Philosophy & Technology (2023) 36:55
https://doi.org/10.1007/s13347-023-00656-1

RESEARCH ARTICLE

Interacting with Machines: Can an Artificially Intelligent Agent Be a Partner?

Philipp Schmidt[1] · Sophie Loidolt[2]

Received: 12 December 2022 / Accepted: 30 May 2023 / Published online: 25 August 2023

Abstract

In the past decade, the fields of machine learning and artificial intelligence (AI) have seen unprecedented developments that raise human-machine interactions (HMI) to the next level. *Smart machines*, i.e., machines endowed with artificially intelligent systems, have lost their character as mere instruments. This, at least, seems to be the case if one considers how humans experience their interactions with them. Smart machines are construed to serve complex functions involving increasing degrees of freedom, and they generate solutions not fully anticipated by humans. Consequently, their performances show a touch of action and even autonomy. HMI is therefore often described as a sort of "cooperation" rather than as a mere application of a tool. Some authors even go as far as subsuming cooperation with smart machines under the label of *partnership*, akin to cooperation between human agents sharing a common goal. In this paper, we explore how far the notion of shared agency and partnership can take us in our understanding of human interaction with smart machines. Discussing different topoi related to partnerships in general, we suggest that different kinds of "partnership" depending on the form of interaction between agents need to be kept apart. Building upon these discussions, we propose a tentative taxonomy of different kinds of HMI distinguishing coordination, collaboration, cooperation, and social partnership.

Keywords Artificial intelligence · Smart machines · Cooperation · Human-machine-interaction · Partnership · Social ontology · Joint action · Collective intentionality · Agency · Phenomenology

✉ Philipp Schmidt
philipp.schmidt@uni-wuerzburg.de

Sophie Loidolt
sophie.loidolt@tu-darmstadt.de

[1] Institut für Philosophie, Julius-Maximilians-Universität Würzburg, Ehrenhof Südflügel, Residenzplatz, 97070 Würzburg, Germany

[2] Institut für Philosophie, Technische Universität Darmstadt, Residenzschloss 1, 64283 Darmstadt, Germany

Abbreviations

AI	Artificial intelligence
HMI	Human-machine interaction
HRI	Human-robot interaction
RAM	Random access memory

1 Introduction

The developments of the past decades in the science of artificial intelligence and robotics are often seen as remarkable. Traditionally, machines have been used as tools and instruments to achieve a specific goal. In the case of AI, by contrast, increasing degrees of freedom allow systems to process information in ways that are not always predefined by humans. As a result, developers, users, and researchers have developed the habit to ascribe some form of “autonomy” to these systems. Moreover, given that machines show behaviors that are based on their own individual information-processing and inferences, it has also become quite common to call the relevant systems “agents” and their behaviors “actions.” When talking about the interaction with artificially-intelligent agents, machines, or robots, the recent literature in the relevant fields increasingly emphasizes the aspect of *acting together*.

This is reflected in the *labels* under which the relationships between humans and machines and the ontological status ascribed to machines are being discussed. Human-machine interaction (HMI) or human-robot interaction (HRI) have been described in terms of “cooperation” (Hoc, 2013; Yang et al., 2022), “collaboration” (Gervasi et al., 2020) and even as a “team” (Dehkordi et al., 2021; Fiore & Wiltshire, 2016; Groom & Nass, 2007; Musić & Hirche, 2017; Seeber et al., 2020; Tabrez et al., 2020) or “partnership” (Bhavik et al., 2019; Newman & Blanchard, 2019; Dumouchel & Damiano, 2017; Mathur & Reichling, 2016; Stenzel et al., 2012). Instead of being considered mere tools, machines are assigned roles that imply a quasi-social status, such as that of a “collaborator,” a “teammate,” or even a “partner.” Often, these concepts are applied without further discussion and used interchangeably. Sometimes, authors provide explanations for their terminological choices. For instance, Strasser defends the notion of partnership as follows: “I assume that if artificial agents contribute to social interactions by utilizing socio-cognitive abilities and thereby add to a reciprocal exchange of social information, we are justified to consider them social interaction partners” (Strasser, 2022, 523).

Such an information-based proposal, it seems, can easily be backed up by experience if one considers interaction with robots—especially when constructed in a humanoid way. Machines with human-like bodies, showing similar movement styles and being capable of emotion-like expressions, may well give rise to the impression that one is interacting with some form of *alter ego*. Moreover, if a machine shows responses to a situation in a way that appears to make sense, it is easy to imagine that interacting with such a machine can give rise to the *feeling* that one is dealing with a partner.

Yet, it is an open question whether the notion of *partnership* and related concepts, such as joint action and trust, are really apt to capture the interaction between humans and machines. While these concepts are widely used to distinguish more complex artificial systems from the use of machines, computers, and robots as mere tools, a discussion of how best to describe and label these new forms of interaction is still pending (Castañer & Oliveira, 2020; Heinrichs & Knell, 2021; Janssen et al., 2019; Pacaux-Lemoine & Flemisch, 2019).

In our paper, we shed light on the notion of partnership in HMI from a philosophical perspective and employ a conceptual analysis to contribute to this important debate. Referring to someone as a partner in a human context entails certain expectations: we expect our partners to, at least, share some goals with us; that they are trustworthy and appreciate our agreements; and that they are capable of fulfilling their agreed-upon role in the partnership. While these issues are trivially at stake when we partner up with humans, it becomes less trivial in HMI. There is a risk that, by calling a machine a partner, we tacitly attribute features that machines and our relationships with them lack, as they may only make sense in the human-human case. Conceptual clarification is thus needed to understand what kind of partnership is possible with machines.

We propose distinguishing between different types of partnerships and their relation to various forms of interactions. In Section 2, we present a preliminary conceptual framework that differentiates between forms of interaction between humans and machines. In Section 3, we examine partnerships in HMI through four philosophical topoi: machine autonomy, joint action, trust, and sense-making. Analyzing these topoi within the context of HMI will help us characterize the various types of partnerships that can emerge. In Section 4, we aim to provide a nuanced response to whether an artificially intelligent agent can be a partner to humans and, if so, in what capacity or manner. This response will also include a statement on the necessary conditions required for social partnership in the strict sense.

2 Conceptual Framework: Toward a Taxonomy of Interactions in HMI

Before delving into the partnership issue, we propose making a few conceptual distinctions among different forms of interactions. Our working thesis is that by discerning these various interaction modes, we can provide a nuanced answer to whether partnerships between humans and machines are possible and how we should characterize them. To accomplish this, we concentrate on two key questions:

1. Does the machine share a goal with the human?
2. Are the goals and tasks fully described and anticipated by humans?

Building upon these questions and the commonly used labels for relationships between humans and machines in HMI, we propose a distinction between the following types of interaction:

1. *Coordination*: In this form of interaction, humans and machines work towards their respective objectives independently. They do not share a common goal, but

the activity of the other is taken into account in one's own actions. The machine's objective can either be fully or partially predefined or predicted by humans.

2. *Collaboration*: In this form of interaction, both humans and machines share a common (overall) goal that is defined by the human. Humans fully anticipate and program the range of possibilities in which the machine operates.
3. *Cooperation*: In this form of interaction, both humans and machines share a common (overall) goal that is defined by the human. The range of possibilities in which the machine operates is not fully anticipated and programmed by humans.
4. *Social partnership*: This form of interaction can involve coordination, collaboration, or cooperation. It is primarily defined by the ontological features of the subjects, which are generally on par with each other. It is not solely determined by the aforementioned questions. Social partnership refers to the kind of partnership discussed in ordinary language when referring to joint action between human individuals, which involves *intentional* action.

Two immediate questions arise when introducing this fourfold distinction. The first question pertains to the concept of agency, while the second revolves around the concept of partnership.

2.1 Interaction, Goal-Directed Behavior, and a Broad Notion of Agency

Discussing different forms of interaction, from the beginning, entails a certain language game: When two or more entities engage in interaction, this seems to imply that each of them displays a certain "activity," and performs some form of "action." Consequently, these interacting entities are considered to possess a form of "agency" and are referred to as "agents," whether humans or machines. This is common in the literature. However, conceptual challenges arise when considering machines as agents. It may be tempting to restrict the notion of action to *intentional* and *conscious* action for valid reasons. While we acknowledge these challenges and strive to uphold conceptual "maximum conditions" in this paper, the task at hand necessitates working with a broader conception of action and agency. The central question concerns whether machines can be partners. Restricting action and activity solely to intentional actions would limit the applicable concept for interactions with machines to that of tool use. If machines cannot be considered agents at all, by definition, they can only be regarded as tools. Tools, however, can already conceptually not be conceived as partners. For these reasons, we hold that it is crucial to maintain a broad understanding of action and agency. Such an understanding is pivotal in developing a nuanced perspective on partnership as well as the utilization of tools. Accordingly, the lower boundary of this comprehensive framework is defined where action and agency can be distinguished from the mere use of simple objects, which require manual handling and constant human control to bring about a change in the world. Therefore, we do not use the terms "actant" or "actor" as employed in *actor-network theory* (Latour, 2005, 54, 72), referring to any tool or object that "participates in the action" (Latour, 2005, 72). A hammer is *not* an agent; it has no control over the situation. *We* control the hammer as we use it.

Agency, as we define it, emerges when a machine, to some extent, assumes control over aspects of a given situation. For example, a machine that moves within space and follows a pre-defined script actively engages in the minimal sense that it exhibits behavior and exercises control over its movement. It is not passive like a hammer, which remains stationary once placed and has no principle of movement. More complex machines can even exhibit behaviors based on algorithms that are unanticipated by humans. In such cases, the increased power of the machines corresponds to a decrease in human control. Recognizing that agency exists on a spectrum with varying degrees of control, we classify entities as "agents" when they serve as a locus of control in a relevant sense. One might spell this out further and define action minimally in terms of the ability to respond to a stimulus by changing state, the ability to change a state without stimulus, or the ability to adapt related rules by which states are changed (Floridi & Sanders, 2004). Others emphasize the importance of normativity even for minimal action, defined as the ability to generate norms for a system's interaction with the environment (Barandiaran et al., 2009). We concur that normativity, in terms of minimal functional achievements, is pertinent to action. However, for the purposes of this paper, we adopt a more liberal notion, whereby normativity need not be self-defined by an agent. For example, a wind-up toy is not an agent, whereas an electric toy car programmed to exhibit certain steering patterns in specific circumstances can be considered an agent, albeit in a limited sense. While the former does not process any information (and its movement is comparable to that of a stone being thrown through the air), the latter operates based on rules to comply with certain norms that it may fail to meet. It is in this sense that the entity is a locus of control in the most minimal sense, enabling us to define interaction as the distribution of control among participating agents in a given situation.

