International Journal of Social Robotics (2024) 16:2223–2241
https://doi.org/10.1007/s12369-024-01184-4

Reframing Deception for Human-Centered AI

Steven Umbrello[1] · Simone Natale[2]

Accepted: 2 October 2024 / Published online: 25 October 2024

Abstract
The philosophical, legal, and HCI literature concerning artificial intelligence (AI) has explored the ethical implications and values that these systems will impact on. One aspect that has been only partially explored, however, is the role of deception. Due to the negative connotation of this term, research in AI and Human–Computer Interaction (HCI) has mainly considered deception to describe exceptional situations in which the technology either does not work or is used for malicious purposes. Recent theoretical and historical work, however, has shown that deception is a more structural component of AI than it is usually acknowledged. AI systems that enter in communication with users, in fact, forcefully invite reactions such as attributions of gender, personality and empathy, even in the absence of malicious intent and often also with potentially positive or functional impacts on the interaction. This paper aims to operationalise the Human-Centred AI (HCAI) framework to develop the implications of this body of work for practical approaches to AI ethics in HCI and design. In order to achieve this goal, we take up the analytical distinction between "banal" and "strong" deception, originally proposed in theoretical and historical scholarship on AI (Natale in Deceitful media: artificial intelligence and social life after the turing test, Oxford University Press, New York, 2021), as a starting point to develop ethical reflections that will empower designers and developers with practical ways to solve the problems raised by the complex relationship between deception and communicative AI. The paper considers how HCAI can be applied to conversational AI (CAI) systems in order to design them to develop banal deception for social good and, at the same time, to avoid its potential risks.

Keywords Human centered AI · Artificial intelligence · Deception · Banal deception · Ethics by design · Applied ethics

1 Introduction

The last two decades of academic research have been marked by a host of discussions dedicated to outlining how to ensure that innovations in digital technologies remain aligned with important human values [24]. However, given the rapid development and deployment of AI systems across most human domains, these systems risk becoming pervasive despite lacking a comprehensive approach to designing them to ensure they embody human values. There is an urgent need to move beyond conceptual investigations of the principles and values that AI systems must be aligned with to be deemed safe and trustworthy and towards practical approaches for designing such systems *for* these human values.

One of the critical problems that communities and policymakers need to address in order to mitigate these risks is the relationship between AI and deception. Recently, increasing efforts have been made to investigate how AI and robots might facilitate deceptive outcomes and which responses are needed from an ethical and technical viewpoint [14, 52]. As technologies such as chatbots, voice assistants, and robots develop the ability to engage in communicative interactions with human users, many worry that this will forcefully lead to deceptive mechanisms, as users are stimulated to anthropomorphise the technology or to overstate the capacity of the machine to socialise and empathise with them.

This places the debate about AI and deception before a conundrum that bears no simple solutions. How can technologies be created to communicate and interact socially with users while keeping a fair and human-centered approach? Since existing AI technologies can only create the impression of empathy and humanity, should technologies such as

✉ Steven Umbrello
steven.umbrello@unito.it

Simone Natale
simone.natale@unito.it

[1] Department of Philosophy and Educational Sciences, University of Turin, Via Giuseppe Verdi, 8, 10124 Turin, Italy

[2] Department of Humanities, University of Turin, Via Giuseppe Verdi, 8, Turin, Italy

social robots and companion chatbots be refused for being inherently deceptive? How can analytical frameworks and practical recommendations orient the design of communicative AI technologies to avoid the risks and problems related to deception?

This paper tackles such questions by proposing a framework that distinguishes between different situations and dynamics in which communicative AI [26] involves deceptive outcomes. In particular, it aims to provide an analytical distinction between situations in which deceptive mechanisms remain beneficial to users, enhancing rather than jeopardising human control and value sensitive design, and deceptive mechanisms that are risky, problematic or unfair [66]. Although deception is colloquially understood as having negative implications, that is not necessarily true. Specific forms of deception can be used towards socially beneficial ends as long as the functionality and productivity of deceptive processes remain controlled by users and designers [1]. Even more significantly, since AI technologies that have become of everyday use such as voice assistants and LLM-based chatbots, simulate human abilities and characters, the emergence of deception may not be wholly obliterated or prevented in human–machine communications. Therefore, understanding the characteristics and outcomes of specific deceptive processes and situations is essential for assessing to what extent such processes and situations are viable and desirable to meet the goal of developing a human-centered approach to AI. While theoretical and historical scholarship on AI has recently advanced a perspective that recognises that deception is an integral component of AI systems programmed to enter into communicative interactions with users [13, 44, 53], this acknowledgement needs to be accompanied by the development of analytical tools that can practically orient design in human–computer interaction and AI—a gap that this article aims to address.

Our proposal recognises that any approach to deception and AI should begin by substantially investigating what deception actually means in the context of AI and the consequences of considering the different understandings of deception for the kind of mediated communication that is implied in human–computer interactions. Two theoretical insights in this regard guide the proposal presented here. First, we argue that deception is not the result of exceptional circumstances but is instead a structural component of interactive AI systems. Human users, in fact, forcefully project their perceptions, reactions and biases on AI systems exhibiting communicative behaviour; designers of these systems can anticipate the dynamics that can result from such acts of projection, thus leading the interaction towards deceptive situations that facilitate the desired outcomes. Second, we propose that it is possible to distinguish between deceptive mechanisms and designs that provide value to users and those that, instead, are of little use and bear, in fact, significant risks. This is particularly important for designing AI systems since it implies a shift of approaches to tackle the problem of deception in AI. Suppose deception is a significant and even integral dimension of interactions with AI. In that case, the question that needs to be asked is not if specific AI technologies facilitate deception but how the underlying structural deception of AI can be adjusted in ways that are functional, fair, ethical, and useful to the user. In this regard, we propose an analytical distinction between "banal" and "strong" deception, originally proposed in theoretical and historical scholarship on AI [44]. This framework can potentially empower designers and developers with a practical way to solve the problem raised by the complex relationship between deception and communicative AI.

By providing practical analytical tools to identify and assess problematic uses of deception in AI, this article contributes to existing efforts to interrogate the complex relationship between AI and ethics and to create theoretical and practical ground that can feed into ongoing efforts to regulate AI systems based on their risks (Artificial Intelligence Act [AI Act]). There have been several recent studies that look at how to ensure that specific AI systems can be designed *for* human values, including sustainability [86], human autonomy [7], privacy [67, 68], control [51], and justice [21], among others. The varied approaches to design have been applied in a host of domains of AI applications such as recommenders systems [28, 43], care robots [55, 79], autonomous vehicles [41, 81], and even military robots [75, 76, 78]. Although various approaches have been proposed for designing AI systems *for* human values (e.g., [2, 3, 78, 80, 85]), these approaches are limited given that generally work to adapt existing design approaches like value sensitive design, participatory design, and universal design, in order to meet the challenges that are posed by AI systems. Although approaches such as these have conceptual merit, they have yet to be tested in real-world instances to determine their effectiveness. This paper draws on the Human-centered AI (HCAI) framework, an approach designed specifically for AI systems to address these issues. The HCAI framework is an approach to ensure that banal deception can be mobilised to achieve better HCAI's goals of producing "successful technologies that augment, amplify, empower, and enhance humans rather than replace them" ([60], p. 4).

The remainder of this paper begins by explaining banal deception, how it differs from other forms of deception, and the technical implications that such forms entail. This is followed by examining how the analytical distinction between banal and strong deception can be incorporated into existing discussions and proposals under the auspices of the HCAI framework. To provide practical examples and illustrations of how the application of the theory can be used to tackle specific design problems and issues, we draw from the case of conversational AI systems (CAI), typically referred to as 'voice assistants', exploring how they can be designed to ensure these beneficial forms of deception are supported while minimising the more deleterious forms. The case of CAI is particularly apt to advance this analytical work since CAI, in contrast with technologies such as social robots or companion chatbots, have found wide application and have shown the potential to integrate functional design principles within a human-centered approach [60]. At the same time, however, the role of deception in CAI presents significant risks, as ongoing discussions about representations of sex, labour, and anthropomorphisation have shown (e.g. [25, 70, 90]). The distinction between banal and strong deception can provide, in this context, a useful resource to assess how to design and develop compelling CAI systems that will ensure that new AI-powered interfaces benefit people and businesses by coupling automation with the firm control of human users, thereby meeting HCAI recommendations and empowering present and future users of such systems.

2 Banal Deception: Theoretical Insight with Technical Implications

As computational systems that enter into communication with human users, such as conversational AI systems, grow in breadth and complexity [26], Human–computer interaction (HCI) and AI researchers face some of the same challenges that characterise human communication. One such challenge is the possibility that the design of interactive systems facilitates or causes users' deception [52]. Similar to deception in human communication [17], deception in HCI can result from several factors, including system design failures, intentionality, and user misinterpretation.

Increasingly in the last few years, a lively debate has ensued regarding the boundaries and definitions of deception in HCI and the theoretical, methodological, and practical tools that could be at the disposal of designers and users to counteract risks and problems associated with deception (e.g. [9]). As some have argued, all responses to this challenge must start with a serious engagement with the concept of deception [13, 14]. It is not only essential to define deception more rigorously but also to develop reliable analytical tools for distinguishing different types of deception in HCI. Distinguishing between different kinds of deception, in fact, not only helps identify and correctly describe situations where deception arises,it is also crucial to develop the ability to anticipate and assess their potential outcomes.

Traditional approaches to the problem of deception in HCI have tended to understand deception as an exception rather than as a structural element of interactions between humans and machines. Since deception is usually given a negative connotation, researchers and developers in areas such as AI and robotics have usually discussed deception as an unwanted outcome [52, 53]. Some, however, have acknowledged that deception might involve a broader range of situations than usually considered. For instance, Adar et al. stress that there are also benevolent forms of deception, which benefit both the developer and the user; benevolent deception would indeed be "ubiquitous in real-world system designs, although it is rarely described in such terms" ([1], p. 1863). Similarly, Chakraborti and Kambhampati [11] observe that the apparent outcome of embedding models of mental states of human users into AI programs is that it opens up the possibility of manipulation. Masters et al. [39] provide a taxonomy of computer deception forms, including imitating, obfuscating, tricking, calculating, and reframing. More recently, the phenomenon of hallucination in Large Language Models (LLMs) has been discussed as a bug but also as an integral feature of this technology [19].[1]

By advancing a different and broader application of the notion of deception in HCI, these and other interventions resonate with perspectives that have been recently developed in areas such as philosophy and cognitive sciences. These perspectives refuse to draw rigid boundaries between deception and "normal" perception, arguing for the need to account for the fact that deception represents an integral and functional aspect of human experience. Mark Wrathall, for instance, points out that "it rarely makes sense to say that I perceived either truly or falsely" ([91], p. 60), given the fact that the possibility of deception is ingrained in the mechanisms of our perception. Similarly, cognitive psychologist Donald D. Hoffman recently argued that evolution has shaped human perception in such a way that we can only navigate the physical world through "useful illusions," which make us perceive external reality in ways that are instrumental to our survival. These illusions are functional to our capacity to navigate the

[1] Recent advancements in generative AI technologies have introduced the phenomenon of AI "hallucination," where AI agents generate plausible but incorrect or misleading information. This capability, particularly prevalent in advanced language models, presents new challenges for managing deception in HCI and HRI. Hallucination can be viewed as a form of unintentional deception that lies on the continuum between banal and strong deception. While it might be tempting to tackle hallucination as an error of the models that generates disinformation effects, hallucination is, in fact, a consequence of the functionality of LLMs. These systems, in fact, do not simply recombine the textual data used for their training. they generate something new that may diverge significantly from the "source" of their training. This capacity to generate something new can be even convenient to users, who can employ the models to create something novel. Hallucination, in this sense, is not a bug but a feature of LLMs. Although it may not be intentional, the impact on users can be significant, especially when users are unable to discern the inaccuracies. Techniques such as confidence scoring and real-time fact-checking are crucial to mitigate the risks associated with AI hallucination, aligning with the principles of HCAI to maintain user trust and control [5, 42]. For further discussion, see [30, 64, 94].

external world but can also be manipulated through technology, advertising, and design [31].[2]

Transposed into HCI, these approaches suggest that deception does not represent the exception but is, instead, a structural component of interactions between users and machines. While deception has usually been conceptualised as intentional and malicious, there are actually more nuanced ways in which deception enters the experience of users as well as the work of designers and engineers.

