenda.” Journal of Business Research 122 (January): 889–901. https://doi.org/10.1016/j.jbusres.2019.09.022.

Vial, Gregory. 2019. “Understanding Digital Transformation: A Review and a Research Agenda.” The Journal of Strategic Information Systems, SI: Review issue, 28 (2): 118–44. https://doi.org/10.1016/j.jsis.2019.01.003.

Wallach, Wendell, and Colin Allen. 2009. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195374049.001.0001.

Wallach, Wendell, Stan Franklin, and Colin Allen. 2010. “A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents.” Topics in Cognitive Science 2 (3): 454–85. https://doi.org/10.1111/j.1756-8765.2010.01095.x.

Wen, Tanya, and Shulan Hsieh. 2015. “Neuroimaging of the Joint Simon Effect with Believed Biological and Non-Biological Co-Actors.” Frontiers in Human Neuroscience 9.

Wenke, Dorit, Silke Atmaca, Antje Holländer, Roman Liepelt, Pamela Baess, and Wolfgang Prinz. 2011. “What Is Shared in Joint Action? Issues of Co-representation, Response Conflict, and Agent Identification,” 27.

Wheeler, Michael. 2019. “The Reappearing Tool: Transparency, Smart Technology, and the Extended Mind.” AI & SOCIETY 34 (4): 857–66. https://doi.org/10.1007/s00146-018-0824-x.

Williams, Garrath. 2013. "Sharing Responsibility and Holding Responsible." Journal of Applied Philosophy 30 (4): 351–64. https://doi.org/10.1111/japp.12019.

Wilson, Catherine. 2004. Moral Animals: Ideals and Constraints in Moral Theory. Oxford University Press on Demand.

Wilson, Robert A. 2010. "Extended Vision." In Perception, Action and Consciousness, edited by N Gangopadhyay, M Madary, and F Spicer, 277–90.

Wischert-Zielke, Moritz, Klemens Weigl, Marco Steinhauser, and Andreas Riener. 2020. "Age Differences in the Anticipated Acceptance of Egoistic Versus Altruistic Crash-Control-Algorithms in Automated Vehicles." In Proceedings of the Conference on Mensch Und Computer, 467–71. Magdeburg Germany: ACM. https://doi.org/10.1145/3404983.3409992.

Wright, Thomas D., and Jamie Ward. 2018. "Sensory Substitution Devices as Advanced Sensory Tools." In Sensory Substitution and Augmentation, 219:188–204. Proceedings of the British Academy. The British Academy. https://doi.org/10.5871/bacad/9780197266441.003.0012.

Wright, Thomas, and Jamie Ward. 2013. "The Evolution of a Visual-to-Auditory Sensory Substitution Device Using Interactive Genetic Algorithms." Quarterly Journal of Experimental Psychology 66 (8): 1620–38. https://doi.org/10.1080/17470218.2012.754911.

Xiaodong Wang, and H. V. Poor. 1998. "Blind Multiuser Detection: A Subspace Approach." IEEE Transactions on Information Theory 44 (2): 677–90. https://doi.org/10.1109/18.661512.

Yampolskiy, Roman V. 2021. "AI Personhood: Rights and Laws." Chapter. Machine Law, Ethics, and Morality in the Age of Artificial Intelligence. https://www.igi.global.com/chapter/ai-personhood/www.igi-global.com/chapter/ai-personhood/265710; IGI Global. https://doi.org/10.4018/978-1-7998-4894-3.ch001.

Yonezawa, Tomoko, Hirotake Yamazoe, Akira Utsumi, and Shinji Abe. 2007. "Gaze-Communicative Behavior of Stuffed-Toy Robot with Joint Attention and Eye Contact Based on Ambient Gaze-Tracking." In Proceedings of the 9th International Conference on Multimodal Interfaces, 140–45. ICMI '07. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1322192.1322218.

Young, Aaron J., and Daniel P. Ferris. 2017. "State of the Art and Future Directions for Lower Limb Robotic Exoskeletons." IEEE Transactions on Neural Systems and Rehabilitation Engineering 25 (2): 171–82. https://doi.org/10.1109/TNSRE.2016.2521160.

Yun, Kyongsik, Katsumi Watanabe, and Shinsuke Shimojo. 2012. "Interpersonal Body and Neural Synchronization as a Marker of Implicit Social Interaction." Scientific Reports 2 (1): 959. https://doi.org/10.1038/srep00959.

Zahavi, D. 2015. "You, Me, and We: The Sharing of Emotional Experiences." Journal of Consciousness Studies 22 (1-2): 84–101.

Zanesco, Julie, Eda Tipura, Andres Posada, Fabrice Clément, and Alan J. Pegna. 2019. "Seeing Is Believing: Early Perceptual Brain Processes Are Modified by Social Feedback." Social Neuroscience 14 (5): 519–29. https://doi.org/10.1080/17470919.2018.1511470.

Zanzotto, Fabio Massimo. 2019. "Viewpoint: Human-in-the-loop Artificial Intelligence." Journal of Artificial Intelligence Research 64 (February): 243–52. https://doi.org/10.1613/jair.1.11345.

Zeng, Fan Gang, Stephen Rebscher, William Harrison, Xiaoan Sun, and Haihong Feng. 2008. "Cochlear Implants: System Design, Integration, and Evaluation." IEEE Reviews in Biomedical Engineering. https://doi.org/10.1109/RBME.2008.2008250.

Zhang, Jingling, Jane Conway, and César A. Hidalgo. 2022. "Why Do People Judge Humans Differently from Machines? The Role of Agency and Experience." https://doi.org/10.48550/ARXIV.2210.10081.

Zhang, Wenqiang, Bin Gao, Jianshi Tang, Peng Yao, Shimeng Yu, Meng-Fan Chang, Hoi-Jun Yoo, He Qian, and Huaqiang Wu. 2020. "Neuro-Inspired Computing Chips." Nature Electronics 3 (7): 371–82. https://doi.org/10.1038/s41928-020-0435-7.

Zhao, Xuan, and Bertram F. Malle. 2022. "Spontaneous Perspective Taking Toward Robots: The Unique Impact of Humanlike Appearance." Cognition 224 (July): 105076. https://doi.org/10.1016/j.cognition.2022.105076.

Zheng, Guan, and Hong Wu. 2019. "Collusive Algorithms as Mere Tools, Super-tools or Legal Persons." Journal of Competition Law & Economics, September, nhz010. https://doi.org/10.1093/joclec/nhz010.

Artificial Intelligence (AI) profoundly affects how people communicate, work, and perceive the world. While autonomous AI systems are the focal point in societal and academic discussions, advisory AI systems, which influence human decisions but don‘t undertake independent actions, often remain unexplored. Examples range from automated purchase recommendations to medical diagnoses. This dissertation seeks to understand what advisory AI systems truly are. Are they capable of autonomous, human-like action? Or can they be reduced to inert tools? And what happens when advisory AI systems are closely linked with human perception, especially through Augmented Reality and sensory augmentation? Does their ontological status change? This dissertation concludes that, regardless of their implementation, advisory AI systems occupy an ontological status between tools and humans. They are more than just tools but less than humans.

Louis Longin ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Philosophie des Geistes an der Ludwig-Maximilians-Universität München, wo er 2023 mit der vorliegenden Dissertation promoviert wurde. Seinem Interesse gilt der wachsende Einfluss der Künstlichen Intelligenz auf den menschlichen Nutzer, besonders in den Bereichen der sozialen Interaktion, Ethik und sensorischen Augmentation.

49,00 €
ISBN 978-3-487-16692-6