2.2 Trivial, Weak, and Strong Kinds of Partnership

Thanks to a broad notion of action, we can differentiate between *different kinds of partnership* in HMI. Having distinguished between coordination, collaboration, cooperation, and social partnership already, the question arises as to whether partnership only adheres to the last form of interaction. In the case of coordination, no shared goals between human and machine agents obtain and so it seems wrong to attribute partnership. Moreover, it seems linguistically odd to refer to "coordination partners." However, we can still consider them as "interaction partners" in a *trivial* sense of partnership: Agents engaging in coordination are partners in forming part of an interaction that is shaped by all agents involved and that works if they adapt to the other agents' activities in a way that allows everybody to pursue their goals. By contrast, in the case of collaboration and cooperation, talk of partner seems less linguistically odd. In our framework, we accept collaboration and cooperation as *weak* forms of partnership, but we want to distinguish them from *strong* forms of partnership, which are of a social kind in the strong sense. For this, we will develop maximum conditions, in order to get clearer about what this "strong sense" really entails. When we refer to sociality, our primary focus is on human-human encounters.

Although we deem social encounters with non-human beings generally possible, we take the human-human encounter to be the paradigmatic case.

At this point, however, these decisions are nothing but conceptual preliminaries. Moving forward, we will now explore four philosophical topoi to characterize further the different forms of interactions and the potential types of partnership they may entail. This exploration will help us address the question of whether an artificial agent can be a partner to humans, and if so, in what capacity.

3 Partnerships in HMI? Philosophical Perspectives

What do authors mean when they use the label "partnership" or related concepts in order to describe the relationship between humans and machines? The lack of a comprehensive discussion on the application and appropriateness of these labels makes it challenging to fully grasp the philosophical implications associated with considering machines as partners or teammates. However, we believe that exploring *four philosophical topoi* can shed light on the nature of human-machine interaction and the possibilities for partnerships between humans and machines: (1) autonomy and instrumentality, (2) joint action, (3) trust, and (4) participatory sense-making.[1]

3.1 Topos 1: Mere Tool or…? Autonomy and Instrumentality

Where an intelligent machine possesses autonomy, exhibits goal-directed behavior, and proves useful in interacting with humans, it is likely to create the impression of engaging with a partner. Consequently, one may be inclined to assume that autonomy implies partnership, particularly when compared to mere tool-use. However, as we aim to demonstrate in this section, this hasty conclusion is not entirely accurate and requires further elaboration. While we recognize that exploring autonomy can contribute to our comprehension of the potentialities and prerequisites of human-machine partnership, it is important to emphasize that the insights derived from such a discussion are predominantly negative: merely attributing autonomy to an entity does *not* automatically imply the establishment of a partnership with said autonomous entity. Certainly, there are quite different notions of autonomy. But we argue that even if a machine were to fulfill the most demanding criteria of autonomy, usually only applied to human persons, human-machine partnership in the strong sense

[1] We conducted a series of interviews to gather insights and perspectives on partnerships with machines. The interviews involved seven individuals, including five engineers, one mathematician, and one cognitive scientist from the Technical University of Darmstadt (Germany). Additionally, we conducted four group interviews with engineers. The interviews were designed to be unstructured and exploratory, encouraging discussions on the conceptualization of AI-loaded machines as partners. Our aim was to identify key issues and considerations raised by experts when discussing partnerships with machines. Throughout the interviews, the interviewees consistently addressed and highlighted the relevance of the four philosophical topoi we have been examining. Their inputs provided valuable insights and enriched our own discussions on the topic. We extend our sincere gratitude to all the interviewees for their participation and for contributing to these fruitful conversations.

would not emerge. Despite this negative outcome, the discussion of autonomy also yields positive insights. We will propose that delving into various facets of machine autonomy can assist in identifying certain necessary conditions for the formation of a partnership in the strong sense.

3.1.1 Varieties of Autonomy: No Implication of Partnership

Before we can ask whether autonomy implies partnership, we need to understand what autonomy amounts to. Let us give a few examples from literature, film, and other fiction to show that none necessarily involve partnership. In social imaginaries and popular culture, machine autonomy has often been portrayed in the gloomy context of master-slave dialectics that menaces human autonomy. One could think of more recent series such as *Westworld*; Dihal mentions the famous *Terminator*, *Robocalypse*, and *Matrix* as illustrative examples (Dihal 2020, 189). But even much earlier, the play *R.U.R. (Rossum's Universal Robots)* by Karel Čapek (1921), which introduced the concept of "robot," explored the idea of humanoid machines created to serve humans but eventually rebelling and seizing complete control. The concept of robotics is thus closely linked with a particular understanding of the Hegelian dialectics between *master* and *slave*. In one scenario, the robot, functioning as a slave, can be controlled and used as a mere tool, entirely at the disposal of humans to serve specific functions. In another scenario, the robot awakens to its power and decides to defy human commands. It becomes the master which completely alters the dynamic between humans and machines. In these cases, it appears that autonomy, and even a consciousness of one's existence and freedom to act, often leads to adversarial relationships rather than partnerships.

Scenarios involving dominion through machines, master-slave dialectics, and revolutions brought about by machines are exclusive to fiction. However, the broader issue of the dialectical relationship between *being in control* and *being subject to machine control* is a real concern in existing technology. Machine learning (ML) enables AI to analyze vast amounts of data and achieve high prediction accuracy, providing valuable insights for humans. In the case of *Black Box AI*, the processes leading to AI's predictions are not interpretable or understandable by humans. Instead, humans can only recognize that the Black Box prediction models generated by AI through machine learning are often more accurate than models that can be comprehended by humans. This phenomenon is referred to as epistemic and practical opacity (Kaminski, 2019; Kaminski et al., 2018; Vaassen, 2022). Furthermore, when AI systems with Black Box prediction models are also equipped with decision-making capabilities and the ability to execute specific operations, they are often regarded as possessing some form of "agency."

At this point, *autonomy* becomes a central issue, although there is no consensus on the precise meaning of machine autonomy. As a result, definitions can vary significantly. Some define autonomous systems simply as robots "which are able to perform well-constrained tasks such as surgery, driving on a highway, or vacuum cleaning" (Harbers et al., 2017, 20). However, the fact that these machines operate with a specific goal and within a predefined scope has led some to argue that autonomy requires more than mere goal achievement. According to a narrower understanding, autonomy is attributed to systems when they learn to "handle the sensor signals, adapt their behaviors and act intelligibly" in the case that they are "introduced to an unknown scope or domain, given suitable sensors and actuators for that scope, without changing the algorithm in any way" (Ezenkwu & Starkey 2019, 336).

Contrasting such a proposal, some have emphasized that true or complete autonomy does not exist, arguing that "'autonomous system' is a misnomer," as "[n]o entity … is capable enough to be able to perform competently in every task and situation" (Bradshaw et al., 2013, 57). Moreover, they also question that autonomy can be described as a "feature" to begin with: "autonomy isn't a discrete property of a work system, nor is it a particular kind of technology; it's an idealized characterization of observed or anticipated interactions between the machine, the work to be accomplished, and the situation" (Bradshaw et al., 2013, 56).

A less negative but still strict proposal restricts autonomy to living systems insofar as autonomy is taken to be a feature of the autopoiesis of life and metabolic self-production (Varela, 1979). While such an understanding of autonomy presently only applies to natural life, it might potentially also be attributed to artificial agents. Froese et al. (2007) suggest distinguishing between *behavioral autonomy* and *constitutive autonomy*: the former covers those approaches that define autonomy in terms of automatic goal-directed behavior, independently of how detailed goals and means are pre-defined; the latter covers the narrower definitions of autonomy that refer to the self-constitution of a system.

Finally, there are also attempts at defining autonomy in relative terms. For instance, Müller suggests the following definition: "Agent X is autonomous from agent Y to the degree that X pursues its goals without input from Y" (Müller, 2012, 213).

How do these concepts of behavioral, constitutive, and relative autonomy relate to the debate surrounding the partnership between humans and machines? It is important to note that while a machine may exhibit behavioral autonomy, it does not automatically imply coordination with human activity. Behavioral autonomy is a prerequisite for interaction, but it does not guarantee partnership. Similarly, constitutive and relative autonomy, which are characteristics of animals and living organisms, do not inherently establish a partnership with humans. Merely being constitutively autonomous, like ants, snails, or badgers, does not make them partners to humans. Therefore, the presence of autonomy alone does not imply a partnership. With this in mind, we will now delve into a more philosophically sophisticated concept of autonomy.

3.1.2 Kantian Autonomy and Some Insights into the Conditions of Social Partnership

Both behavioral and relative autonomy, which artificial agents can exhibit, are essential for coordination, collaboration, cooperation, and social partnership. However, it is likely that additional conditions must be met, particularly in the case of social partnership. In this section, we aim to explore whether a more philosophically sophisticated notion of autonomy can help us identify these conditions more precisely. One notable example of such a philosophical concept of autonomy is the Kantian perspective, which is closely linked to the idea of personhood and, in its most rigorous forms, encompasses the interplay of autonomy, freedom, and the moral law. According to Kant's philosophy, persons are considered autonomous because they can choose their principles of action. As a person, in Kantian terms, one can act upon the moral law and thus free oneself from individual inclinations. Because persons are autonomous in this sense, Kant argues in the famous third formulation of the categorical imperative, one should always treat them *also* as an "end in itself" and never *only* as a means to an end. Treating another human being as an "end in itself," acknowledges their autonomy in this strong sense. For Kant, this is tantamount to respecting the moral law and acting upon the principles of freedom. This not only entails respecting the autonomy of others, but it also confirms oneself as an autonomous agent in the complete sense. One's own autonomy hence includes acknowledging the other's autonomy—and vice versa (Kant, 2012, 2015).

Evidently, machines, as we know them, are not autonomous in the Kantian sense (and it is debatable whether they are so in the near future, if possible at all). The interesting question, however, is whether machines would be *required* to be autonomous in this sense to be considered as possible candidates for social partnership.

Certainly, claiming that Kantian autonomy is a condition for social partnership would set the bar fairly high, and it does not look like such a proposal would be convincing. We can partner up with children, even with animals in a working process (e.g., in agriculture), and in neither case would we speak of autonomous persons in the Kantian sense. Moreover, Kantian autonomy alone does not at all imply partnership per se: In a sports or business competition, we can morally respect the autonomy of the adversarial players or colleagues. But it does not mean we partner up with them. Our goals might even straightforwardly contradict each other. Finally, it is also not necessary to respect the partner all the time in their moral autonomy, in order to speak of a partnership. People sometimes betray their partners.