To develop the practical implications of this, we draw from the conceptual distinction between "strong" and "banal" deception advanced by Natale [44], which can be usefully applied to practical approaches in HCI. In a historical study of AI, Natale moves from the observation that deception plays a fundamental and structural role in all AI technologies programmed to communicate with humans. Already at the origins of the field, Alan Turing's proposal of the Imitation Game or Turing Test framed the AI question in terms of deception: the computer passes the test if able to deceive human users into believing it is not a computer but a human. In the following decades, AI researchers realised that whenever AI technologies create interactions with humans, users can interpret the functioning and behaviour of the AI systems. Consequently, the possibility of deception is structurally incorporated in AI systems. For instance, in the case of CAI, users might overestimate the system's linguistic proficiency if the assistant replies to a query with an ironic response. Although such a response would be probably scripted by professional writers who work at companies such as Amazon, Apple, Microsoft, and Google for systems including Alexa, Siri, Cortana, and Google Assistant [93], and therefore the result of relatively simple algorithms at a technical level, the fact that irony is usually thought as evidence of human intelligence encourages users to overestimate the complexity of the process that resulted in such a response [25]. Such misconceptions can have significant implications for how users perceive CAI and, therefore impact habits and behaviours of interaction.

Another example is the hints at the sex that CAI convey through sexed voices and names. Considering the range of stereotypes and representations associated with sex, these are highly likely to inform the ways users approach these tools [54]. Notably, at least to some extent, these dynamics can also be anticipated by HCI designers, as Deborah Harrison, who was one of the "personality designers" of Microsoft's voice assistant Cortana, acknowledged in an interview [93]. Communicative AI software such as chatbots, social interfaces and social robots have been associated with theatrical characters, recognising that users tend to attribute characterisation to artificial agents programmed to display communicative behaviour (e.g. [35, 87]).

These examples demonstrate that deception entails a broader range of dynamics than is usually acknowledged in AI and HCI. The problem, however, is the distance between the more explicit and the more nuanced manifestations of deception. This is evident if one compares, for instance, the experience of the victim of a fraud who exchanged the automated email they received for a message from a real person with the experience of users of Alexa and Siri who are led to exaggerate the "intelligence" of the machine.

In order to make sense of such diversity, Natale [44, 45] proposes a distinction between two types of deception in AI, each representing one end of the spectrum in a continuum that goes from the most evident to the most nuanced case of deception. On the one hand, "strong" deception occurs when users are led to misunderstand the artificial nature of AI software. Examples of strong deception may include social media bots that are exchanged for human users [23], media reports produced by natural language generation software that pass as if they were authored by humans [29], and Google's project Duplex, an extension of Google Assistant that was meant to carry out phone conversations on behalf of users without disclosing its mechanical nature [48]. Three criteria can characterise it: (1) lack of transparency, (2) an intent to mislead, and (3) a potential to undermine user control or ethical considerations. On the other end of the spectrum is "banal" deception [44]. This type of deception, while less overt, still impacts user interactions and responses due to specific strategies and features embedded in the technology [62]. Instances of banal deception are typified by users reacting to programmed elements in AI systems, such as character voices, which encourage stereotyping, empathy, and projection without necessarily misconceiving the AI as human. Examples of banal deception include the reactions evoked by specific choices in the characterisation of CAI, but also by companion chatbots such as Replika [65] or social robots [10]. Although such features do not mislead users into exchanging machines for humans, they are designed to stimulate responses that may lead to specific outcomes. Three defining attributes of banal deception include (1) the activation of inherent human tendencies such as agency attribution or personification, (2) the shaping of user reactions and outcomes, and (3) the potential to deliver value or benefits for the user despite its deceptive underpinnings.

Banal deception is aptly named due to its ubiquity and seeming ordinariness. It is a part of everyday interactions with AI technologies, leveraging inherent psychological responses to facilitate particular outcomes. However, it should not be dismissed as insignificant due to its potential to shape user experiences and actions in powerful ways [4, 44].

[2] This is not unlike how theatre or cinema works where audience members suspend their disbelief and achieve Aristotle's catharsis through fear and pity [69].

Strong deception corresponds to the situation envisioned in the Turing Test when a human is led to believe that their conversation partner is a human. However, it is actually a computer programmed to lead conversation in a chatroom. However, a similar situation is relatively rare to be found in the most common interactions with machines programmed to communicate with humans. The most common experiences with contemporary AI systems do not entail this kind of strong deception: users, for instance, are usually well aware that Alexa or Siri are not real persons but just AI applications. This does not mean, however, that deception is absent. Banal deception evoked by different elements of characterisation in the assistant's voice and in the conversation inputs can still exercise a powerful influence on the interaction. These banal deceptions can offer pragmatic benefits by enabling seamless integration of AI tools into domestic or professional settings, facilitating playful interactions, or helping users become accustomed to the tools' functionality.

In contrast to 'strong' deception, banal deception can have, at least potentially, a value for the user. The fact that elements of the characterisation of CAI invite users to activate social habits and behaviours when interacting with CAI, for instance, may have pragmatic benefits. For example, it can improve users' disposition to integrate these tools into their domestic and professional environments, or it may provide occasions for playful interaction that helps users experiment and become accustomed to the functioning of these tools.

This potential for functionality makes the mechanism of banal deception of particular relevance for HCI applications to communicative AI—i.e., AI technologies programmed to enter into communicative interactions with users [26]. While it can benefit users, the dynamics of banal deception can also bear risks and create problematic issues. Moreover, as mentioned above, the distinction between banal and strong deception is not binary, any single AI system can have cases and dynamics that oscillate between banal and strong deception: the projection of empathy stimulated by voice assistants and robots, for instance, could be activated for commercial advertisement and political propaganda, raising ethical questions [18] (see Fig. 1).

Figure 1 illustrates the continuum between banal and strong deception. On the left end, "banal" deception corresponds to CAI, such as Google Assistant, in which features including the humanlike voice invites a degree of anthropomorphisation, helping users to more easily integrate the technology into their everyday lives and environments. At the middle of the continuum, companion chatbot Replika represents a case where the impression of sociality created by the chatbot invites a higher degree of projection and emotional engagement from the part of the user; however, the user is still invited to maintain the difference between Replika and a human person. On the right end, "strong" deception corresponds to cases such as Google Duplex, a project that aimed to use CAI to make phone calls in which the assistant was programmed to deceive interlocutors into believing it was a real person.

It is important to note that Google Duplex is not a totally different technology from Google Assistant: they are both CAI, designed by the same company using similar resources and technologies, yet different design choices and applications result in different deceptive outcomes [46]. In fact, the boundaries between strong and banal deception are not rigid. Rather than a binary opposition, the categories of strong and banal deception sit on a continuum, which in certain cases might make it difficult to clearly categorise deception within a specific AI system. Social robots or companion chatbots such as Replika, for instance, are located at a midpoint between banal and strong deception: their appearance of sociality enhances their capacity to offer companionship to their users, therefore adding a potential value, but at the same time the emotional character of the relationships that ensue raises significant questions in terms of users' capacity to keep full control of the experience [27]. This, however, does not undermine the usefulness of the distinction between banal and strong deception for HCI researchers and practitioners,as we will show in the next section of the paper, if combined with protocols from Human-Centered AI, positioning an AI system within the banal-strong deception continuum provides crucial analytical resources to assess potential benefits and risks of deceptive dynamics in AI. The framework of Human-Centered AI has the potential to provide helpful orientation and practical guidelines on how to implement practical and functional benefits of banal deception while avoiding the more problematic implications of this phenomenon and guaranteeing fairness and explainability (i.e., [15]) for the user as well as clearer audit trails supporting accountability when things go awry.

3 Human-Centered AI and the Question of Deception

Human-Centered AI (HCAI) synthesises Artificial Intelligence (AI) algorithms and human-centred thinking. This approach combines research on AI algorithms with user experience design methods to shape technologies that amplify, augment, empower, and enhance human performance ([60], p. 4). Researchers and developers for HCAI systems value meaningful human control, putting people first by serving human needs, values, and goals (c.f., [51]). This means that HCAI is occupied primarily not with technologies that do the work or replace the work of human agents but instead with technologies that enable humans to carry out their work far more effectively than they were able to do before. Systems like the Internet, email, digital navigation,

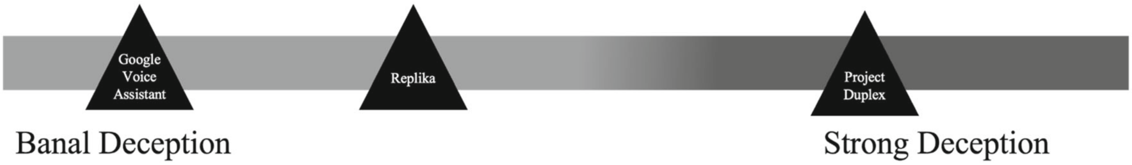

Fig. 1 Continuum of deception

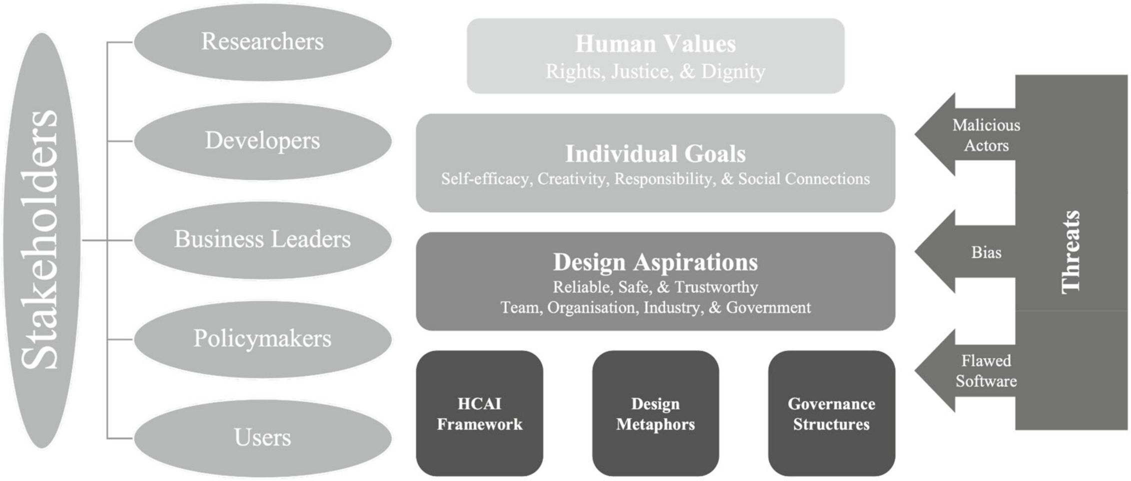

Fig. 2 Human-Centered AI [59]

and digital photography are examples of technological processes that enhance human abilities rather than substituting them [59] (Fig. 2).

The HCAI framework begins not just with the technology in question but also with human values. It asks how technology can support human rights, as defined by statutes such as the United Nations Declaration of Human Rights [83]. Other questions like how we provide equal access to justice and equity across different groups (i.e., minority groups) and how we can support human dignity. These are the driving questions that underlie the HCAI framework for how engineers can design AI systems *for* human values rather than relegating them to afterthoughts or sidelining them altogether.