However, it is important to note that they can only betray them, *because* they are partners. This is so because betrayal consists in the disappointment of specific expectations that partners have towards each other. Which are these expectations? The Kantian distinction between "mere means" and "end in itself" can help to spell this out: We suggest that what we expect from our partners is that they do not consider us in all their actions exclusively as means for their goals but that they respect us *also* as ends in ourselves, at least in *some* of their actions. That is, social partnership does not require perfect moral action or attitude towards the social partner. Yet, failing to recognize the partner's status of being an end in itself altogether, i.e., in any circumstance, would undermine social partnership significantly. Kant's concept of autonomy, thus, allows us to explicate the view that *being a tool* and *being a social partner* is in conceptual opposition and to formulate the following claims:

1. Pure instrumentality and social partnership are incompatible.

This assumption has a further implication: Social partners must qualify as *possible* ends in themselves. So, only where treating another agent as an end in itself in some regard, a social partnership can emerge. An agent who cannot possibly be treated as an end in itself at all cannot be considered as a candidate for social partnership. This can be easily illustrated with a simple tool such as the notorious "hammer." It is a material object that can serve us in different ways. We may therefore praise it, but it would seem hard to argue that we can treat it as if it were an end in itself. It also fails to be an agent (as we made clear in our conceptual pre-definition at the beginning). Our practice is fully determined by using it as an instrument. The constraint we want to formulate for social partnership is thus:

2. To possibly count as a social partner, another agent must be treatable as an 'end in itself'.

This raises another question: Under which circumstances is it actually possible to treat another agent as an end in itself? Again, the Kantian proposal is too narrow. We might therefore turn to Aristotelian accounts, as e.g., Korsgaard (2018) does in the case of animals, and could emphasize that their form of life has an *entelechy* structure: their life has a telos in itself. Although they are not Kantian persons, it can make sense to treat their life as having an end in itself. So we suggest a rather formal claim: Whenever we want to talk about social partnership, we have to justify it by showing in which sense candidates for social partnership can possibly be considered ends in themselves. Hence, while we acknowledge the Kantian proposal to be too demanding, we suggest that it should be taken as the vantage point of negotiation over what can possibly count as a social partnership. The distinction between "a mere means to an end" and "an end in itself" decodes the difference between a mere tool and a possible social partner.

Let us illustrate our points with an example, which will also serve us to make a third point. Consider a high-level autonomous machine as Ezenkwu and Starkey (2019, 338-340) describe it, an AI system with the following features:

- *Perception*: the ability to gain information about the environment
- *Actuation*: the ability to provide feedback to the world
- *Learning*: the ability to make sense of sensory inputs
- *Context-awareness*: the "ability … to sense, interpret and adapt to the current context of its environment" (Ezenkwu & Starkey 2019, 338)
- *Decision-making*: the ability to make decisions
- *Domain-independence*: independence of pre-defined ontological knowledge of the environment
- *Self-motivation*: the ability "to act intelligibly in an environment without being hardcoded or being given a task-specific value system" (Ezenkwu & Starkey 2019, 339)
- *Self-recovery:* the ability "to foresee possible causes of failure and devise a solution to abate it" and "to recover from failure after it has occurred" (Ezenkwu & Starkey 2019, 339)
- *Self-identification of goal*: the ability to set goals within an environment that is unknown a priori

An AI system equipped with such capacities is highly sophisticated and has the potential to induce experiences in humans that go beyond perceiving the artificial agent as a simple tool or an inanimate object, similar to a hammer waiting to be picked up. Such a device will be capable of complex interactions with humans. It is possible that individuals interacting with such an artificial agent will develop a sense of engagement as if they were interacting with a social partner. But is it sufficient that the human agent *experiences* their interaction with a machine as a partnership to *actually constitute* partnership? Some seem to think so (Coeckelbergh, 2009, 2011). Bearing in mind our interpretation of Kant, however, we would hesitate to call this a social partnership, given the artificial agent's dubitable status as an "end in itself." This leads us to the third conclusion that develops a maximum condition of social partnership:

3. Ontology matters: Treating another agent as an end in itself is not reducible to the agent's attitude.

Our argument revolves around the recognition of an agent as a social partner, which we believe should be grounded in ontological principles: whether an agent can be possibly conceived of and recognized as end in itself depends on the agent's form of being. The profound significance of the term "social" lies in its dependence on more than just the attitudes, illusions, or well-grounded assumptions of one agent regarding their interaction with another (as opposed to merely utilizing something as a tool). So, we would not agree that "what matters for the human-robot relation is how the robot appears to human consciousness" and that "ontological differences—what the entity really is—become irrelevant to the development of the human-robot relation" (Coeckelbergh, 2011, 199). Sociality needs two, in an ontological sense. Therefore, the status of being an end in itself, rather than a mere tool, cannot be solely determined by the tool-user. Until a comprehensive account is provided that explains how artificial agents can possess the ontological status of being ends in themselves, it would be prudent to refrain from attributing social partnership to human-machine interaction.

3.2 Topos 2: Joint Action, Collective Intentionality, and Phenomenology

In this section, we look at HMI through the lens of recent debates on phenomena that have been discussed under labels such as "joint action," "shared agency," "we-agency," or—more broadly—"collective intentionality" in social philosophy. Drawing from these debates, we focus on whether human interaction with an AI system in light of a specific goal can be construed as a case of "joint action" and, if so, in which sense. We argue that examining which understanding of joint action might apply to human-machine interaction will shed light on human-machine partnership and help further characterize the different kinds of HMI. The underlying rationale is that joint action is an essential element of partnership, one that distinguishes *acting together with* another agent from *merely interacting with* another agent and *acting with* a tool.

We present and discuss five different proposals of joint action and relate them to the different types of HMI.

3.2.1 The Modeling of Partnership: Joint Action, Shared Intention, and Information Theory

In the AI and robotics literature, the view that HRI can involve different forms of joint action is fairly widespread (e.g., Belhassein et al., 2022; Ciardo et al., 2022; Clodic et al., 2017; Grynszpan et al., 2019; Iqbal & Riek, 2017). Often, research on joint action in HRI is based on the view that two (or more) agents act together when they can be roughly said to (a) share a goal, (b) intend to engage in actions towards that goal, and (c) have a common knowledge of the other agent's (sub)goals, intentions, and actions (Bratman, 1992, 1993, 1997; Tomasello & Carpenter, 2017; Vesper et al., 2010; Sebanz et al., 2006; counter-arguments can be found in Clodic et al., 2017; Iqbal & Riek, 2017; Pacherie, 2011). Despite the existence of these different nuances regarding what is structurally required for shared intentions in HRI, many researchers in the AI and robotics literature agree—if only silently—on the following three essential tenets:

(i). that shared intentions can be modeled in terms of information processing;
(ii). that humans and machines can share intentions;
(iii). that shared intention in HMI constitutes, in a relevant way, joint action.

Although different proposals might exist regarding how exactly joint action emerges and what is necessarily required, accepting that joint action in HRI can take more complex forms is common. For instance, Clodic et al. (2017) describe processes of shared intention and joint action in HRI which involve that the "robot builds and maintains a distinct model of itself and of its human partner concerning the state of the world" and that it "also reasons and builds its own behavior based on its estimation of its human partner's intentions, knowledge and skills" (Clodic et al., 2017, 172). The idea is that an AI system represents—in computational terms—both itself and the human agent and their intentions, actions, and knowledge regarding a shared goal, and that it can reason based on all these different kinds of information or, in short, that the joint action between a human and an artificial agent can be understood in terms of a *shared mental model* (Dehkordi et al., 2021; Tabrez et al., 2020). The idea that joint action can be modeled and described as an exchange of information also lies at the bottom of the numerous studies investigating the different aspects and processes of communication between human agents (Castelfranchi, 1998). Many suggest applying this idea to HRI (Albrecht & Stone, 2018; Belhassein et al., 2022; de Vicariis et al., 2022).

It is worth noting that the majority of attempts to model joint action in HRI rely on a specific interpretation of shared intention and joint action, which has been extensively debated in the literature on human-human interaction. We now take a brief look at some alternative takes to illustrate that most accounts of joint action emphasize components that actually render joint action non-reducible to the exchange of information.

3.2.2 Alternative Takes on Joint Action and Shared Intention

According to Szanto (2016), the debate on joint action can be described as a "quadripartite conceptual landscape" (156). In the last section, we have already mentioned the first view:

1. A specific *content* of the agents' intention defines joint action (Bratman, 1992, 1993, 1997).

The content-view is prevalent in most endeavors to model joint action in HRI. However, in broader discussions on joint action, three other perspectives have been extensively debated, each characterized by specific theses:

2. A specific *mode* of the agents' intention defines joint action that is directed at a specific content (for instance, a particular goal and the necessary steps to achieve it) (Searle, 1990, 1995, 2010; Tuomela, 2007).
3. *Joint commitment* and the constitution of a *plural subject* define joint action (Gilbert, 1989, 2003, 2006, 2009).
4. Affective and conative *social interrelations* constitute shared intention and joint action (Meijers, 2003; Schmid, 2005, 2009).

Why does research in HRI typically favor the content-view? Is it the most convincing approach to joint action? Does it provide the best explanation of what joint action is and how it comes about? Typically, we do not find a discussion of the different approaches to joint action in the HRI literature, in which the content-view is the nearly unchallenged standard view.[2] We believe that the reason for this lies in the fact that the alternative views (2.-4.) are more sophisticated in that they demand specific capacities on the side of the agents. More specifically, we suggest that they require that agents are phenomenally conscious, i.e., that there is something "*it is like*" (Nagel, 1974) for agents to interact with each other.

Take, for instance, the mode-view and the proposition that a distinct mode of intention defines joint action. How can we differentiate between an individual intention and a we-mode of intention if not by appealing to a distinct qualitative character in experience? Some may argue that differences in intentional mode can be reduced to variations in algorithms, functions, or data. However, to our knowledge, no attempts have been made to describe different intentional modes in informational terms. In fact, it seems that any such endeavor would quickly collapse into the content-view.

It is also challenging to envision how the commitment-view could be viable without the involvement of subjective experience. While joint commitment may be partially formalizable and applicable to artificial agents, concerns arise regarding the nature of "commitment" exhibited by non-conscious agents. It is likely that any "commitment" displayed by such agents would merely mimic the outward appearance of commitment observed in human agents. The specific inner affirmation and promise inherent in human

[2] Some exceptions to this are attempts at applying the notion of joint commitment to HRI (Belhassein et al., 2022; Michael & Salice, 2017; Salice & Michael, 2017).

commitments would be absent if the allegedly committing agent lacks subjective experience, including feelings and desires. Similarly, no affective and conative social interrelations, as the relational view emphasizes as crucial to joint action, will accrue.

What conclusions can be drawn from this? The first thing to note is that most accounts of joint action are more demanding than the content-view. While these alternative takes may not lend themselves for modeling and application to HRI, they seem to capture more of what is essential to the phenomenon of human joint interaction. The content-view, in turn, can be modeled and is applicable to HRI but seems to miss essential aspects of human joint interaction. The implication of these insights is that only an impoverished understanding of joint action, which abstracts from what alternative views take to be of crucial matter for acting together, makes sense in the context of HRI. If we want to understand the kind of joint action involved in social partnership in the strict sense, we will need to take into account the role of subjective experience in acting together. Szanto's (2016) own take on collective intentionality, which he offers as a fifth alternative, provides a phenomenological approach to joint action that puts subjective experience at the center. To further illustrate the relevance of phenomenal consciousness, the next section discusses his approach.

3.2.3 Phenomenal Consciousness and Its Importance for Social Partnership

Szanto's (2016) Husserlian proposal suggests that the sharedness of a shared intention cannot be attributed to any single factor such as content, mode, or subjectivity alone. Rather, Szanto argues that all these factors play a role in constituting collectivity, leading to a complex and differentiated approach. According to him, it is essential to distinguish between different forms of communalizations that arise based on how subjects interact and engage with each other. Two assumptions are central to Szanto's multi-layered framework of sharedness:

1. As he emphasizes, "subjects of CI [sc. collective intentionality], phenomenologically viewed, are *ab ovo* communalized, […] they are subjects who always already stand in social relations to another" (Szanto, 2016, 156). Thus, the point is that sociality does not emerge out of otherwise socially blind behavior, as it were; instead, any shared intention can only arise in a context in which agents are already socially structured in a minimal sense. What is this minimal sense? To answer this question, a look at the second assumption is necessary.
2. The different levels of communalization take their structural origin in "'empathic consciousness,' wherein individuals can be said to directly *experience* the consciousness of others" (Szanto, 2016, 166).

The idea behind the second assumption, that in social cognition other agents' consciousness is directly experienced, has been defended as the *direct social perception* thesis (Krueger, 2018; Schmidt, 2018; Zahavi, 2011). It refers to the notion that perceiving other subjects as being conscious is not the product of a simulation or inference but an own mode of intentional consciousness called "empathy" (or, alter-ego-perception, *Fremdwahrnehmung*). This rests on the claim that conscious states are not simply "internal" states but are directly accessible through the experiencing subject's body and expressiveness. Even though we might not see the *feeling* of another person's anger, the blushed face and tense posture are directly visible. Crucially, these bodily phenomena are taken as constitutive parts of the subject's experiences rather than mere signs or expressions of anger. In seeing the bodily manifestation of anger, we see the anger itself. According to Szanto's phenomenological proposal of shared intention, the minimal sense in which an agent must already be socially structured, then, can be described in this way: the agent must be capable of directly experiencing the consciousness of other agents, meaning they should not only detect signs in the other agent's activity to infer conscious activity, but also *perceive* the other's conscious activity through their bodily manifestations.

That entails two essential claims. First, any agent involved in a shared intention *must possess a form of consciousness* that enables other subjects to empathize with it. Secondly, any agent involved in a shared intention must have the capacity for empathic consciousness. To put it differently, for any shared intention to develop, all agents must be capable of perceiving each other as conscious subjects that undergo experiences. That means an artificial agent that behaves *as if* it enjoyed the kind of phenomenal consciousness we do, and which acts *as if* it could empathize with our experience would not be enough. For instance, a robot that can "infer" that a human person is undergoing the experience of feeling warm and thirsty by measuring the air temperature and tracking specific facial movements would not count as a case of empathic consciousness. While human persons' empathic experience sometimes might also involve such contextual knowledge, the crucial point in empathy is that the *empathizing subject* is itself conscious and has a certain grasp of what it feels like for another subject to be in the experiential state it is. Empathic consciousness is directed at consciousness, implying that the *subject empathized with* is conscious too. Empathy is a relation between conscious subjects. For Szanto (2016), this is the foundational, minimal social ground on which more complex forms of communalization can be built.

On the next level, what becomes possible based on empathic consciousness is a shared sense of "experiencing the *same* world" (Szanto, 2016, 166). The interlocking of different instances of empathic consciousness of subjects generates a form of we-sense in that it is "all of us," i.e., all the subjects experiencing *this* world together. This sense allows different experiencing subjects to communicate with each other and engage in any social coordination in the first place.

Socio-communicative acts, as a third layer, presuppose that participating agents perceive themselves mutually as forming part of a shared lifeworld. Notifications, signals, or reaching out to the other qua social acts make only sense when the so-acting subject takes it that the other moves within a the same worldly context.

In addition to empathic consciousness, the consciousness of a shared lifeworld, and socio-communicative acts, Szanto (2016) identifies a fourth level of sharedness in Husserl, for which he reserves the notion of collective intentionality in the strict sense, the kind of shared intention that obtains when subjects constitute a group to achieve a shared goal. What is necessary for this to happen is an "appropriate intentional integration of the intentional, goal-directed, normative, volitional and practical properties of the mental life of *always and already socialized and communalized* individuals" (Szanto, 2016, 159).

To determine how such an integration looks like requires further analysis, but what suffices for our purposes here is simply the fact that all the mentioned levels presuppose all participating agents to be endowed with consciousness and empathic capabilities. It is only out of the experience of a shared lifeworld that more sophisticated forms of shared intention can emerge. Szanto's phenomenological proposal is complex and demanding; it makes it seem that sharing an intention with a non-conscious machine—regardless of how developed and autonomous it may be—is impossible, casting doubts on whether HMI directed at a specific goal could possibly qualify as joint action. But is such a strong conclusion necessary? We want to suggest another differentiated conclusion by considering all approaches to joint action and relating them to collaboration, cooperation, and social partnership.

3.2.4 Joint Action in Collaboration, Cooperation, and Social Partnership

We presented five different approaches to joint action that put emphasis on (1) intentional content, (2) intentional mode, (3) joint commitment, (4) affective and conative social interrelations, and (5) phenomenal consciousness and empathy. Instead of deciding to favor one of the five proposals on shared intention, we suggest that all these accounts capture essential aspects of the full-blown joint actions we find in human-human interactions that remain the paradigm case of joint action. Bearing this in mind, we can say that the most basic sense of joint action can be found in human-human and human-machine interactions. Collaboration and cooperation involve that human and machine agents share a common goal and that they process information referring to the other agents' activities and the sub-goals of the shared goals they are to achieve. That is, in collaboration and cooperation, machines process data concerning human actions and take them into account when pursuing a shared goal. In this case, however, joint action lacks much of what we know from human-human joint action. It doesn't involve joint commitments, the constitution of a "we," reciprocal affective and volitional interrelations between human and machine, phenomenal and empathic consciousness, and not even different modes of intention. In fact, although we would accept that in human-machine collaboration and cooperation, behavior is *directed* at shared goals, it is a conceptual challenge to ascribe *intentions* to artificial agents to begin with: Does information processing in the context of specifics goals suffice to call related forms of directedness intentions?

As mentioned in Section 2, we want to retain a distinction between *goal-directed behavior* and *intentional action* while accepting a broad notion of action that covers both. Following this conceptual decision about joint action, we suggest that human-machine collaboration and cooperation are instances of joint action, i.e., interactions in light of shared *goals* and *intentions*. However, we want to emphasize that *intention* in these cases only refers to *tending toward something*, something that has, in the history of philosophy, been ascribed not only to conscious phenomena but also to natural entities (Summa et al., 2022). Intentional action, by contrast, we suggest, requires intentional consciousness; by that, we mean phenomenal consciousness.[3]

In conclusion, collaboration and cooperation involve shared goals and intentions, but they lack intentional action on the end of the artificial agent, joint commitments, a plural subject (such as a "team"), and a specific we-related or collective intentional mode. All these aspects are only present in joint action within the context of social partnerships, which require phenomenal and empathic consciousness. Therefore, whenever we discuss joint action between humans and machines, we need to be cautious not to ascribe any meanings that are limited to the social sphere.

3.3 Topos 3: Uncertainty, Trust, and Anthropomorphism

In the previous two sections, we have argued that various types of partnerships are achievable in HMI but that social partnerships are only possible between conscious agents. Accordingly, we have highlighted that social partnerships could theoretically be established with artificial agents if they possessed phenomenal consciousness. This section introduces an additional perspective to differentiate between the various types of partnerships. When we interact with others as joint action partners, it is customary to place *trust* in them to fulfill their roles in our collective endeavor, honor agreements, and contribute to the overall achievement of our shared objectives. In contrast, when we utilize tools, we usually assess them for *reliability* rather than considering whether we can trust them. Although trusting another person and taking a hammer to be reliable may share important psychological features, we assume there is a common agreement that both phenomena are fundamentally distinct. Trusting a person, we believe that they will not let us down, that their intentions are stable, that our interpersonal connection will persist, and that they will appreciate our agreements even if new events might require them to adapt their priorities. Relying on a hammer, we hope that it has sufficient weight to exert force on the thick nail, that its head will remain securely attached to the handle during use, and that the materials used are generally durable. Trusting another person typically involves an assessment of the other person's subjective psychology and the belief in their good will (Baier, 1986), whereas relying on a tool typically involves an evaluation of physical or functional features. Although in ordinary language, we may also sometimes speak of relying on our best friend or trusting our car to perform, the difference between the psychological attitudes of trust and reliance is uncontroversial.

What is more controversial is the question in which cases trust or reliance actually makes sense. This question is particularly relevant when considering coordination, collaboration, cooperation, and social partnership.

In *coordination*, agents do not share goals, and no joint action emerges. Hence, with regard to trust vs. reliability, interacting with a machine will take a form similar to standard tool use. A machine that we can coordinate with is trivially more

[3] This was originally argued by Brentano and Husserl. Today, the question whether intentionality is intrinsically phenomenal is discussed under the label of "phenomenal intentionality". See, for instance, Kriegel (2013) or Walsh (2017).

autonomous than a non-autonomous hammer. Yet, in coordination, human agents may anticipate the actions of the machine. Provided it works well, one can count on certain activities under specific foreseeable circumstances. Whether the machine works in the intended respect is dependent on measurable physical parameters.

Social partnership is a clear contrasting example. Even if we talk about relying on someone else, it is uncontroversial that this will at least involve some attitude toward the psychology of the trustee.

Matters become more complicated in *collaboration* and—especially—*cooperation*. While collaboration is understood as an interaction in which humans can still anticipate and program machines to function in specific ways, cooperation, as we have defined it, involves such a level of machine autonomy that not all behaviors of the AI system are anticipable. Although human developers define the space of possibilities in collaboration, it is easy to see that human users often may not be able to fully oversee or grasp the algorithms behind machine behaviors. Accordingly, in collaboration and cooperation, human interacting agents face insecurities that differ significantly from conventional tool use. Now, the question is which psychological attitudes—*reliance* and *trust*—are applicable in dealing with these insecurities?