The HCAI framework identifies and aims to counteract potential threats that may jeopardise the human-centered character of AI technologies. This includes the agency of *malicious actors* such as adversaries, criminals, spammers, terrorists, and oppressive political regimes, which must be accounted for to determine how they can appropriate the technology and misuse it [60]. We must also be aware of the *bias* that emerges from both our conscious and unconscious activities. This involves exploring means of reducing bias as much as possible, becoming aware of bias, and determining that the data used is unbiased and that such is used in an unbiased way. Finally, we need to ensure that the software we build is actually *reliable, safe*, and *trustworthy* (RST), ensuring that it has fewer bugs and can help achieve our goals.

The possibility of deception represents one of the critical elements of risk that approaches to AI design and ethics need to consider. While traditional approaches to this problem have privileged a binary distinction by which users are either deceived or not deceived, the concept of banal deception reframes the question in a way that is more sensitive to the fact that deception is not opposed but closely aligned to human perception [31, 91]. This means that deception, in other words, is a structural component of human experience that can have negative and positive outcomes. In this sense, acknowledging that deception is also "banal" [44] is in continuity with an approach that *re*-places humans at the very centre of reflections and practical work in technology and design, such as HCAI.

The HCAI provides the tools to frame the implications of banal deception in the context of AI to permit desired outcomes. In particular, HCAI provides a unique framing on meaningful human control, how designers employ metaphors for understanding systems, and the various nexuses for governing these systems over their lifespan. The following three subsections outline how HCAI frames these elements concerning the issue of deception.

3.1 HCAI, Deception, and the Issue of Control

One of the characteristics usually attributed to deception is that it implies an asymmetry between the actors involved:

the deceiver knows something that the deceived does not know [8]. Yet scholars have pointed out that each actor engaged in deception is active, including the deceived party [82]. The concept of banal deception underlines that users can exploit their liability for deception in active ways and that design value can be accrued from such a dynamic. For instance, spectators of a fiction movie profit within a safe and human-centred environment from the impression of reality produced by the representation on the screen, as it helps them to participate emotionally in the events depicted in the movie. Although traditional approaches to deception could dismiss this situation as different from deception, there is value in acknowledging the continuity between "banal" and "strong" deception in such contexts. The same effect of reality that contributes to the appeal of audiovisuals as a form of entertainment and information is also activated by openly deceptive uses of this technology—such as deepfakes, which profit from the statute of the photographic image as evidence.

To ensure that banal deception mechanisms remain reliable, safe, and trustworthy, design protocols need to be developed that allow users to maintain control over their experience and interaction. As highlighted by HCAI [60], traditional paradigms of discussing machine automation placed automation and human control on a spectrum where each existed at opposite poles (e.g., Fig. 3).

This means that having more of one would lead to less of the other, a zero-sum game [56]. This simple and compelling model remains structural in the thinking of the design of autonomous systems (e.g., [33, 63]). However, recent developments in design and HCI have shifted away from this dilemmatic paradigm of *balancing* autonomy and control to *ensuring* how more automation can lead to more meaningful human control (i.e., see [61]). This entails that machine automation and human control levels are not on the same axis but distinct axes (see Fig. 4) [57].

Although the upper-right quadrant in Fig. 4 is not always possible to attain, it is the goal of the HCAI framework. Most RST systems can be found on the upper-right side of the matrix, whereas systems with highly predictable tasks, like automatic lane assist systems in many modern cars, can be found on the lower-right side. What distinguishes the two right-side quadrants is that the upper-right side is required for those types of automatic systems that involve complex behaviours with highly dynamic contexts of use.

An excellent example of this would be the current debate on lethal autonomous weapons systems. Such systems would be feasibly placed in war theatres with highly complex and changing environments and need to make similarly complex and vital decisions. Because such decisions have lethal consequences, it would be ethically problematic and practically dangerous to deputise them to machines. Hence a high level of human control is required (see [75, 76, 78]). However, the context of use can become ever more standardised over time, allowing a system to have even greater control over that context, thus permitting greater automation over those tasks to occur [57]. For example, elevators and the digital cameras found in most modern smartphones are based on numerous highly complex AI algorithms despite their seemingly simple behaviour and high levels of human control [59]. There are dangers, however. There can be too much automation, like in the famed Tesla crash of 2016 [92] and the Boeing 737 MAX's MCAS system [47]. Additionally, too much human control can sometimes lead to egregious human errors like those which occur with morphine drips for patients [73].[3]

The balance between computer automation and human control is also a crucial issue for CAI. For instance, features such as the wakeword (e.g. Hey Siri) ensure that the user exercises control over the interface, as the tool is activated only upon the user's decision. In comparison, companion chatbots such as Replika are programmed to communicate also when not asked to, for instance, by asking the user how she feels or by expressing willingness to talk; this is meant to stimulate users to interact more often with the chatbot, but locates Replika in the lower right corner, as a high level of automation is coupled with less control by the user [27].

The two-dimensional framework used to distinguish the approach of HCAI through the coordinates of human control/computer automation [57] can be fruitfully adapted to applying the notion of banal deception. As shown below, HCAI should not only comprise banal forms of deception and avoid strong deception altogether,it should also couple banal deception with high levels of user control. This means that HCAI is firmly placed in the upper-right quadrant of Fig. 5.

Our paper underscores the existence and implications of deception within HCI. Deception, in this context, can range from banal to strong deception, with each having its own dynamics and implications. Here, we propose outlining an analytical tool to assess these types of deception within HCI systems, alongside the level of human control involved. To start, banal deception is characterised by subtly obscuring the nature of the system, yet in a way that does not forcefully harm or negatively impact the user experience. An example could be an audiovisual that enhances the user's emotional engagement without misleading about the reality it represents. Strong deception, on the other hand, is when the system intentionally gives a false impression of its capabilities or purpose, such as the infamous deepfake technology that misrepresents reality with an intention to deceive.

Our analytical tool will include a set of questions to determine:

[3] For further examples of systems in each of the quadrants, as well as examples of excessive automation and excessive human control, see Shneiderman [59] and Shneiderman [60].

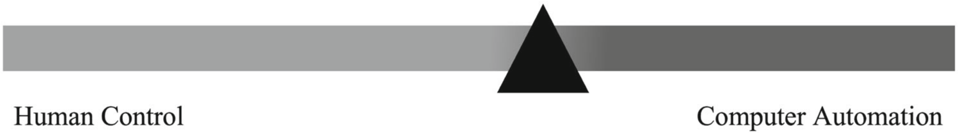

Fig. 3 One-dimensional thinking suggests that designers must choose between human control and computer automation—source: [57]

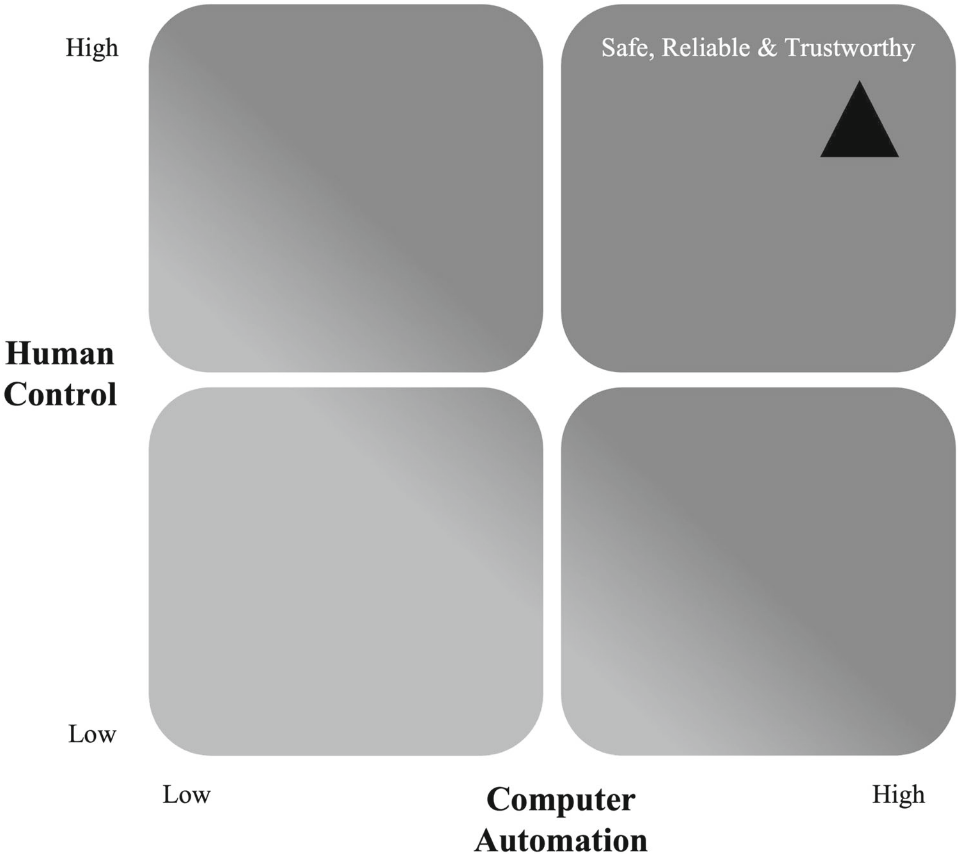

Fig. 4 Two-dimensional framework with the goal of being Reliable, Safe, & Trustworthy, achieved by a high level of human control and a high level of computer automation (black triangle). Source: [57]

(1) The level of banal deception: To what extent does the system benignly obscure its true nature? Are there clear disclosures about the system's functionality and purpose?
(2) The level of strong deception: To what extent does the system intentionally mislead about its capabilities or function? Are there instances where the system represents itself inaccurately or falsely?
(3) The level of human control: How much control does the user have over the system's functionality and operation? Can the user control when and how the system operates?

These questions are meant to operationalise the concepts of deception and human control within HCI. For instance, a system that allows the user complete control over its operations and clearly discloses its nature and capabilities would score high on human control and low on both types of deception.

Our goal is to create an analytical framework that provides measurable and precise characteristics of the concepts involved. This framework makes it possible to approximate place a system on a multidimensional scale based on these three aspects: banal deception, strong deception, and human control.

In defining banal and strong deception, necessary elements include a clear understanding of the system's nature and purpose and how these are presented to the user. Is the system honest about its capabilities and function, or does it intentionally mislead? Furthermore, to classify an HCI system's level of banal or strong deception, minimal questions would include inquiries about the system's representations, disclosures, and the user's understanding of these. Therefore, the analytical tool we're proposing aims to create a structured, multi-dimensional framework for HCI, which would help to accurately measure and place a system on the HCAI-based scale.