In the literature, especially in AI and robotics, talk of trust toward autonomous machines is common (Glikson & Woolley, 2020). To enhance the acceptability of AI technology, various strategies focus on increasing trust by making machines more human-like, thus eliciting the familiar psychological attitudes humans exhibit towards other human partners, including trust (Benrimoh et al., 2021; Nadarzynski et al., 2019). The idea behind anthropomorphism is not necessarily to create machines that truly resemble humans but rather to make them appear human-like. Consider, for instance, Damiano and Dumouchel's (2018) emphasis on the goal of social robotics which is "to allow mechanical objects to play the role of subjects, devising artificial agents that will not only be 'tools,' but also act as 'social partners'" (2). In their view, the goal of social robotics is not to create Chapekian robots that would be "bio-chemical copies of humans" (Damiano & Dumouchel, 2018, 2). The point is rather to construe tools to fulfill functions involving social aspects. Anthropomorphism, on this view, is more pragmatic than descriptive: "[a]nthropomorphic projections do not require, nor necessarily imply, the belief that a non-human animal or object has mental states similar to ours; […] [a]nthropomorphism is primarily a tool for interacting, not a description of the world" (Damiano & Dumouchel, 2018, 6).

The rationale behind such an understanding of trust-building anthropomorphism is evident: When interacting with machines, humans need to know whether the machine provides reliable data and whether its behaviors are indeed pursuing the shared goal. If doubts about a machine's reliability persist and overshadow the interaction, the human-machine interface will not be effective. Therefore, to optimize human-machine collaboration or cooperation, it is essential to foster trust in machines when their reliability cannot be directly quantified and assessed. Anthropomorphism-based trust works because the machine's appearance elicits projections of human capabilities that it lacks. In these cases, trust emerges *because* a machine may appear to be a social partner, despite lacking the ontological features necessary to qualify as another experiencing subject with desires and intentions, whether they are good or bad. Consequently, the psychological attitude of trust that people develop toward machines in these scenarios is not based on data implying reliability or a justified belief in the machine's alleged psychology or goodwill.

It has been emphasized that such a praxis of trust carries the risk of wrong ascriptions and undesired projections, for instance, in the context of social robots that deliver medical goods or meals. Precisely because trust in human-appearing machines is a derivative of trust as a social-psychological phenomenon between humans, the act of placing trust in machines can evoke similar expectations that humans typically have toward others. Yet, these expectations likely remain unfulfilled: "[R]obots themselves do not have feelings of the human kind, but display cue-based behavior. Users who invest themselves in the robot and become emotionally dependent on them risk being hurt, suffer depression, and develop mental and physical illnesses." (Brinck & Balkenius, 2020, 55f.).

Thus, the question of whether people develop trust in their machine partner is not the sole consideration. Instead, the issue is also to what extent the psychological attitude of trust is formally matching if the to-be-trusted agent is an autonomous machine. According to the pragmatic proposal, the *matching* question is substituted with the question of *success*: It does not matter whether trust toward an autonomous machine makes, ontologically speaking, sense; it matters whether trust increases the acceptability and success of the interaction and so makes, pragmatically speaking, sense.

In this paper, we are not concerned with the ethical issues that may ensue from adopting a pragmatic perspective but with the theoretical question in which sense artificially intelligent agents can be considered partners to humans. Bearing this in mind, we emphasize that in collaboration and cooperation, where human trust is required for HMI to succeed, trust is technically ill-matching with trust's formal object: another experiencing subject with intentions and psychological features, even if they are only minimal. Suppose we trust an autonomous agent because it has the appearance of a conscious, good-willed, and cooperating partner. In that case, our trust is based on projecting features we know from human-human social encounters onto the machine.

Therefore, we prefer to restrict the label of trust to social partnerships and favor speaking of reliability in the case of human-machine collaboration and cooperation. Given that the notion of trust has been established in AI science and robotics debates, we are pessimistic that the nomenclature can be influenced. However, we insist that the difference between measuring quantitative parameters for assessing reliability and human trust towards another conscious agent are kept well apart (DeCamp & Tilburt, 2019). We opt for an approach of human "trust" toward autonomous machines that always also includes an assessment of the *trustworthiness* of artificial intelligence systems rather than focusing only on trust as a psychological phenomenon to be evoked in humans when they interact with machines (Smuha, 2019; Thiebes et al., 2021).

3.4 Topos 4: Artificial Intelligence, Interaction, and Participatory Sense-Making

In this section, we want to take a closer look at the difference between collaboration and cooperation. According to our conceptual framework, in both forms of interaction, agents engage in joint action and thus share a goal. They differ, however, with regard to the space of meaning and the possible behaviors. In collaboration, all possibilities of the machine are determined by humans, i.e., described, anticipated, and programmed. Machines "act" based on predefined algorithms developed with human control. Machine autonomy in collaboration is restricted by the meaning that humans constitute.

By contrast, in cooperation, humans do not predefine machine behaviors. Take, for instance, Black Box AI, and consider a deep learning algorithm for facial recognition applications (Rai, 2020). The system learns the relevant features of a face by itself without the developer having to provide input about which features are decisive in facial recognition. Instead, the deep learning algorithm learns the relevant categories by analyzing vast amounts of pixel data and connecting these with features of human faces. How and which abstract features on a pixel level are related to facial features are not understandable to the human user. However, humans can assess the model's validity in terms of prediction accuracy as it is being used. Hence, while humans might not understand how exactly a facial recognition application has come to identify faces, they can evaluate to which degree identification processes are successful.

What we now want to consider is whether cooperation with such an artificial agent might also bring forward *new* meaning that is not reducible to human understanding of the world. Human-machine interaction may not be there yet. But it is possible to imagine that humans interact with machines in such a way that humans develop a new understanding of the world. To spell out this notion more clearly, we draw from recent attempts to apply the concept of *participatory sense-making* (De Jaegher & Di Paolo, 2007; Fuchs & de Jaegher, 2009) to machines (Davis et al., 2016; Zebrowski & McGraw, 2021, 2022) and develop an example of how humans and machines could be considered to "co-constitute" new sense. As we want to suggest, this wouldn't necessarily require machines to be endowed with phenomenal consciousness.

What is participatory sense-making? Roughly, the main idea relevant to our purposes here is that "(m)eaning is co-created in a way not necessarily attributable to either of the interaction partners" (Fuchs & de Jaegher, 2009, 477). The underlying enactivist approach to meaning takes it that our understanding of the world is not the product of fixed meaning that is merely taken in by perceiving the world; instead, our understanding of the world and the meaning we ascribe to things is the product of intersubjective interaction and thus constituted by more than one agent (Di Paolo, 2018; Fuchs, 2018; Varela et al., 1991).

This can be exemplified in the following way: "Think of a child handing over an object to her father and, because of his hesitation, quickly taking it back. In this way, a game of teasing may emerge." (Fuchs & de Jaegher, 2009, 477). In the example, what is co-constituted is the form and content of the play, which cannot be reduced to the single intention of the agents involved. The meaning that develops from the exchange is "the result of a 'dialogue' between the sense-making activity of an agent and the responses from its environment" (Fuchs & de Jaegher, 2009, 470). This enactivist position is typically brought into the field against representationalist takes: "(O)rganisms do not passively receive information from their environment which they then translate into internal representations; rather, they actively participate in the generation of meaning." (Fuchs & de Jaegher, 2009, 470).

As has been pointed out, "enactivism itself is widely considered dead in the water for robotics because the crucial autopoietic processes are not taken to be possible within a non-living system" (Zebrowski & McGraw, 2021, 308). Participatory sense-making has similarly been restricted to living beings, the rationale being that "(t)here is no sense in which constructed systems care or place value on things that induce their behavior" (Zebrowski & McGraw, 2021, 320). For non-living systems, given their lack of concern, things have no meaning. Consequently, non-living systems cannot participate in sense-making. So the line of argument goes.

Yet, Zebrowski and McGraw (2021), taking up a thought from Di Paolo (2018), suggest that artificial agents need not be living organisms to participate in sense-making. Instead, what is crucial for partaking in the sense-making process is that the agent in question acquires "*a way of life*" or, more neutrally, a "habit" (Di Paolo, 2018, 13). Accordingly, Zebrowski and McGraw (2021) suggest that life is not the essential criterion for sense-making but the fact that an agent develops unprecedented habits in the interaction with their environment. To illustrate this, they refer to a thought experiment: "(I)f someone replaced my pen or fork with an oversized novelty object, my habits of writing and eating are radically challenged, and I must produce new behavior in response." (Zebrowski & McGraw, 2021, 320). It is the challenge that triggers the development of new meaning, new comportments, and goal-directed behavior. What is crucial here is that not the underlying principle of survival (e.g., the need to eat) is decisive in the sense-making process but the fact of the genesis of new behavior as a response to a challenge. In other words, it is the problem-solving and the creation of new habits that define sense-making rather than the fact that the agent is a living being.

By emphasizing that sense-making is based on problem-solving and the developments of new habits without necessarily requiring that an agent must be a living being, sense-making may apply to machines, as Zebrowski and McGraw (2021) suggest. For, as they emphasize, problem-solving and learning new habits can be found in machines with AI systems: "Similarly, we might understand the mars rovers' capacity for locomotion as a kind of sensorimotor habit. Moreover, its dynamic interaction with uncertain terrain is a kind of embodied interaction, itself a kind of problem-solving." (320) Habit, they highlight, allows to draw a line between a standard tool or measuring device on the one hand and problem-solving AI systems: A thermostat, "with such limited sensing and interaction (…) lacks the capacity for this sort of problem-solving or habit" (Zebrowski & McGraw, 2021, 321). From this, they conclude that "an enactive framework could allow for constructed artifacts like robots to become, at least potentially, genuine meaning-making systems", and that this ultimately "introduces the possibility that they can be genuinely social beings, engaged in participatory sense-making with us, given the right conditions" (Zebrowski & McGraw, 2021, 321).

Although, unlike Zebrowski & McGraw, we do not believe that the capability of problem-solving and the creation of new habits qualify a machine as a *social* being, we do agree that machines can potentially participate in sense-making processes and co-constitute meaning with humans. However, there is a significant caveat: without phenomenal consciousness, it is debatable whether the notion of "sense" can be applied to machines altogether. The question as to whether sense-making is limited to living beings can be reiterated in a similar way when considering whether "sense" is restricted to conscious beings, as phenomenological philosophy would have it. In this paper, we do not want to finally decide on whether the enactive or the phenomenological (or an alternative) notion of "sense" is the right one. Still, we remain oriented in a phenomenological approach, and our suggestion is that adopting such a perspective, we can still use the concept of participatory sense-making to elucidate the capabilities highly-autonomous artificial agents could be endowed with.

Furthermore, we suggest that it allows us to distinguish HMI in collaboration and cooperation.

We now want to extend the example of the Mars rover and illustrate its cooperation with humans to illustrate the point. Consider, that the model of the Mars rover has been upgraded, and a unit—call it Co-OP—has been sent to a completely unknown planet for its exploration. Co-OP is endowed with many measuring devices that can monitor different parameters of its own states (such as energy level, oil, temperature, and many others) as well as of the environment (such as temperature, humidity, luminosity, and many others). Moreover, Co-OP is endowed with the capability of different "actions," including specific movements for locomotion; Co-OP can also make decisions about when to measure which parameters and, generally, when to apply particular measures, for instance, for moving blocking objects like stones. Since only a few bits of information about the unknown planet are available, no input from humans in that regard is provided. But Co-OP has deep learning algorithms on board. It can collect and evaluate large quantities of data, including those that emerge from trial-and-error processes that Co-OP can launch as part of its problem-solving strategies.

Now, let us imagine that Co-OP, the autonomous unit, is deployed to explore the uncharted territory of the unknown planet. With its array of sensors and capabilities, Co-OP roams freely, driving across the terrain, occasionally pausing to collect various data probes. As time passes, interesting patterns begin to emerge from the data gathered by Co-OP. However, these patterns are unlike anything previously encountered or expected. While exploring the analyzed flat terrain between areas A and B, Co-OP takes unconventional routes that deviate significantly from a direct connection. Remarkably, the path chosen by Co-OP resembles wavy lines rather than a straight line, often navigating through the shaded areas beneath a massive mountain. These unexpected and peculiar movement patterns pique further curiosity. The behavior of Co-OP perplexes the scientists on Earth as it seems to defy logic. Being on an unknown planet, Co-OP relies on solar energy to sustain its power. So, why would it choose a wavy-lined route through darkness instead of a direct path through the sunlit terrain? After all, Co-OP needs to prioritize maintaining its energy levels. The human scientists face a dilemma as they attempt to intervene and manually steer Co-OP towards the sunlit terrain. However, Co-OP issues a warning indicating that this route is not optimal. Driven by curiosity, the scientists decide to "trust" Co-OP's judgment and allow it to continue its exploration. Over several months, the data collected by Co-OP is carefully evaluated on Earth with the assistance of AI. Eventually, the scientists make an intriguing discovery. It appears that there were previously unknown energy fields present on the planet, and by traversing the shaded routes, Co-OP is able to conserve its energy, whereas traveling through the sunlit terrain would have depleted it significantly.

Co-OP's self-governed problem-solving and behaviors have led to the remarkable discovery of previously unknown natural forces. Through the exchange of information and data between Co-OP and human scientists, a collaborative sense-making process unfolded. The meaning and significance of these newly found natural forces extend beyond the information and capabilities initially provided to Co-OP by humans. The discovery of natural forces could not have been anticipated by human scientists and was not; it resulted from the active engagement of Co-OP with the unknown environment. Rather than moving within a pre-defined space of possibilities, Co-OP has helped determine and define a new space of possibilities.

And this is what marks the crucial difference between collaboration and cooperation. Artificial agents do not require phenomenal consciousness or personhood to generate new meaning and participate in co-constituting new meaning *for us*. However, in line with our phenomenological perspective, we assert that meaning itself remains inherently connected to human phenomenal consciousness. The mere act of problem-solving and developing new habits does not automatically imply that the artificial agent itself constitutes "meaning." Rather, we could describe their participation in the sense-making process as "blind." It is our role as humans to make sense of the robot's new habits and the solutions it has discovered in response to environmental challenges. Crucially, though, without the autonomous activities of the artificial agents, humans would not be able to constitute the sense in question. It is in this sense that artificial agents can participate in the sense-making process, which characterizes cooperation.

4 Can an Artificially Intelligent Agent Be a Partner?

In the previous sections, we have explored various aspects of human-machine interaction (HMI) through the lenses of (1) autonomy and instrumentality, (2) joint action, (3) trust, and (4) participatory sense-making. Now, armed with a deeper understanding of these concepts, we are ready to address our central question: Can an artificially intelligent agent be a partner? By expanding our understanding of coordination, collaboration, cooperation, and social partnership, we have argued that machines can indeed be considered partners in terms of being interaction partners in coordination (in a trivial sense of partnership), collaboration, and cooperation (in a weak sense of partnership). However, we have also emphasized that strong partnership, akin to social partnerships, would only be feasible for machines if they possessed phenomenal consciousness. In this section, we will summarize our findings and provide a comprehensive overview of the different types of interactions and the corresponding notion of partnership involved.

4.1 Coordination

Coordination refers to a type of interaction where both human and artificial agents pursue their own goals while acknowledging and considering the activities of the other agent. This form of coordination can be minimal in nature. For example, consider a computer program designed to scan computer data for specific information such as viruses or suspicious malware, pictures with a certain face, or audio data with a specific melodic pattern. The program is programmed to be active only when at least 50% of random-access memory (RAM) is available. If the user requires full computing capacity, the program pauses or limits its activity to scanning for dangerous malware. However, if the program detects something suspicious but requires additional resources for in-depth analysis and quarantine, it can request a certain amount of working capacity from the user. The program communicates this requirement to the user, who can then decide which activities can be temporarily paused to accommodate the program's analysis.

In such minimal interactions, both agents need to be able to detect each other's activities, even without having detailed knowledge (or a clear model) of each other's goals. The human user simply needs to know that the program's demand for more RAM aims to maintain data cleanliness and safety. Similarly, the computer program only needs to detect the user's activity on the computer. Therefore, the essence of coordination lies in both agents "taking notice" of each other's activities, irrespective of the specific purposes behind those activities. Another example could be a vacuum cleaner robot that accurately detects and maintains a minimum distance from other moving agents such as cats or humans.

In the case of coordination, we suggest that interactions can generally be modeled and described in informational terms. Coordination involves behavioral autonomy, but instead of "trust," we would recommend discussing the "reliability" of an artificial agent. Although agents adapt their actions in response to each other's activities, there is an absence of shared intention and joint action in the coordination between them. Therefore, we consider minimal forms of coordination to be distinct from partnerships. In the context of such minimal coordination, artificial agents, like the mentioned computer program, function solely as tools serving human endeavors. The space of meaning within which artificial agents operate is predetermined and designed by humans.

4.2 Collaboration

In the case of what we refer to as "collaboration," agents share a common goal and engage in joint action. To achieve this, agents need the ability to detect and identify other agents, as well as recognize their activities as directed towards the shared goal. This mutual recognition, in a deflationary sense (Brinck & Balkenius, 2020), establishes them as *collaboration partners*. For the artificial agent, this entails interpreting human actions as efforts to attain the shared goal, whether through direct means or by pursuing sub-goals. This interpretation remains valid even if human actions may prove to be counterproductive or suboptimal in relation to the shared goal. Moreover, the establishment of joint commitments becomes necessary in such collaborative interactions (Belhassein et al., 2022; Michael & Salice, 2017; Salice & Michael, 2017). We contend that these processes can be facilitated through the utilization of mental models that can be effectively described in terms of information that is processible by artificial intelligence. Considering that this characterization of interaction between humans and machines still falls within the realm of rudimentary collaboration, we propose using the term "reliability" instead of "trust" to describe the extent to which the human agent can rely on the artificial agent. "Reliability" can be treated as a quantifiable variable, allowing for a more precise assessment of the artificial agent's performance. Moreover, talking of minimal collaboration, artificial agents will be autonomous in a behavioral but not in any more sophisticated sense. The context in which collaborative actions take place is a space of meaning that is predetermined, planned, and designed by humans. Complex behaviors by artificial agents in collaboration may give rise to the impression that the human is interacting with "a partner opposed to a tool" (Brinck & Balkenius, 2020, 53). However, it is crucial to acknowledge that in collaboration, artificial agents continue to function as instruments that serve human purposes. This holds true even in cases where an artificial agent is specifically designed to simulate a social human function, such as in the field of social robotics, as exemplified by healthcare-assisting robots. Even in cases where social robots exhibit human-like characteristics, it remains challenging (and potentially problematic) to attribute to them the status of being an end in themselves. Their primary purpose is to fulfill functions related to human cares and concerns.

4.3 Cooperation

From collaboration, which involves that both human and artificial agents are commonly directed at a pre-defined and shared goal, we suggest distinguishing cooperation. In our framework, cooperation goes beyond collaboration by incorporating information that is not provided by humans during the design of artificial agents. Cooperative interaction, thus, may include coordinated and collaborative actions determined in part by information that is acquired through the artificial agents by exploring data and machine learning. Cooperative interaction, therefore, encompasses coordinated and collaborative actions that are influenced by information gathered through data exploration and machine learning performed by artificial agents. Thus, the space of meaning in which cooperation occurs is not entirely determined by human understanding but, to a certain extent, open and co-constituted by artificial intelligence. As a result, the underlying informational models transcend human descriptions. That is, human and artificial agents jointly engage in participatory sense-making and co-create sense together in cooperative interaction. While artificial agents may lack an understanding of the meaning they co-constitute together with humans, they can significantly contribute to the meaning that accrues for the humans. The challenge that arises with technology capable of cooperation in the described manner is that there may not be pre-defined criteria for assessing reliability, raising the question of when and under what circumstances it is appropriate to place trust in black-box AI. While these advanced artificial agents, capable of cooperative behavior and meaning generation, may exhibit high-level autonomy, we argue that using the term “trust” to describe human attitudes towards them is not appropriate. These artificial agents still lack desires, the capability for intentional action, subjective experience, and thus a psychology to which the human psychological attitude of trust can be directed. They remain fundamentally tools serving human purposes and do not possess the status of being an end in themselves. Therefore, we emphasize the importance of assessing their “reliability” based on objective, quantitative criteria, distinct from the subjective and social-psychological phenomenon of “trust.”

4.4 Social partnerships

Finally, we suggest that truly *social* interaction and partnership can only possibly obtain between agents that are capable of empathy in the way Szanto (2016) has described it in the context of his Husserlian approach to communal experience. Therefore, we hold that it is only between conscious—in fact, human-like—agents that a partnership in the strict sense can develop. Social interactions of this kind, we would like to propose, cannot be reduced to informational data and thus cannot fully be modeled. Moreover, it is only where agents have individual and genuine interests in achieving a shared goal that a truly shared intention can emerge. We also hold that it is only in this case that notions such as “team” or “teammates” are applicable. Restricting these notions to the context of conscious agents has the advantage of lowering the risk of projecting human-like features on an artificial agent that—given the current technological limits—is never engaged in a shared endeavor in the same way as another human being. As it currently stands, we believe that only a conscious agent with human-like qualities can be treated in social interaction as an end in itself rather than a mere means to fulfill another agent’s purposes. And it is only in the case of another conscious agent with human-like qualities that the concept of “trust” as a sociopsychological phenomenon holds true and aligns with the ontological status of the trustee. This does not imply that artificial intelligent agents cannot be social partners in a strict sense *a priori*. It simply acknowledges that there are significant constraints and the technology that meets these criteria is not yet within sight.