To ensure the HCAI framework's utility and reliability, it is vital that banal deception aligns with high user control levels. This means that the system remains transparent about its operations while still enhancing the user experience. Strong

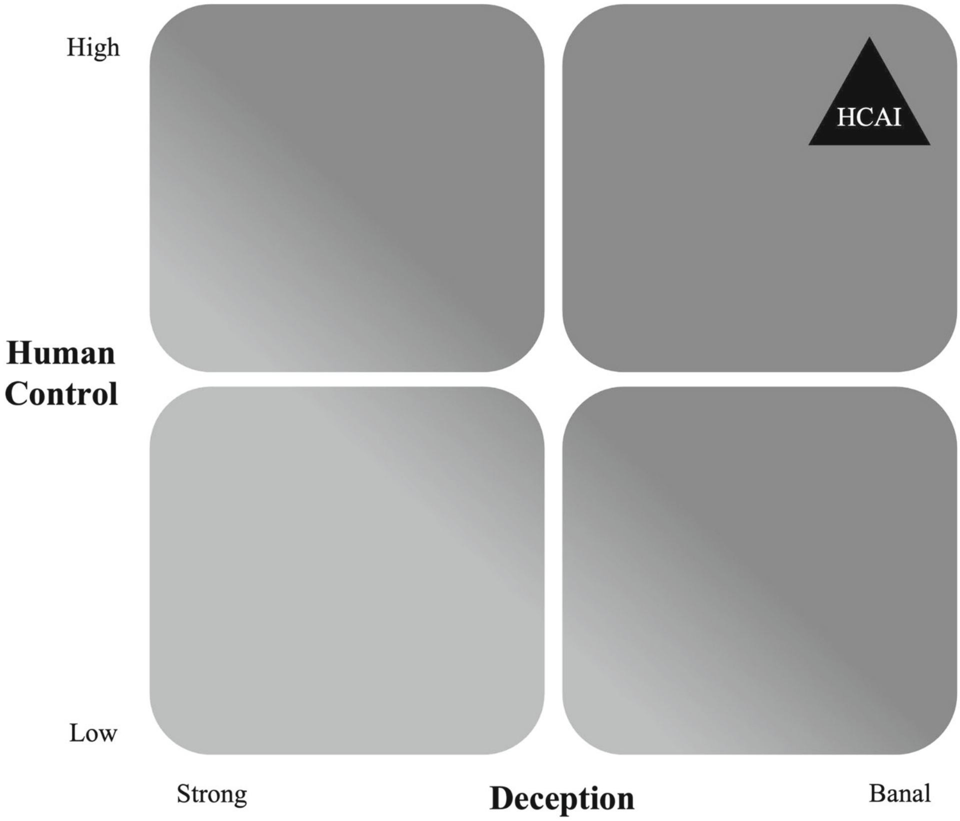

Fig. 5 Two-dimensional framework of HCAI and the continuum of computer deception

deception, on the other hand, should ideally be minimised or eradicated completely.

Focusing on the case of Conversational AI (CAI), we posit various hypothetical scenarios representing different combinations of deception and control. The placement of CAI features in each of the four quadrants of Fig. 5 is based on the type and degree of deception and control. However, it's important to note that these placements are not fixed, and the perceived value for the user may vary depending on individual circumstances and needs.

Strong deception can easily be activated in CAI and other conversational agents; in fact, these can potentially be programmed to pass as humans, as Alan Turing [75] already envisioned for the design of the imitation game or Turing Test. In contrast, the computer passes the test if it deceives a human interrogator into believing it is human. Therefore, in the lower left quadrant, we find CAI purportedly programmed to hide their computational nature and defy identification as computational agents, combining strong deception with a low level of human control. Although one may point out that this will be difficult to achieve, the experience of textual-based chatbots shows that users can be tricked even by relatively simple systems in specific situations and contexts [44].

In the upper left quadrant, we find applications that combine a high level of human control with the delivery of strong deception. An example is Google Duplex, a project pursued and then discontinued by Google in 2018. Google aimed to develop a new functionality of its voice assistant to make phone calls on behalf of the user, for instance, to book a table at a restaurant. Since a restaurant might decide not to take automated reservations by phone, Google made explicit efforts to make Duplex sound more realistic, such as programming the assistant to pause and hesitate at specific moments during the conversation, mimicking human conversation [48]. This move, which was criticised as "straight up, deliberate deception" [74], combined a high level of human control—since the user can ask the assistant to make the call on their behalf -with the delivery of strong deception. Google Duplex can, therefore, be located in the upper left quadrant.

CAI can also be programmed so that banal forms of deception are coupled with low levels of human control, corresponding to the lower right quadrant of Fig. 5. These are the cases when banal deception ceases to be human-centred and is mobilised as part of a design framework that does not enhance but instead limits human control. Examples of such dynamics include CAI, such as Alexa always listening while appearing inert and defying the expectations of users [36]. More broadly, CAIs are also web interfaces that provide access to information available online,concerns have been raised about the fact that they may direct users predominantly to the cloud services of the respective companies, further eroding the distinction between the web and proprietary cloud services and limiting users' control over their navigation of the web [89], p. 120, [46].

To meet recommendations advanced by the HCAI framework, banal deception should be combined with a high level of user control, i.e. the upper right quadrant of Fig. 5. This corresponds to a situation in which banal deception adds value to the user while enhancing or, in any case, not jeopardising control. For instance, using a humanlike voice rather than a synthetic-sounding voice has facilitated the introduction of CAI in everyday life and domestic environments, as it is perceived as more natural by users [40]. Therefore, this design choice employs positive banal deception to impact user experience while maintaining the level of control and, consequently, the human-centred orientation of the software.

To clarify our definitions, 'banal deception' refers to subtle forms of deception that are not immediately apparent to the user and may even enhance the user experience. 'Strong deception,' on the other hand, refers to overt attempts by the AI to mimic human behaviour and trick the user into believing they are interacting with a human. 'High control' refers to situations where the user maintains a significant degree of control over the AI's actions, while 'low control' indicates that the AI operates more autonomously.

A user-based evaluation could be performed to assess these theoretical placements. This could involve asking users to interact with each type of CAI and provide their assessment on the level of perceived deception and control. These assessments could then be compared with the expected placement in the framework. Such an evaluation would provide insights into user perceptions and validate the assumptions made in the framework. However, as this type of user-based evaluation has yet to be performed, the potential placements of different CAIs in the framework remain theoretical.

3.2 Incorporating Perspectives from Persuasive Technology and Digital Nudge Frameworks

Recent research in persuasive technology and digital nudging provides valuable insights that can complement the discussion on deception in AI. Persuasive technology involves designing interactive systems to change users' attitudes or behaviours [22]. Digital nudging, a concept derived from behavioural economics, employs subtle design features to guide user behaviour in digital environments without restricting options [72]. These frameworks can be particularly relevant to HRI. Persuasive technologies have been effectively used in applications ranging from health behaviour change to environmental sustainability. Similarly, digital nudges can be integrated into AI systems to promote beneficial behaviours without misleading users.

One pertinent example is using digital nudges in conversational AI systems to encourage healthy lifestyle choices. These nudges can be designed to appear as casual suggestions from a voice assistant, thus employing banal deception to foster positive outcomes. This aligns with the concept of benevolent deception discussed earlier in this paper, where the goal is to enhance user experience while ensuring that the user remains in control.

3.3 Design Metaphors

Scholars have stressed the close relationship between interface design and deception. Lori Emerson argues that human–computer interaction systems "also inevitably acts as a kind of magician's cape, continually revealing (mediatic layers, bits of information, etc.) through concealing and concealing as it reveals" (2014, p. x). Similarly, Wendy Chun [12] observes that an interface builds a paradox between invisibility and visibility, making the system "transparent" to users at the surface level but simultaneously opaque at the level of its inner functioning. Graphic user interfaces, for instance, "offer us an imaginary relationship to our hardware: they do not represent transistors but rather desktops and recycling bins" [12, p. 43]. Creating digital interfaces entails the use of a range of metaphors that provide "something known and of our making, or at least at our choosing, that we put stand for, and so to help us understand, something unknown and not of our making" ([49], p. 30). Corresponding this to banal deception, the choice of design metaphors can positively impact users' experience [35]. However, to achieve this, the interface needs to activate appropriate design metaphors that can enhance human control even with a high level of computer automation in the underlying systems.

The metaphors chosen to make AI systems accessible to users are never neutral. As Lakoff and Johnson [34] famously observed, using specific metaphors to describe things and events always results in orienting attitudes towards them and guiding future actions. This means that designers need to take great care and attention to the choice of metaphors to describe computing systems and the choice of metaphors incorporated within interfaces. For instance, a failed attempt to develop social interfaces, such as Microsoft's Ms. Dewey featured a video of a woman commenting on a user's search engine queries, constructed sexed and racialised representations of interface "assistants" that ultimately impacted the functionality and fairness of such systems [70].

HCAI emphasises developing better metaphors to describe and empower human-centered technology. Despite the popularity of speaking about 'social robots', Shneiderman [60] observes that many of the existent enterprises that have ventured into creating social robots, often in the form of humanoid systems, have failed and proposes that such metaphors should give way to a more innovative paradigm of thinking about 'super tools'. Rather than simulating human characteristics, tools extend human capabilities, such as in

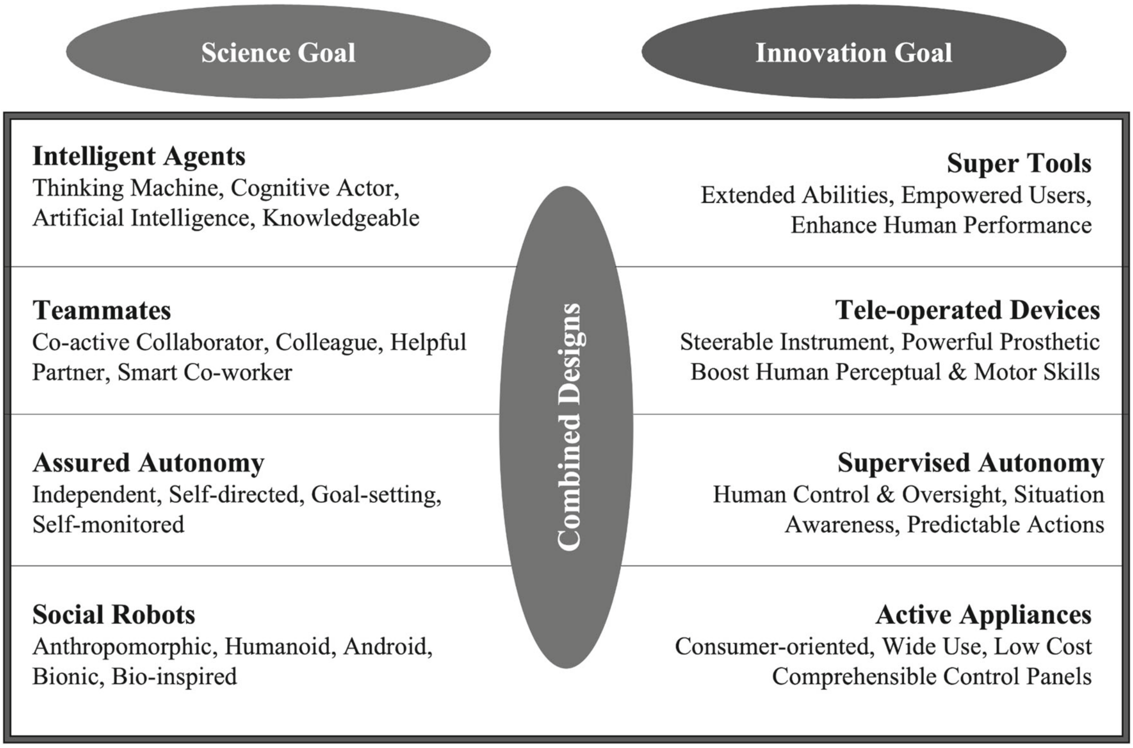
Science Goal
Innovation Goal
Intelligent Agents
Thinking Machine, Cognitive Actor, Artificial Intelligence, Knowledgeable
Super Tools
Extended Abilities, Empowered Users, Enhance Human Performance
Teammates
Co-active Collaborator, Colleague, Helpful Partner, Smart Co-worker
Tele-operated Devices
Steerable Instrument, Powerful Prosthetic Boost Human Perceptual & Motor Skills
Assured Autonomy
Independent, Self-directed, Goal-setting, Self-monitored
Supervised Autonomy
Human Control & Oversight, Situation Awareness, Predictable Actions
Social Robots
Anthropomorphic, Humanoid, Android, Bionic, Bio-inspired
Active Appliances
Consumer-oriented, Wide Use, Low Cost Comprehensible Control Panels
Combined Designs

Fig. 6 Design Metaphors. Source: Shneiderman [58]

teleoperated devices and active appliances. Hence, to arrive at the SRT system, the metaphors used when referring to these systems must make a similar shift. In Fig. 6, we can see that the innovation goal in shifting metaphors refers to devices with supervised autonomy to ensure human control and oversight. Supertools can include devices like digital camera tools, which permit users to choose and make the photos that they want to make. Navigation systems allow users to select between route choices to a destination. In these systems, one becomes the beneficiary of AI systems' power while also maintaining predictive and prospective control. The user is allowed to choose *before* the system acts.

The idea of active appliances is best illustrated by devices like rice cookers, coffee makers, dishwashers, dryers, etc., which have hidden behind their seemingly simple interfaces a host of sensors and automation that let the user make choices regarding their function. Even the once highly automatic pacemakers are shifting to become more human-centred by permitting more levels of human control over their behaviours via apps that allow the user to control and collect the data and alter the pacemaker's parameters, enabling healthcare experts to access the data [71].

The case of CAI well exemplifies how the careful selection of design metaphors suggested by HCAI can ensure that the negative implications of deception are prevented, and that banal forms of deception remain beneficial to users. The lively debate about representations of sex embedded in the characterisation of CAI is a case in point. Hints at sex identity include the female name chosen for services such as Alexa or Cortana and the sexed voice. Many of these services were initially launched with the female voice as default, creating concerns about the overlaps between sex stereotypes and the design metaphor of the clerical worker through which assistants are offered to users [90]. The extent of this problem can be fully considered if examined through the lens of banal deception: the effect of anthropomorphisation is facilitated both by the sexed character and the metaphor of the assistant or servant [90], which are meant to "give people something to acclimate to" (as a developer involved in the design of CAI conceded, see [93], p. 117). The problem emerged due to the interaction between these two forms of representation, which activated associations between sex and hierarchical roles. HCAI provides a potential solution to the conundrum. In terms of user control, the ability of the user to select their own name and sexed voice can place the user in control of the sex representations activated by the assistants,in terms of design metaphors, moreover, conceptualising CAI as supertools counteracts the risk of creating implicit connections with professional figures.

While traditional understandings of deception do not consider the potential risk of design metaphors in creating misleading and unfair conceptions about AI tools, the banal deception framework provides the conceptual means to identify metaphors as an area in which the fine line between banal and strong deception can be crossed, or where banal deception mechanisms can cease to be placed at the service of the user. Such an agenda is ideally aligned with HCAI's efforts to counteract metaphors that humanise CAIs, such as the metaphor of the assistant or servant. The metaphor of an instrument or tool can help ensure that elements of characterisation of CAI, such as sex, provide value to users. Elements of personification such as sex/personality have been shown to benefit users' capacity to successfully integrate CAI within their everyday lives and environments [38] while avoiding problematic implications and incorporating stereotyping practices within the tool's design.

3.4 Governance Structures

The potential and the implications of AI technologies are often discussed in terms of their technical characteristics and what this allows us to achieve, overlooking the role of the broader institutional structures within which these technologies are designed and situated. The HCAI framework provides a useful corrective to this problem, placing the institutional and professional structures that inform the governance of specific AI tools at centre stage. In particular, the HCAI framework identifies a three-level structure concerning the governance structures.

Starting from the innermost elliptical, this domain encompasses a large portion of software engineering teams. Unsurprisingly, these teams are responsible for engaging in the actual technical practices that each project requires. Consequently, these teams form part of the broader organisation (middle elliptical). Here, approaches concerning creating and implementing a "safety culture" impact the project teams. The most prominent domain is where independent oversight boards review organisations in the same industry (the middle elliptical). This permits oversight bodies to gain a more comprehensive understanding of that domain and facilitate disseminating successful practices [57, 58].

Although corporations and governments regularly boast about their concerted efforts to ensure that stakeholder needs and values are met by their practices, these are often sacrificed or sidelined when leaders are forced to make tough choices concerning power and money. More often than not, leaders may choose their own personal needs or cede to political pressures and stockholder expectations if pitted against stakeholder needs [32].[4] Despite effective Human Rights and Corporate Social Responsibility campaigns, the decisions concerning technology design are often made by software engineers, managers, and review boards. Consequently, it is them who have to be guided by principles and operational recommendations. If we consider transformative technologies like AI, this becomes of particular importance, given that both government leaders and corporate managers may require information as to the available range of options that are open to them.

What Fig. 7 aims to do, then, is to highlight simple yet practical ways forward. These steps are built on familiar existing practices; however, they must also be modified to allow less traditional and more transformative technologies like HCAI systems. Fundamentally, they are geared towards clarifying who the actors are and, thus, who is consequentially responsible for those actions (c.f., [6]). Naturally, each of these ideas requires research and testing to determine if they are indeed effective. Still, they are directed to designing HCAI systems that are *reliable, safe*, and *trustworthy*. Doing so will bring boons to individuals, organisations, and society [37, 88]. Because this governance structure outline is a point of departure, newer frameworks will undoubtedly be required as technologies develop or when market forces and public opinion guide the products and services that become successful.

For the organisational and business side, there are both challenges and opportunities. Part of this is that AI is global, highlighting the importance of considering specific values that are culturally and socially specific. This implicates the need to contextualise AI within the cultural and social environments in which they are designed, deployed, and pervasive. This is further complicated by the various stakeholders who use the conversational AI systems and those involved in their design. This is rarely undertaken from beginning to end by a single firm or individual, but many hands muddy these systems and the labour dedicated to their creation.

For example, we can use the popular conversational AI system Amazon Alexa as our case study. The two primary domains in which the system is used are (1) the industries that use Amazon's underlying system to create what are called "Skills" (i.e., addons) and (2) companies using Amazon Alexa's API to build things tailored to them. The industries that use Amazon's underlying system to create skills are also further stratified into the industries that hire other agencies to build these skills and add-ons (i.e., building the underlying AI system of these add-ons) and the voice designers and product individuals who influence how it is used in the world. These actors often do not directly work side-by-side and are extracted from their labour. For this reason, traditional co-creation methodologies like participatory design, universal design, and inclusive design are not well equipped to address the dissociated nature characterising the conversational AI landscape.

Figures 8 and 9 best illustrate the fragmented nature of the CAI landscape. Figure 8 describes the various actors concerning CAI that the typical person is most familiar with, the

[4] The Z-Inspection®, discussed below, provides a explciit means of addressing value tensions, particularly those of economic value. See Zicari [96].

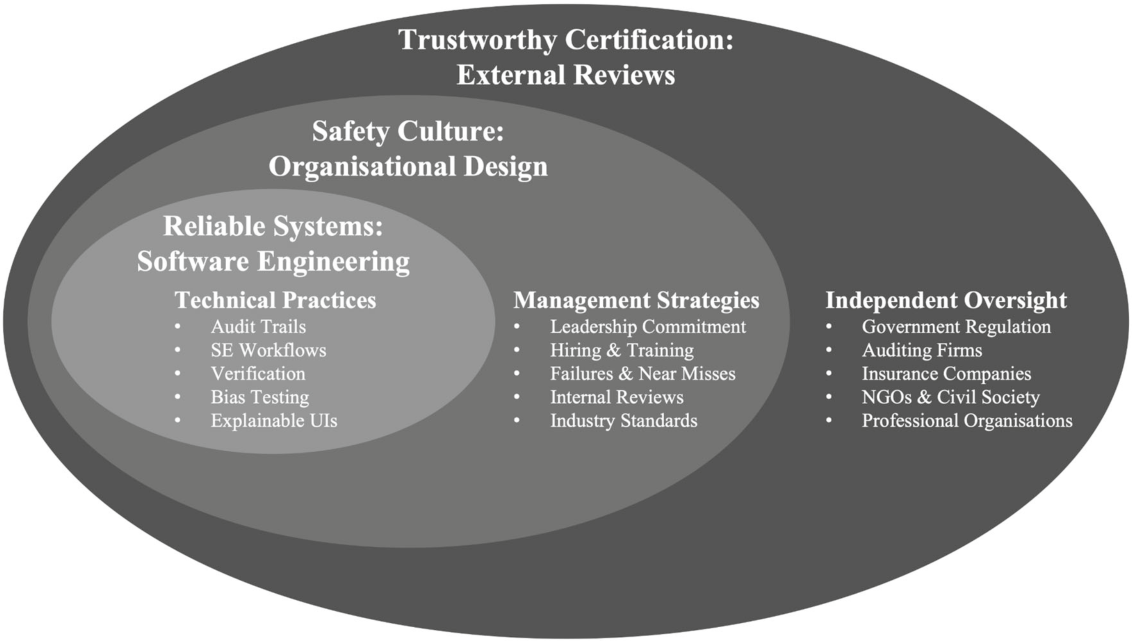

Fig. 7 Governance structures for human-centered AI have three levels: reliable systems based on software engineering (SE) practices, a well-developed safety culture based on sound management strategies, and trustworthy certification by external review. Source: [57, 58]

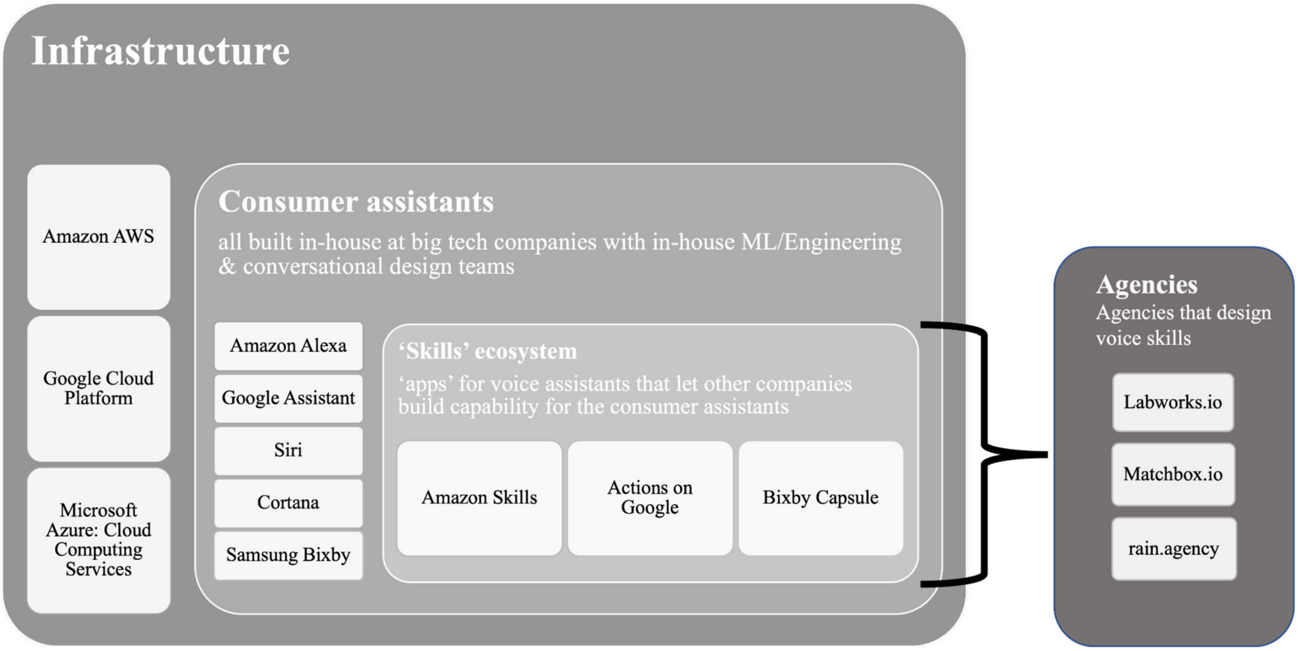
Infrastructure
Amazon AWS
Google Cloud Platform
Microsoft Azure: Cloud Computing Services
Consumer assistants
all built in-house at big tech companies with in-house ML/Engineering & conversational design teams
Amazon Alexa
Google Assistant
Siri
Cortana
Samsung Bixby
'Skills' ecosystem
'apps' for voice assistants that let other companies build capability for the consumer assistants
Amazon Skills
Actions on Google
Bixby Capsule
Agencies
Agencies that design voice skills
Labworks.io
Matchbox.io
rain.agency

Fig. 8 The consumer side of the CAI landscape

consumer CAI systems that people buy and use in their homes like Alexa and Cortana. The business-to-business domain is the other modality of describing the CAI landscape (Fig. 9). This landscape is described as the CAI systems that companies build for specific applications, such as controlling a robot, ordering a taxi, getting banking information, providing access to information for workers in the field, etc.