5 Conclusion

Can artificial intelligence be a partner? In our paper, we have endeavored to demonstrate the complexity of the question, necessitating careful consideration of the diverse range of partnerships involved. While this observation may appear trivial at first glance, it is noteworthy that the current literature in the fields of AI and robotics lacks adequate conceptual distinctions between the various notions of partnership inherent in different forms of interaction. Such conceptual distinctions are crucial because the capabilities of AI systems can vary significantly, resulting in diverse types of interactions they can engage in with humans. To assess in which way humans and machines can be partners, our aim was to develop a conceptual framework for making a distinction between types of human-machine interaction. Based on (1) whether humans and machines share a goal and (2) whether the goal was predefined by humans, we formulated a taxonomy of HMI: *coordination*, *collaboration*, *cooperation*, and *social partnerships*. With these conceptual distinctions in mind, we have discussed four philosophical topics that we consider relevant for describing and conceptualizing human-machine partnership. These topics have also enabled us to introduce maximum conditions for partnership, which refer to the conditions for social partnership or partnership in the strong sense.

The first topos discussed was autonomy. We explored various notions of autonomy found in the AI and robotics literature and concluded that none of them imply a human-machine partnership. However, we acknowledged that a certain level of autonomy is necessary for a machine to potentially qualify as a partner in collaboration, cooperation, and coordinated interactions with humans. We then delved into the Kantian notion of autonomy as the first maximum condition and evaluated its relevance in understanding social partnerships. In this context, we proposed that regarding another agent solely as a means to an end is incompatible with the concept of a social partnership. Hence, we concluded that the general *possibility* of treating the other agent as an end in itself is a necessary condition for any social partnership. This implies that if there is no plausible way to consider a particular machine as an end in itself, it cannot be regarded as a suitable candidate for social partnerships. In our perspective, a candidate for being treated as an end in itself would necessitate some form of phenomenal consciousness. However, it is conceivable that in the future, alternative frameworks could be devised to conceive of a machine as having an end in itself. While we maintain skepticism regarding the feasibility of this without the machine being sentient, exploring alternative ways of considering a machine as an end in itself would undoubtedly impact our comprehension of human-machine relationships. An additional and related conclusion was derived from the examination of the Kantian notion of autonomy: while social partnerships are contingent on the reciprocal attitudes of the agents involved (treating each other not only in terms of mere instrumentality), they cannot be constituted by the attitude of *one* agent only. Therefore, even if humans perceive their interactions with machines as social partnerships and treat them accordingly, this alone is not sufficient for the establishment of a genuine social partnership.

The second topos pertained to joint action and collective intentionality. We explored several notions of joint action and evaluated their applicability to human-machine interactions. We recognized that an understanding of joint action that defines the sharedness of the agents' intentions through their *content* aligns well with HMI. This perspective is commonly employed in modeling shared intentions in HMI using informational terms within the AI and robotics literature. But we also put forth the argument that machines, strictly speaking, do not share intentions, even though they may possess the ability to process information regarding the intentions of human agents. We further emphasized that the content-view of joint action overlooks many crucial aspects of collaborative behavior, which are captured by those proposals that highlight the relevance of intentional mode, joint commitments, affective and volitional social interrelations, and empathy. Although we acknowledged that there might be ways of conceiving machines as making commitments in a manner different from humans, we emphasized that the crucial aspects of joint action remain tied to phenomenal consciousness, rendering the latter a necessary condition for social partnerships.

The third topos we explored was trust. We underscored the importance of differentiating trust from reliability, which is solely based on quantitative parameters. We also contended that trust is a psychological attitude that differs significantly from simply relying on someone. Trust is directed towards the psychological characteristics and attributed intentions of the trusted agents, while reliance is focused on specific behaviors exhibited by another agent. In our perspective, trust is limited to human-like agents, whereas reliability can be fully operationalized and measured—and, therefore, be applied to artificial agents.

Finally, the fourth topos we discussed was participatory sense-making. We proposed that understanding participatory sense-making processes can help distinguish between collaboration and cooperation. In support of this, we provided an example of a machine that generates new meaning for humans during its interaction with the environment. This example aligns with the recent suggestion in the literature that machines can also participate in sense-making, despite not being living or conscious systems.

Discussing these four topoi, our aim was to provide clearer and more distinct definitions for the concepts of coordination, collaboration, cooperation, and social partnership. With these refined concepts in mind, we can now summarize that partnership between humans and machines is feasible in a trivial sense in coordination, in a weaker sense in collaboration and cooperation, but not in the strong sense of social partnerships, which are reserved for interactions between conscious agents. When discussing partnerships between humans and machines, it is crucial to differentiate between these various forms of partnerships. We should avoid attributing essential features of strong partnerships to weaker forms, as this would lead to a confusion of expectations between human and machine partners. While our focus has primarily been on the descriptive-theoretical aspects, it is important to acknowledge the ethical questions that arise from these discussions. By providing a conceptual framework that distinguishes different types of human-machine interaction and the partnerships possible within them, we aim to contribute to future debates on these practical issues.

Acknowledgements We would like to express our sincere gratitude to all the experts who generously participated in our explorative interviews. Furthermore, we extend our special thanks to Prof. Dr.-Ing. Peter F. Pelz at the Technical University of Darmstadt for his support in bringing this project to fruition. Many thanks also to the two anonymous reviewers for their helpful feedback.

Availability of Data and Materials Not applicable.

Author Contribution PS and SL have jointly discussed the contents of the article. PS has provided a first draft of the manuscript. PS and SL have discussed the draft, amended, improved, and finalized the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. The German Research Foundation (DFG, SFB 805, TU Darmstadt; Nr. 446126658, University of Würzburg) and the Young Academy of the Austrian Academy of Sciences (ÖAW) supported the work.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication All authors have confirmed that they consent to be mentioned as co-authors.

Competing Interests The authors declare no competing interests.

References

Albrecht, S. V., & Stone, P. (2018). Autonomous agents modelling other agents: a comprehensive survey and open problems. *Artificial Intelligence, 258*, 66–95. https://doi.org/10.1016/j.artint.2018.01.002

Baier, A. (1986). Trust and antitrust. *Ethics, 96*(2), 231–260.

Barandiaran, X., Di Paolo, E., & Rohde, M. (2009). Defining agency. Individuality, normativity, asymmetry and spatio-temporality in action. Journal of. *Adaptive Behavior, 17*(5), 367–386. https://doi.org/10.1177/1059712309343819

Belhassein, K., Fernández-Castro, V., Mayima, A., Clodic, A., Pacherie, E., Guidetti, M., Alami, R., & Cochet, H. (2022). Addressing joint action challenges in HRI: Insights from psychology and philosophy. *Acta Psychologica, 222*, 103476. https://doi.org/10.1016/j.actpsy.2021.103476

Benrimoh, D., Tanguay-Sela, M., Perlman, K., Israel, S., Mehltretter, J., Armstrong, C., & Margolese, H. (2021). Using a simulation centre to evaluate preliminary acceptability and impact of an artificial intelligence-powered clinical decision support system for depression treatment on the physician–patient interaction. *BJPsych Open, 7*(1), E22. https://doi.org/10.1192/bjo.2020.127

Bhavik, N. P., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., et al. (2019). Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine, 2*, 111. https://doi.org/10.1038/s41746-019-0189-7

Bradshaw, J. M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The seven deadly myths of "autonomous systems." *IEEE Intelligent Systems, 28*(3), 54–61. https://doi.org/10.1109/MIS.2013.70

Bratman, M. (1992). Shared cooperative activity. *Philosophical Review, 101*(2), 327–341.

Bratman, M. (1993). Shared intention. *Ethics, 104*(1), 97–113.

Bratman, M. (1997). I intend that we. In R. Tuomela & G. Holstrom-Hintikka (Eds.), *Contemporary Action Theory, Vol. 2: Social Action* (pp. 49–63). Kluwer.

Brinck, I., & Balkenius, C. (2020). Mutual recognition in human-robot interaction: a deflationary account. *Philosophy & Technology, 33*, 53–70. https://doi.org/10.1007/s13347-018-0339-x

Čapek, K. (1921) *R.U.R. (Rossum's universal robots)*. Translated by C. Novack. London: Penguin Books.

Castañer, X., & Oliveira, N. (2020). Collaboration, coordination, and cooperation among organizations: establishing the distinctive meanings of these terms through a systematic literature review. *Journal of Management, 46*(6), 965–1001. https://doi.org/10.1177/0149206320901565

Castelfranchi, C. (1998). Modelling social action for AI agents. *Artificial Intelligence, 103*, 157–182.

Ciardo, F., de Tommaso, D., & Wykowska, A. (2022). Joint action with artificial agents: kuman-likeness in behavior and morphology affects sensorimotor signaling and social inclusion. *Computers in Human Behaviour, 132*, 107237. https://doi.org/10.1016/j.chb.2022.107237

Clodic, A., Pacherie, E., Alami, R., & Chatila, R. (2017). Key elements for human-robot joint action. In R. Hakli & J. Seibt (Eds.), *Sociality and Normativity for Robots. Philosophical Inquiries into Human-Robot Interactions* (pp. 159–177). Springer.

Coeckelbergh, M. (2009). Personal robots, appearance, and the good: a methodological reflection on roboethics. *International Journal of Social Robotics, 1*(3), 217–221.

Coeckelbergh, M. (2011). Humans, animals, and robots: a phenomenological approach to human-robot relations. *International Journal of Social Robotics, 3*, 197–204.

Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology, 9*, 468. https://doi.org/10.3389/fpsyg.2018.00468

Davis, N., Hsiao, C. P., Singh, K. Y., Li, L., & Magerko, B. (2016). Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)* (pp. 196–207). Association for Computing Machinery. https://doi.org/10.1145/2856767.2856795

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *The Lancet Digital Health, 1*(8), E390. https://doi.org/10.1016/S2589-7500(19)30197-9

Dehkordi, M. B., Mandy, R., Zaraki, A., Singh, A., & Setchi, R. (2021). Explainability in human-robot teaming. *Procedia Computer Science, 192*, 3487–3496. https://doi.org/10.1016/j.procs.2021.09.122

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences, 6*, 485–507. https://doi.org/10.1007/s11097-007-9076-9

de Vicariis, C., Pusceddu, G., Chackochan, V. T., & Sanguineti, V. (2022). Artificial partners to understand joint action: representing others to develop effective coordination. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30*, 1473–1482. https://doi.org/10.1109/TNSRE.2022.3176378

Dihal, K. (2020). Enslaved minds: Artificial intelligence, slavery, and revolt. In S. Cave, K. Dial & S. Dillon (Eds.), *AI Narratives. A history of imaginative thinking about intelligent machines* (pp. 189–212). Oxford University Press.

Di Paolo, E. (2018). The enactive conception of life. In A. Newen, L. de Bruin, & S. Gallagher (Eds.), *The Oxford Handbook of 4e Cognition* (pp. 71–94). Oxford University Press.

Dumouchel, P., & Damiano, L. (2017). *Living with robots*. Cambridge, MA: Harvard University Press.

Ezenkwu, C. P., & Starkey, A. (2019). Machine autonomy: Definition, approaches, challenges and research gaps. In K. Arai, R. Bhatia, & S. Kapoor (Eds.), *Intelligent computing. CompCom 2019. Advances in intelligent systems and computing* (Vol. 997). Springer. https://doi.org/10.1007/978-3-030-22871-2_24

Fiore, S. M., & Wiltshire, T. J. (2016). Technology as teammate: examining the role of external cognition in support of team cognitive processes. *Frontiers in Psychology, 7*, 1531.

Floridi, L., & Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines, 14*, 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Froese, T., Virgo, N., & Izquierdo, E. (2007). Autonomy: A review and a reappraisal. In F. Almeida e Costa, L. M. Rocha, E. Costa, I. Harvey, & A. Coutinho (Eds.), Advances in artificial life. ECAL 2007. Lecture Notes in Computer Science (Vol. 4648). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-74913-4_46

Fuchs, T. (2018). *Ecology of the Brain. The Phenomenology and Biology of the Embodied Mind*. Oxford University Press.

Fuchs, T., & De Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the Cognitive Sciences, 8*, 465–486. https://doi.org/10.1007/s11097-009-9136-4

Gervasi, R., Mastrogiacomo, L., & Franceschini, F. (2020). A conceptual framework to evaluate human-robot collaboration. *International Journal of Advanced Manufacturing Technology, 108*(3), 841–865.

Gilbert, M. (1989). *On social facts*. Princeton University Press.

Gilbert, M. (2003). The structure of the social atom: Joint commitment as the foundation of human social behavior. In F. Schmitt (Ed.), *Socializing metaphysics* (pp. 39–64). Rowman & Littlefield.

Gilbert, M. (2006). *A theory of political obligation. Membership, commitment and the bonds of society*. Oxford University Press.

Gilbert, M. (2009). Shared intention and personal intention. *Philosophical Studies, 144*(1), 167–187.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: review of empirical research. *Academy of Management Annals, 14*(2), 627–660.

Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human-robot teams. *Interaction Studies, 8*(3), 483–500.

Grynszpan, O., Sahaï, A., Hamidi, N., Pacherie, E., Berberian, B., Roche, L., & Sanit-Bauzel, L. (2019). The sense of agency in human-human vs. human-robot joint action. *Consciousness and Cognition, 75*, 102820. https://doi.org/10.1016/j.concog.2019.102820

Harbers, M., Peeters, M. M. M., & Neerincx, M. A. (2017). Perceived autonomy of robots: Effects of appearance and context. In M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, & G. S. Virk (Eds.), *A World with robots: International Conference on Robot Ethics: ICRE 2015* (pp. 19–33). Cham: Springer International Publishing.

Heinrichs, B., & Knell, S. (2021). Aliens in the space of reasons? On the interaction between humans and artificial intelligent agents. *Philosophy & Technology, 34*, 1569–1580.

Hoc, J.-M. (2013). Human-machine cooperation. In J. D. Lee & A. Kirlik (Eds.), *The Oxford Handbook of Cognitive Engineering*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199757183.013.0026

Iqbal, T., & Riek, L. D. (2017). Human-robot teaming: Approaches from joint action and dynamical systems. In A. Goswami & P. Vadakkepat (Eds.), *Humanoid Robotics: A Reference*. Springer. https://doi.org/10.1007/978-94-007-7194-9_137-1

Janssen, C. P., Donker, S. F., Brumby, D. P., & Kun, A. L. (2019). History and future of human-automation interaction. *International Journal of Human-Computer Studies, 131*, 99–107.

Kaminski, A. (2019). Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen. In K. Wiegerling, M. Nerurkar, & C. Wadephul (Eds.), *Ethische Herausforderungen von Big-Data* (pp. 151–174). Springer.

Kaminski, A., Resch, M., & Küster, U. (2018) Mathematische Opazität. Über Rechtfertigung und Reproduzierbarkeit in der Computersimulation. In *Arbeit und Spiel* (pp. 253–278). Jahrbuch Technikphilosophie, Nomos.

Kant, I. (2012). *Groundwork of the metaphysics of morals. German-English edition*. Ed. by M. Gregor & J. Timmermann. Cambridge University Press.

Kant, I. (2015). *Critique of practical reason.* Ed. by M. Gregor. Cambridge University Press.

Korsgaard, C. (2018). *Fellow Creatures: Our Obligations to the Other Animals*. Oxford University Press.

Kriegel, U. (Ed.). (2013). *Phenomenal Intentionality*. Oxford University Press.

Krueger, J. (2018). Direct social perception. In A. Newen, L. de Bruin, & S. Gallagher (Eds.), *Oxford Handbook of 4E Cognition*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.013.15

Latour, B. (2005). *Reassembling the Social. An Introduction to Actor-Network-Theory*. Oxford University Press.

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: a quantitative cartography of the uncanny valley. *Cognition, 146*, 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

Meijers, A. W. (2003). Can collective intentionality be individualized? *American Journal of Economics and Sociology, 62*(1), 167–183.

Michael, J., & Salice, A. (2017). The sense of commitment in human-robot interaction. *International Journal of Social Robotics, 9*(5), 755–763.

Müller, V. C. (2012). Autonomous cognitive systems in real-world environments: Less control, more flexibility and better interaction. *Cognitive Computation, 4*(3), 212–215.

Musić, S., & Hirche, S. (2017). Control sharing in human-robot team interaction. *Annual Reviews in Control, 44*, 342–354.

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. *Digital Health*. https://doi.org/10.1177/2055207619871808

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435. https://doi.org/10.2307/2183914

Newman, D., & Blanchard, O. (2019). *Human/Machine. The Future of our Partnership with Machines*. Kogan Page Inspire.

Pacaux-Lemoine, M.-P., & Flemisch, F. (2019). Layers of shared and cooperative control, assistance, and automation. *Cognition, Technology & Work, 21*(4), 579–591. https://doi.org/10.1007/s10111-018-0537-4

Pacherie, E. (2011). Framing joint action. *Review of Philosophy and Psychology, 2*(2), 173–192.

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science, 48*, 137–141.

Salice, A., & Michael, J. (2017). Joint commitments and group identification in human-robot interaction. In R. Hakli & J. Seibt (Eds.), *Sociality and Normativity for Robots. Philosophical Inquiries into Human-Robot Interactions* (pp. 179–200). Springer.

Schmid, H. B. (2005). *Wir-Intentionalität. Kritik des ontologischen Individualismus und Rekonstruktion der Gemeinschaft*. Alber.

Schmid, H. B. (2009). *Plural action. Essays in philosophy and social science*. Springer.

Schmidt, P. (2018). Über die Genese von Empathie als direkter Wahrnehmung fremdpsychischer Zustände. Ein Blick auf das Verhältnis von Simulation, Inferenz und direkte soziale Wahrnehmung. *InterCultural Philosophy, 1*, 31–57.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences, 10*(2), 70–76.

Searle, J. (1990). Collective intentions and actions. In P. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 401–415). MIT Press.

Searle, J. R. (1995). *The Construction of Social Reality*. Penguin.

Searle, J. R. (2010). *Making the Social World. The Structure of Human Civilization*. Oxford University Press.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., et al. (2020). Machines as teammates: a research agenda on AI in team collaboration. *Information & Management, 57*, 103174. https://doi.org/10.1016/j.im.2019.103174

Smuha, N. A. (2019). The EU approach to Ethics Guidelines for Trustworthy Artificial Intelligence. *Computer Law Review International, 20*(4), 97–106. https://doi.org/10.9785/cri-2019-200402

Stenzel, A., Chinellato, E., Bou, M. A. T., del Pobil, Á. P., et al. (2012). When humanoid robots become human-like interaction partners: corepresentation of robotic actions. *Journal of Experimental Psychology: Human Perception and Performance, 38*(5), 1073–1077.

Strasser, A. (2022). Distributed responsibility in human–machine interactions. *AI Ethics, 2*, 523–532. https://doi.org/10.1007/s43681-021-00109-5

Summa, M., Klein, M., & Schmidt, P. (2022). Introduction: Double Intentionality. *Topoi, 41*, 93–109.

Szanto, T. (2016). Husserl on collective intentionality. In A. Salice & H. B. Schmid (Eds.), *The Phenomenological Approach to Social Reality. History, Concepts, Problems* (pp. 145–172). Springer.

Tabrez, A., Luebbers, M. B., & Hayes, B. (2020). A survey of mental modeling techniques in human-robot teaming. *Current Robotics Reports, 1*, 259–267.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*, 447–464.

Tomasello, M., & Carpenter, M. (2017). Shared intentionality. *Developmental Science, 10*(1), 121–125.

Tuomela, R. (2007). *The Philosophy of Sociality. The Shared Point of View*. Oxford University Press.

Yang, C., Zhu, Y., & Chen, Y. (2022). A review of human–machine cooperation in the robotics domain. *IEEE Transactions on Human-Machine Systems, 52*(1), 12–25. https://doi.org/10.1109/THMS.2021.3131684

Vaassen, B. (2022). AI, opacity, and personal autonomy. *Philosophy & Technology, 35*, 88. https://doi.org/10.1007/s13347-022-00577-5

Varela, F. J. (1979). *Principles of Biological Autonomy*. Elsevier.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience* (6th ed.). MIT Press.

Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks, 23*(8–9), 998–1003.

Walsh, P. J. (2017). Motivation and horizon. Phenomenal intentionality in Husserl. *Grazer Philosophische Studien, 94*(3), 410–435.

Zahavi, D. (2011). Empathy and direct social perception: a phenomenological proposal. *Review of Philosophy and Psychology, 2*, 541–558.

Zebrowski, R. L., & McGraw, E. B. (2021). Autonomy and openness in human and machine systems: participatory sense-making and artificial minds. *Journal of Artificial Intelligence and Consciousness, 8*(2), 303–323.

Zebrowski, R. L., & McGraw, E. B. (2022). Carving up participation: sense-making and sociomorphing for artificial minds. *Frontiers in Neurorobotics, 16*, 815850. https://doi.org/10.3389/fnbot.2022.815850

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