What both the two landscapes (Figs. 8 and 9) illustrate is that the people building the core AI/ML technology (i.e., the ASR, NLU and TTS) are usually separate from the people designing the overall experience (i.e., thinking about how conversations will flow). This means that it is rare to find a CAI system built entirely from the ground up by any given team. This implicates the governance mechanisms that are

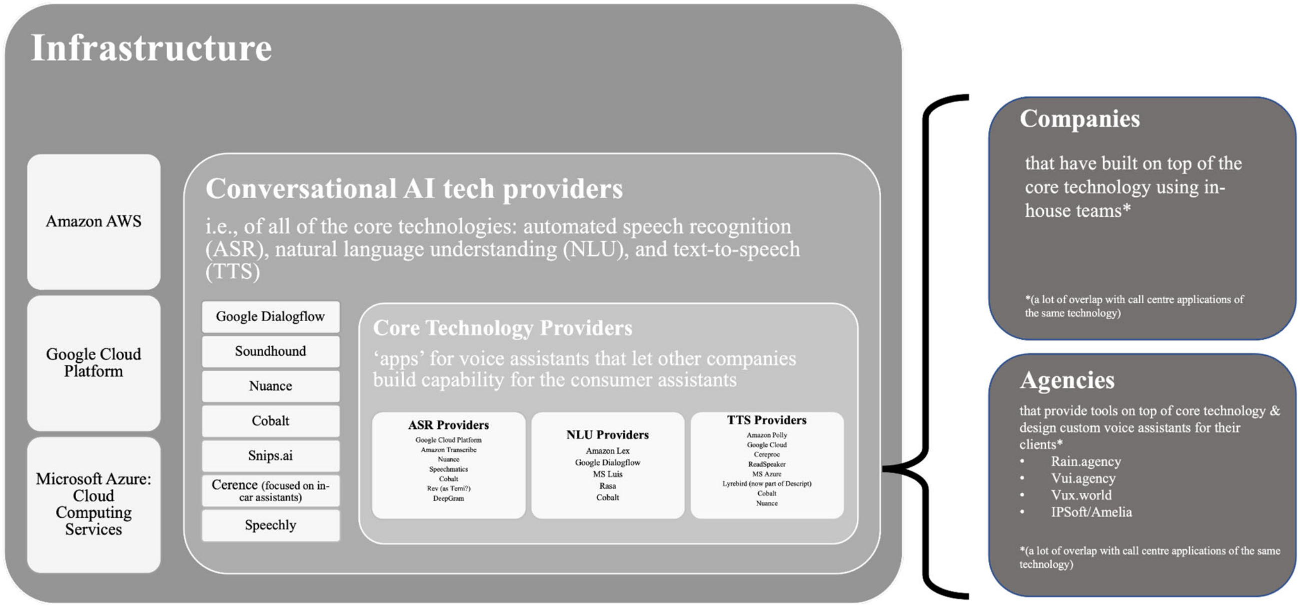
Infrastructure
Amazon AWS
Google Cloud Platform
Microsoft Azure: Cloud Computing Services
Conversational AI tech providers
i.e., of all of the core technologies: automated speech recognition (ASR), natural language understanding (NLU), and text-to-speech (TTS)
Google Dialogflow
Soundhound
Nuance
Cobalt
Snips.ai
Cerence (focused on in-car assistants)
Speechly
Core Technology Providers
'apps' for voice assistants that let other companies build capability for the consumer assistants
ASR Providers
Google Cloud Platform
Amazon Transcribe
Nuance
Speechmatics
Cobalt
Rev (as Temi?)
DeepGram
NLU Providers
Amazon Lex
Google Dialogflow
MS Luis
Rasa
Cobalt
TTS Providers
Amazon Polly
Google Cloud
Cereproc
ReadSpeaker
MS Azure
Lyrebird (now part of Descript)
Cobalt
Nuance
Companies
that have built on top of the core technology using in-house teams*
*(a lot of overlap with call centre applications of the same technology)
Agencies
that provide tools on top of core technology & design custom voice assistants for their clients*
• Rain.agency
• Vui.agency
• Vux.world
• IPSoft/Amelia
*(a lot of overlap with call centre applications of the same technology)

Fig. 9 The business-to-business CAI landscape

supported by the HCAI framework. In particular, it begs the question of how oversight can be achieved in such a fragmented landscape. This problem of many hands makes it harder to track the governance structures. However, where lies the problem also lies the solution. This fragmented CAI landscape allows us to discuss what can be done to achieve what HCAI argues to be good governance. This is not particular to CAI; big companies build infrastructures, and it is up to other companies to use these infrastructures to develop their own specific tools. This paper does not aim to solve this problem here; however, it does draw attention to the issue that governance of these technologies is not straightforward, and discussing the nature of the CAI landscape is a necessary condition for discussing salient governance solutions.

HCAI, unlike other parallel approaches to technological design, like value sensitive design, is not predicated on a unitary understanding of the technical design landscape. For this reason, HCAI provides us with the modality of understanding how CAIs can be human-centred despite their many constituent parts and actors being entirely dissociated from one another. HCAI provides the vehicle in which more specific processes particular to AI design can be delivered across various boundaries.

We propose that the HCAI framework is particularly apt for ensuring that socially beneficial banal deception is designed *for* via the appropriation of the Z-Inspection® process for trustworthy AI design [97]. This is for at least two reasons. Firstly, the Z-Inspection process provides a unifying methodology across all three levels of the HCAI governance structure (Fig. 7). Secondly, the process has proven to be efficacious in achieving trustworthy AI using real-world case studies via co-designing all the identified stakeholders that HCAI delineates (Fig. 2), see [98].

The process begins by determining the specific preconditions of the AI technology in question. Concerning CAIs, for example, this entails looking at the *legal admissibility and the absence of conflict of interest*. This would mean ensuring congruency between the various stakeholders and acknowledging potential malicious actors and threats like strong deception (i.e., 'Threats' in Fig. 2). This is followed by creating an *initial team of multidisciplinary experts* who function as stakeholders, internal reviewers, and industry-standard professionals and consulting *best practices* to learn from and avoid 'failures and near misses' (see 'Safety Culture' in Fig. 7). This multidisciplinary team then creates a *protocol (log) of the process,* which is used to record information about the inspection process over time, ensuring that *audit trails* and *verification* are possible (see 'Reliable Systems' in Fig. 6) as well as to provide the necessary resources for *independent oversight* to be possible (see 'Trustworthy Certification' in Fig. 7). The process then aims to define both the *boundaries and context of the assessment*. This entails considering the entire ecosystem or landscape (like that of the CAI landscapes in Figs. 8 and 9) and understanding that such a landscape plays a vital role in defining the boundaries of the assessment. The *metaphor* here is that the landscape is less a field to be played upon, but one that has agential qualities to it, one that co-constitutes the design of the technologies and systems in which they are embedded. Aligning directly with HCAI's understanding of governance, Z-Inspection® understands these landscapes as "a set of sectors and parts of society, level of social organisation, and stakeholders within a political and economic context" [60, 95, 967; c.f.,

In bringing together the various stakeholders, HCAI via Z-Inspection® aims to *Identify Ethical Issues and Tensions*. This begins through metaphor work, viz., the co-construction amongst stakeholders of socio-technical scenarios. These involve scenarios where the possible uses of the AI system in question can be imagined in various contexts and applications (e.g., using a male/female-voiced CAI in a children's hospital). These scenarios are used to highlight potential ethical issues and tensions. Here, the consensus amongst stakeholders is the goal to determine value priorities and potential solutions to moral overload (see [84]). Likewise, this exercise aids stakeholders in concept building.

Consequently, stakeholders can "build a shared understanding of key concepts that acknowledges and resolves ambiguities, and bridges disciplines, sectors, stakeholders and cultures" [95, 96]. These ethical issues and tensions can then be *mapped onto the ethical categories established by the EU´s Guidelines for Trustworthy AI* (see European [20]). The appropriation of these guidelines aligns with HCAI's need for reliability at the technical level, the industry standards at the organisational level, and the government regulation at the higher certification level (see also AI Act) [50]. Once a path(s) is defined, the means to execute that path are outlined, and feedback is provided to the system's designers. At this point, ethical maintenance begins, and any emerging ethical issues or tensions are addressed in situ. This latter auditing and monitoring phase is necessary to ensure that the AI system fulfilled the Trustworthy AI requirements when deployed and continues to do so over its lifecycle. This is particularly important when considering deception. As we mentioned, deception is not either banal or strong but exists on a continuum which changes based on the context. Different applications can change the degrees of deception. For example, the use of a humanised voice can be helpful to users to be able to integrate the CAI system into their everyday life. Such a system would ideally not be unfamiliar or uncanny compared to a metallic or synthetic-sounding voice. However, the same CAI can be used to create strong deception, i.e., to make a user think that they are talking to a person rather than a software system. The issue is that because deception is so integral to HCI technologies powered by AI, governance structures and scrutiny are needed at multiple levels to capture the various levels of where these varied technologies originate. The same voice-based CAI has very different deception implications when applied in the line at a bank and when placed in a kindergarten class. The former may be banal, while the latter may be effectively strong. Therefore, this requires audit trails and validation to ensure one does not become the other without clear lines of tracking and accountability.

This process entails a high bar regarding what can be considered reliable, safe and trustworthy AI; however, it ensures that AI systems' emergent behaviour does not go unchecked over their lifecycle. Fundamentally, this higher standard permits greater autonomy and, consequently, greater and more meaningful forms of human control over these systems. Concerning the CAI landscape, this approach does not aim to 'bring together' the various actors that constitute the fragmented landscape in the sense that it does not seek to frame the landscape as homogenous and thus governable from that perspective. Instead, HCAI, as a vehicle for processes like Z-Inspection®, not only permits but supports the creation of reliable, safe, and trustworthy CAI systems across all levels of abstraction[5] by targeting each level rather than a unitary top-down approach that would be, at best, ineffective if not outright impossible to achieve.

4 Conclusion

This paper has explored the connection between various forms of deception in human–computer interaction, particularly how AI might facilitate deceptive outcomes. Likewise, it explored the distinction between strong deception and banal deception and how that has implications for the design of conversational AI (CAI) systems. Although deception of AI systems can indeed have negative impacts, they can also be geared towards socially beneficial outcomes. This paper confronts challenges by proposing a framework that distinguishes between different situations and dynamics in which communicative AI may involve deceptive outcomes. More specifically, it proposed an analytical distinction between situations where deceptive mechanisms remain beneficial to users. Doing this permits us to enhance rather than jeopardise human control and value sensitive design by distinguishing between deceptive mechanisms that are beneficial and those that remain risky, problematic, or unfair.

In tackling these issues, this paper argues that deception is not the result of exceptional circumstances but a structural component of interactive AI systems. This is a consequence of human users forcefully projecting their own perceptions, reactions, and biases on AI systems exhibiting communicative behaviour. Designers of these systems can anticipate the dynamics resulting from such acts of projection, thus leading the interaction towards deceptive situations that facilitate desired outcomes. Likewise, this paper also proposes that it is possible to distinguish between deceptive mechanisms and designs that provide value to users and those that are of little

[5] Infrastructure, consumer assistants, and 'skills' ecosystem for the consumer side of the CAI landscape, and the infrastructure, conversational AI tech providers, and core technology providers for the business-to-business CAI landscape.

use and bear significant risks. This is particularly important for designing AI systems since it implies a shift of approaches to tackle the problem of deception in AI.

This paper draws on the Human-centered AI (HCAI) framework, an approach designed specifically for AI systems to ensure that banal deception can be mobilised to achieve better HCAI's goals of producing "successful technologies that augment, amplify, empower, and enhance humans rather than replace them". CAI is drawn on as an example of how HCAI can be used towards this end. At the same time, however, the role of deception in CAI presents significant risks, such as those present in the ongoing discussions about representations of sex, labour, and anthropomorphisation. In this context, there is utility in distinguishing between banal and strong deception, given that it can provide a useful resource for assessing how to design and develop compelling CAI systems that will ensure that new AI-powered interfaces benefit people and businesses by coupling automation with the firm control of human users, thereby meeting HCAI recommendations and empowering present and future users of such systems. If appropriate, the HCAI may prove to be a potential solution for designing CAI systems in what can be described as nothing other than a challenging and fractured design landscape.

The unique aspect of our framework lies in the differentiation between "banal" and "strong" deception, a nuanced perspective that extends beyond traditional binary understandings of deception in AI. By situating deception on a continuum and integrating it within the HCAI framework, we provide a comprehensive tool for assessing and designing AI systems that balance user engagement with ethical considerations. This distinction empowers designers to create AI systems that use benign deception to enhance user experience while maintaining transparency and control. To validate our framework, we propose the following executable method. (1) Select a range of conversational AI systems and categorise their deceptive mechanisms as either banal or strong using our framework; (2) Conduct user studies to assess the impact of these deceptive mechanisms on user trust, engagement, and control, which could involve surveys, interviews, and interaction logs; (3) Implement the HCAI principles to evaluate whether the identified deceptive mechanisms align with ethical standards and enhance the user experience without compromising ethical integrity; and (4) Use the findings to refine the AI systems iteratively, minimising strong deception and enhancing beneficial forms of banal deception.

A significant advantage of the framework proposed here is its applicability to various AI-based technologies. For instance, the tension between banal and strong deception lies at the core of ongoing discussions about anthropomorphisation in social robotics. It is not surprising to presume also that the integration of conversational AI in social robots enhances their ability to engage in meaningful interactions with users. However, this also brings about the challenge of managing deception in these interactions. The relationship between deception in conversational AI and social robots is critical, as social robots are designed to build rapport and trust with users. Deception, whether banal or strong, can significantly impact these relationships.

Social robots that use conversational AI to simulate empathy might create stronger emotional bonds with users, which could be beneficial or harmful depending on the interaction's context and transparency. Empirical studies in HRI have shown that users often anthropomorphise robots, attributing human-like qualities to them, which can lead to deceptive experiences. De Graaf and Allouch [16] highlight that users may develop attachments to social robots, perceiving them as companions rather than machines. This anthropomorphism can enhance user experience but also risks misleading users about the robot's true capabilities and nature.

In addition, we fully acknowledge the richness of Masters et al.'s [39] work and appreciate the multiple types of deception recognised in their study. Our choice of focusing on the overarching categories of strong vs banal deception in the current framework stems from our intention to propose a simplified yet powerful tool for analysing deception within the scope of CAI systems. These broad categories allow us to incorporate a wide range of deceptive practices, thus making our framework more universally applicable and straightforward for initial adoption. However, we agree that the intricate landscape of deception, including distinct modes such as tricking, calculating, confuscating, and more, deserves a detailed exploration within our framework. Understanding how each of these types fits into the matrix of control and deception can provide a more granular perspective and refine the process of evaluating CAI systems. Although complex and challenging, this task presents an exciting opportunity for future research and could lead to a more nuanced analytical tool.

Future research is also required, with specific case studies of this approach, on emerging or already deployed examples of CAI to determine its effectiveness. However, this conceptual analysis has made a case for how the HCAI approach does address some of the structural issues of AI more broadly (i.e., forms of deception) and those of the CAI landscape (its fractured nature). The actual means of realising HCAI may change and evolve as the technologies evolve. Still, this paper aimed to provide a means of framing these issues and the actual means by which these issues have been and can continue to be addressed.

Acknowledgements We would like to thank Ben Shneiderman for the helpful comments and feedback on an earlier draft of this paper.

Data Availability There is no associated data for this manuscript.

References

1. Adar E, Tan DS, Teevan J (2013) Benevolent deception in human computer interaction. In: Proceedings of the SIGCHI conference on human factors in computing systems. https://doi.org/10.1145/2470654.2466246
2. Baker-Brunnbauer J (2021) TAII framework for trustworthy AI systems. ROBONOMICS J Automat Economy. https://journal.robonomics.science/index.php/rj/article/view/17
3. Bench-Capon TJ (2020) Ethical approaches and autonomous systems. Artif Intell 281:103239. https://doi.org/10.1016/j.artint.2020.103239
4. Billig M (1995) Banal nationalism. Sage, London
5. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? Proceedings of the 2021 ACM conference on fairness. Accountabil Transparency. https://doi.org/10.1145/3442188.3445922
6. Burton S, Habli I, Lawton T, McDermid J, Morgan P, Porter Z (2020) Mind the gaps: assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artif Intell 279:103201. https://doi.org/10.1016/j.artint.2019.103201
7. Calvo RA, Peters D, Vold K, Ryan RM (2020) Supporting human autonomy in AI systems: a framework for ethical enquiry. In: Ethics of digital well-being (pp 31–54). Springer, Cham. https://doi.org/10.1007/978-3-030-50585-1_2
8. Castelfranchi C, Poggi I (1998) Bugie, finzioni, sotterfugi: Per una scienza dell'inganno. Milano, Carocci
9. Castelfranchi C, Tan Y (2001) Trust and deception in virtual societies. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-3614-5
10. Caudwell C, Lacey C (2019) What do home robots want? The ambivalent power of cuteness in robotic relationships. Convergence 41(8):1176–1191. https://doi.org/10.1177/1354856519837792
11. Chakraborti T, Kambhampati S (2018) Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-AI collaboration. arXiv:1801.09854
12. Chun WHK (2004) On software, or the persistence of visual knowledge. Grey Room 18:26–51.
13. Coeckelbergh M (2018) How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. Ethics Inf Technol 20(2):71–85. https://doi.org/10.1007/s10676-017-9441-5
14. Danaher J (2020) Robot betrayal: a guide to the ethics of robotic deception. Ethics Inf Technol 22(2):117–128. https://doi.org/10.1007/s10676-019-09520-3
15. Dazeley R, Vamplew P, Foale C, Young C, Aryal S, Cruz F (2021) Levels of explainable artificial intelligence for human-aligned conversational explanations. Artif Intell 299:103525. https://doi.org/10.1016/j.artint.2021.103525
16. de Graaf MMA, Allouch SB (2013) Exploring influencing variables for the acceptance of social robots. Robot Auton Syst 61(12):1476–1486. https://doi.org/10.1016/j.robot.2013.07.007
17. DePaulo BM, Kirkendol SE, Kashy DA, Wyer MM, Epstein JA (1996) Lying in everyday life. J Personal Soc Psychol 70(5):979–995. https://doi.org/10.1037/0022-3514.70.5.979
18. Donath J (2018) The robot dog fetches for whom? In: Papacharissi Z (ed) A networked self and human augmentics, artificial intelligence, sentience. Routledge, London, pp 10–24
19. Emslie K (2024) LLM hallucinations: a bug or a feature? Communications of the ACM. Retrieved 11 June 2024, from https://cacm.acm.org/news/llm-hallucinations-a-bug-or-a-feature/
20. European Commission, Directorate-General for Communications Networks, Content and Technology (2019) Ethics guidelines for trustworthy AI, Publications Office. https://doi.org/10.2759/346720
21. Floridi L, Cowls J (2021) A unified framework of five principles for AI in society. In: Floridi L (eds) Ethics, Governance, and Policies in Artificial Intelligence. Philosophical Studies Series, vol 144. Cham, Springer. https://doi.org/10.1007/978-3-030-81907-1_2
22. Fogg BJ (2003) Persuasive technology: using computers to change what we think and do. Morgan Kaufmann, Burlington
23. Gehl RW, Bakardjieva M (2016) Socialbots and their friends: digital media and the automation of sociality. Routledge, London
24. Golbin I, Axente M (2021) 9 ethical AI principles for organizations to follow. World economic forum. Retrieved 18 February 2022, from https://www.weforum.org/agenda/2021/06/ethical-principles-for-ai/#:~:text=The%20landscape%20of%20ethical%20AI,nine%20core%20ethical%20AI%20principles
25. Guzman AL (2015) Imagining the voice in the machine: the ontology of digital social agents. PhD Dissertation, University of Illinois at Chicago
26. Guzman AL, Lewis SC (2019) Artificial intelligence and communication: a human–machine communication research agenda. New Media Soc 22(1): 70–86. https://doi.org/10.1177/1461444819858691
27. Hakim FZM, Indrayani LM, Amalia RM (2019) A dialogic analysis of compliment strategies employed by replika chatbot. Adv Soc Sci Educ Hum Res 279:266–271. https://doi.org/10.2991/icalc-18.2019.38
28. Helberger N, Karppinen K, D'acunto L (2018) Exposure diversity as a design principle for recommender systems. Inf Commun Soc 21(2):191–207. https://doi.org/10.1080/1369118X.2016.1271900
29. Henrickson L (2021) Reading computer-generated texts. Cambridge University Press, Cambridge
30. Herzfeld N (2023) Is your computer lying? AI and deception. Sophia, 1–14.
31. Hoffman D (2019) The case against reality: why evolution hid the truth from our eyes. Norton & Company, New York
32. Kalluri P (2020) Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature 583(7815):169–169. https://doi.org/10.1038/d41586-020-02003-2
33. Kircher K, Larsson A, Hultgren JA (2013) Tactical driving behavior with different levels of automation. IEEE Trans Intell Transp Syst 15(1):158–167. https://doi.org/10.1109/TITS.2013.2277725
34. Lakoff G, Johnson M (1980) Metaphors we live by. University of Chicago Press, Chicago
35. Laurel B (1991) Computers as Theatre. Addison-Wesley, Boston
36. Lei X, Tu GH, Liu AX, Ali K, Li CY, Xie T (2017) The insecurity of home digital voice assistants-Amazon Alexa as a case study. arXiv preprint arXiv:1712.03327
37. Leveson N (2011) Engineering a safer world: systems thinking applied to safety. MIT Press, Cambridge
38. Lopatovska I, Williams H (2018) Personification of the Amazon Alexa: BFF or a mindless companion. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, pp. 265–268. https://doi.org/10.1145/3176349.3176868
39. Masters P, Smith W, Sonenberg L, Kirley M (2021) Characterising deception in AI: a survey. In: Sarkadi S, Wright B, Masters P, McBurney P (eds) Deceptive AI. DeceptECAI DeceptAI 2020 2021. Commun Comput Inf Sci, vol 1296. Cham, Springer. https://doi.org/10.1007/978-3-030-91779-1_1
40. Mclean G, Osei-frimpong K (2019) Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. Comput Hum Behav 99:28–37. https://doi.org/10.1016/j.chb.2019.05.009
41. Mecacci G, Santoni de Sio F (2020) Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. Ethics Inf Technol 22(2):103–115. https://doi.org/10.1007/s10676-019-09519-w

42. Marcus G, Davis E (2020) GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about. MIT Technology Review. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/
43. Milano S, Taddeo M, Floridi L (2020) Recommender systems and their ethical challenges. AI Soc 35(4):957–967. https://doi.org/10.1007/s00146-020-00950-y
44. Natale S (2021) Deceitful media: artificial intelligence and social life after the turing test. Oxford University Press, New York
45. Natale S (2023) AI, human-machine communication and deception. In: Guzman A, McEwen R, Jones S (eds) The Sage Handbook of Human-Machine Communication. Sage, London, pp 401–408
46. Natale S, Cooke H (2021) Browsing with Alexa: interrogating the impact of voice assistants as web interfaces. Media Cult Soc 43(6):1000–1016. https://doi.org/10.1177/0163443720983295
47. Nicas J, Kitroeff N, Gelles D, Glanz J (2019) Boeing built deadly assumptions into 737 max, blind to a late design change (Published 2019). Nytimes.com. Retrieved 22 February 2022, from https://www.nytimes.com/2019/06/01/business/boeing-737-max-crash.html
48. O'Leary DE (2019) Google's duplex: pretending to be human. Intell Syst Account Finance Manag 26(1):46–53. https://doi.org/10.1002/isaf.1443
49. Olney J (1972) Metaphors of self: the meaning of autobiography. Princeton University Press, Princeton
50. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 final
51. Santoni de Sio F, Van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Robot A I:15. https://doi.org/10.3389/frobt.2018.00015
52. Sarkadi S, Wright B, Masters P, McBurney P (2021) Deceptive AI. Springer, Cham
53. Sætra HS (2021) Social robot deception and the culture of trust. Paladyn J Behav Robot 12(1):276–286. https://doi.org/10.1515/pjbr-2021-0021
54. Schiller A, McMahon J (2019) Alexa, alert me when the revolution comes: gender, affect, and labor in the age of home-based artificial intelligence. New Polit Sci 41(2):173–191. https://doi.org/10.1080/07393148.2019.1595288
55. Schoenhofer SO, van Wynsberghe A, Boykin A (2019) Engaging robots as nursing partners in caring: nursing as caring meets care-centered value-sensitive design. Int J Human Car 23(2): 157–167. https://doi.org/10.20467/1091-5710.23.2.157
56. Shneiderman B (1986) Designing the user interface: strategies for effective human-computer interaction, 1st edn. Addison-Wesley, Boston
57. Shneiderman B (2020) Human-centered artificial intelligence: reliable, safe & trustworthy. Int J Hum Comput Interact 36(6):495–504. https://doi.org/10.1080/10447318.2020.1741118
58. Shneiderman B (2020) Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interact Intell Syst (TiiS) 10(4):1–31. https://doi.org/10.1145/3419764
59. Shneiderman B (2021) Human-Centered AI: Realiable, Safe & Trustworthy [Video]. Retrieved 22 February 2022. from https://www.youtube.com/watch?v=o5XwczIERvM
60. Shneiderman B (2022) Human-centered AI. Oxford University Press, Oxford
61. Shneiderman B, Plaisant C, Cohen M, Jacobs S, Elmqvist N (2016) Designing the user interface: strategies for effective human-computer interaction, 6th edn. Pearson, Boston
62. Schuetzler RM, Grimes GM, Giboney JS (2019) The effect of conversational agent skill on user behavior during deception. Comput Hum Behav 97:250–259. https://doi.org/10.1016/j.chb.2019.03.033
63. Seppelt B, Reimer B, Angell L, Seaman S (2017) Considering the human across levels of automation: Implications for reliance. In: Proceedings of the ninth international driving symposium on human factors in driver assessment, training and vehicle design (pp 228–234). https://doi.org/10.17077/drivingassessment.1640
64. Sison AJG, Daza MT, Gozalo-Brizuela R, Garrido-Merchán EC (2023) ChatGPT: More than a "weapon of mass deception" ethical challenges and responses from the human-centered artificial intelligence (HCAI) perspective. Int J Hum Comput Interact. https://doi.org/10.1080/10447318.2023.2225931
65. Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB (2021) My chatbot companion: a study of human-chatbot relationships. Int J Hum Comput Stud 149:102601. https://doi.org/10.1016/j.ijhcs.2021.102601
66. Smits M, van Goor H, Kallewaard JW, Verbeek PP, Ludden GD (2022) Evaluating value mediation in patients with chronic low-back pain using virtual reality: contributions for empirical research in value sensitive design. Health Technol. https://doi.org/10.1007/s12553-022-00671-w
67. Stahl BC, Wright D (2018) Ethics and privacy in AI and big data: implementing responsible research and innovation. IEEE Secur Priv 16(3):26–33. https://doi.org/10.1109/MSP.2018.2701164
68. Sugianto N, Tjondronegoro D, Stockdale R, Yuwono EI (2021) Privacy-preserving AI-enabled video surveillance for social distancing: responsible design and deployment for public spaces. Inf Technol People Vol Ahead-of-print No Ahead-of-print. https://doi.org/10.1108/ITP-07-2020-0534
69. Sutton DF (1994) Catharsis of Comedy. Lanham, Rowman and Littlefield
70. Sweeney M (2017) The Ms. Dewey "experience": technoculture, gender, and race. In: Daniels J, Gregory K, McMillan Cottom T (eds) Digital sociologies, pp 401–420. Bristol, Policy Press
71. Tarakji KG, Zaidi AM, Zweibel SL, Varma N, Sears SF, Allred J et al (2021) Performance of first pacemaker to use smart device app for remote monitoring. Heart Rhythm O2 2(5): 463–471. https://doi.org/10.1016/j.hroo.2021.07.008
72. Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, Cambridge
73. Thimbleby H (2020) Fix IT: stories from healthcare IT. Oxford University Press, Oxford
74. Tufekci Z (2018) Google Assistant making calls pretending to be human. Twitter. https://twitter.com/zeynep/status/994233568359575552 (Retrieved 26 April 2022)
75. Turing AM (1950) I—computing machinery and intelligence. Mind, LIX(236), 433–460. https://doi.org/10.1093/mind/lix.236.433
76. Umbrello S (2021) Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach. Ethics Inf Technol 23(3):455–464. https://doi.org/10.1007/s10676-021-09588-w
77. Umbrello S (2021) Towards a Value Sensitive Design Framework for Attaining Meaningful Human Control over Autonomous Weapons Systems (PhD). Northwestern Italian Philosophy Consortium (Consorzio FINO). https://doi.org/10.13140/RG.2.2.20431.41128
78. Umbrello S (2022) The role of engineers in harmonising human values for AI systems design. J Respons Technol 10:100031. https://doi.org/10.1016/j.jrt.2022.100031
79. Umbrello S, Capasso M, Balistreri M, Pirni A, Merenda F (2021) Value sensitive design to achieve the UN SDGs with AI: a case of elderly care robots. Mind Mach 31(3):395–419. https://doi.org/10.1007/s11023-021-09561-y

80. Umbrello S, Van de Poel I (2021) Mapping value sensitive design onto AI for social good principles. AI and Ethics 1(3):283–296. https://doi.org/10.1007/s43681-021-00038-3
81. Umbrello S, Yampolskiy RV (2022) Designing AI for explainability and verifiability: a value sensitive design approach to avoid artificial stupidity in autonomous vehicles. Int J Soc Robot 14(2), 313–322. https://doi.org/10.1007/s12369-021-00790-w
82. Umbres R (2017) Deception as exploitative social agency. In: Enfield NJ, Kockelman P (eds) Distributed agency (pp 243–251). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190457204.003.0025
83. United Nations. General Assembly (1949) Universal declaration of human rights (Vol. 3381). Department of State, United States of America
84. Van den Hoven J, Lokhorst GJ, Van de Poel I (2012) Engineering and the problem of moral overload. Sci Eng Ethics 18(1):143–155. https://doi.org/10.1007/s11948-011-9277-z
85. van de Poel I (2020) Embedding values in artificial intelligence (AI) systems. Mind Mach 30(3):385–409. https://doi.org/10.1007/s11023-020-09537-4
86. van Wynsberghe A (2021) Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics 1(3):213–218. https://doi.org/10.1007/s43681-021-00043-6
87. Weizenbaum J (1966) ELIZA: a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45. https://doi.org/10.1145/365153.365168
88. Wenskovitch J, Zhou M, Collins C, Chang R, Dowling M, Endert A, Xu K (2020) Putting the "i" in interaction: interactive interfaces personalized to individuals. IEEE Comput Graphics Appl 40(3):73–82. https://doi.org/10.1109/MCG.2020.2982465
89. Wilks Y (2019) Artificial intelligence: modern magic or dangerous future? Icon Books, London
90. Woods HS (2018) Asking more of Siri and Alexa: feminine persona in service of surveillance capitalism. Crit Stud Media Commun 35(4):334–349. https://doi.org/10.1080/15295036.2018.1488082
91. Wrathall MA (2010) Heidegger and unconcealment: truth, language, and history. Cambridge University Press, Cambridge
92. Yadron D, Tynan D (2016) Tesla driver dies in first fatal crash while using autopilot mode. The Guardian. Retrieved 22 February 2022. from https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk
93. Young L (2019) 'I'm a cloud of infinitesimal data computation' when machines talk back: an interview with Deborah Harrison, one of the personality designers of Microsoft's Cortana AI. Archit Des 89(1):112–117. https://doi.org/10.1002/ad.2398
94. Zhan X, Xu Y, Sarkadi S (2023) Deceptive AI ecosystems: the case of ChatGPT. In: Proceedings of the 5th international conference on conversational user interfaces, pp 1–6. https://doi.org/10.1145/3571884.3603754
95. Zicari, R (2020) *Definition of the boundaries – Z-Inspection*. Z-inspection.org. Retrieved 16 May 2022. from http://z-inspection.org/ecosystems-what-we-wish-to-investigate/
96. Zicari R (2020) Ethical maintenance–Z-Inspection. Z-inspection.org. Retrieved 16 May 2022. from http://z-inspection.org/ethical-maintenance/
97. Zicari RV, Brodersen J, Brusseau J, Düdder B, Eichhorn T, Ivanov T et al (2021) Z-Inspection®: a process to assess trustworthy AI. IEEE Trans Technol Soc 2(2):83–97. https://doi.org/10.1109/TTS.2021.3066209
98. Zicari RV, Ahmed S, Amann J, Braun SA, Brodersen J, Bruneault F et al (2021) Co-design of a trustworthy AI system in healthcare: deep learning based skin lesion classifier. Front Hum Dyn. https://doi.org/10.3389/fhumd.2021.688152

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Steven Umbrello is currently the Managing Director at the Institute for Ethics and Emerging Technologies and a research fellow at the University of Turin working on the theology of Bernard Lonergan applied to artificial intelligence. He is also an associate researcher at the Collège des Bernardins, where he works on digital humanism, and was previously a research fellow at the Delft University of Technology and the Center for Religious Studies at the Bruno Kessler Foundation. He is the editor of several international academic journals, including the International Journal of Technoethics, the Journal of Responsible Technology, and the Journal of Ethics and Emerging Technologies. He was formerly a Stiftung Südtiroler Sparkasse Global Fellow at Eurac Research, where he worked on the philosophy, religion, and society program. He is the author of several books, including his most recent, Technology Ethics: Responsible Innovation and Design Strategies (2024).

Simone Natale teaches media history and theory at the University of Turin. Before returning to Italy in 2020, he was a researcher and lecturer at international universities such as Columbia University in New York, Loughborough University in the United Kingdom, and Humboldt University in Berlin. He is the author of, among others, Supernatural Entertainments. Victorian Spiritualism and the Rise of Modern Media Culture (Penn State University Press 2016) and oDeceitful Media: Artificial Intelligence and Social Life after the Turing Test (2021) for Oxford University Press.