Towards an Ontology of Simulated Social Interaction

Varieties of the 'As If' for Robots and Humans

Johanna Seibt

Abstract The paper develops a general conceptual framework for the ontological classification of human-robot interaction. After arguing against fictionalist interpretations of human-robot interactions, I present five notions of simulation or partial realization, formally defined in terms of relationships between process systems (*approximating*, *displaying*, *mimicking*, *imitating*, and *replicating*). Since each of the n criterial processes for a type of two-agent interaction $\mathfrak{I}$ can be realized in at least six modes (full realization plus five modes of simulation), we receive a $(6^n \times n) \times (6^n \times n)$ matrix of symmetric and asymmetric modes of realizing $\mathfrak{I}$, called the 'simulatory expansion' of interaction type $\mathfrak{I}$. Simulatory expansions of social interactions can be used to map out different kinds and degrees of sociality in human-human and human-robot interaction, relative to current notions of sociality in philosophy, anthropology, and linguistics. The classificatory framework developed (SISI) thus represents the field of possible simulated social interactions. SISI can be used to clarify which conceptual and empirical grounds we can draw on to evaluate capacities and affordances of robots for social interaction and it provides the conceptual means to build up a taxonomy of human-robot interaction.

Key words: social ontology, classification of social robots, sociality, types of simulated interaction, taxonomy of human-robot interaction

1 Introduction

At first blush, the notion of a 'social robot' does not seem to make much sense. If we use prototype semantics to interpret the terms involved, we end up with a category mistake–the currently prototypical robot, i.e., the industrial robot performing repetitive movements irrespective of context, is not the kind of entity that could engage

Research Unit for Robophilosophy, Department for Philosophy and History of Ideas, School of Culture and Society, Aarhus University, Denmark

in prototypical social interactions such as reaching an agreement or meeting for a cup of coffee. If we set prototype semantics aside and turn to explicit definitions, matters do not seem to improve at all. For, on the one hand, researchers in "social robotics" advise us to abandon all hope with regard to conceptual fixations within a fast changing field–"the concept of robot is a moving target, we constantly reinvent what we consider to be robot'" (Dautenhahn, 2014). On the other hand, while the term 'social' still is taken to be amenable of definition, so far there is only a research path towards a comprehensive account of human sociality but not yet a theory with stabilized terminology. Research on social ontology in philosophy is yet to come into full contact with efforts by anthropologists, linguists, psychologists, and cognitive scientists to create an integrated study of human sociality (Enfield and Levinson, 2006). In short, we have only begun to determine, based on conceptual and empirical research, what could or should be understood by the qualifier 'social,' and the class of items we wish to qualify with this yet rather vague predicate–the reference class of 'robots'–is under construction, even constitutively perhaps.

This is an unusual situation. The philosopher's uneasiness can only increase, however, upon reading that a 'social robot' is designed to enable people to "interact with it as if it were a person, and ultimately as a friend" (Breazeal, 2002, p.xi). This one-liner nicely encapsulates three assumptions that, from the philosophical point of view, raise conceptual issues of breath-taking significance.

The first issue concerns the idea that human-robot interactions could qualify as instances of *social* interactions–are we justified in applying the term 'social interaction' even when the relevant capacities for sociality are not distributed symmetrically over the interacting systems? In short, could there be, in principle, *asymmetric* social interactions?

The second issue centers on the type of capacities that robotic systems could possibly exhibit. Given the current state of robotics, most of those capacites that we consider distinctively human, such as linguistic communication and intelligent behavior, are merely approximated by robotic systems. But can capacities for sociality at all be *approximated*? Which degrees and kinds of approximation are admissible, and can robots have the relevant approximative capacities?

The third issue concerns differences in the normative dimension of classificatory predicates. The predicates (i) *x is a y* and (ii) *x can be interacted with as if it were some y* carry different semantic restrictions and signal different practical implications. Will the practice of human-robot interaction change the semantics and normative significance of ascriptive predicates (e.g., *x is a person*), i. e., will we come to allow for anything normatively to count as a person if our interactions with it resemble interactions with a person?

The philosophical debate about social robotics so far has concentrated on ethical and moral aspects, either sidestepping conceptual issues or addressing them obliquely within a specific context of application. This focus made good sense, given that the basic design idea of social robotics, i.e., the program of building artifacts that engage humans in interaction types that we used to reserve for humans only, generates a host of urgent questions of policy and legislation. Up until recently philosophy of social robotics mainly engaged in "roboethics" (Veruggio,

2006) understood as an area of applied ethics that discusses benefits and disvalues of social robotics applications in elderly care, schools, or therapeutic contexts (cf. e.g. Sullins, 2008; Sharkey and Sharkey, 2012; Sparrow and Sparrow, 2006; Coeckelbergh, 2010; Vallor, 2011).

Increasingly, however, the scope of robo-ethics has widened to take up general methodological issues, exploring the challenges of social robotics to traditional paradigms of moral agency and moral patiency (cf. e.g. Gunkel, 2012; Coeckelbergh, 2012). As these investigations bring out clearly, supported by empirical research on human-robot interaction (cf. e.g. Kahn et al, 2004a), the phenomenon of human-robot interaction contains conceptual ambiguities that are directly linked to ethical judgments.

We need to conceptually clarify the phenomenon of human-robot interactions in all its diversity, in order to make progress in the professional and public ethical debate about social robots. But which methods should we apply if the traditional essential differences between *res extensa* versus *res cogitans* are dissolving? Postphenomenological philosophy of technology recommends the radical step of abandoning ontology. However, there are good epistemological reasons, in my view, why roboethics should rely on both empirical research and ontology to provide a fine-grained analysis of human-robot interactions. Since public policy and legislation are guided by ethical recommendations, it is problematic if ethical recommendations are based on subjective impressions about what a robot 'is' or 'is not'–it may be presumptuous, for example, to suggest without more detailed ontological and empirical investigations that human interactions with carebots are "too undignified for a human to enter into" since carebots are "totally inadequate surrogate products"(Danish Ethical Council, 2010). There is increasing empirical evidence that in certain contexts of therapy and caretaking (e.g., autism and dementia) humans benefit more from the interaction with robots than from human interaction (Cabibihan et al, 2013; Leyzberg et al, 2011).

More generally speaking, the Western democratic conception of political authority is grounded in the epistemology of discourse-theoretic validation; ethical judgements are to be warranted by 'rational discourse,' which in turn presupposes normative classifications that are established by intersubjective methods of empirical and theoretical inquiry. When postphenomenologists dismiss "ontology" in favor of a postphenomenological or "relational" approach focused on subjective 'takings-as,' they miss out on an important ally that can translate the "relational approach" into the conceptual format needed for the sort of rational discourse that in Western societies still counts as epistemic warrant for ethical and political decision making. While–correctly, in my view–ridding themselves of the murky bathwater of traditional realist substance-attribute' ontology, postphenomenologists inadvertently throw out the baby of pragmatist analytical ontology, which alone is currently capable of developing sufficiently differentiated and intersubjectively justifiable normative classifications of interactions.[1]

[1] Note that pragmatist analytical ontology is not committed to the facticity or even the possibility of rational discourse, just to its utility as regulative idea and regulated praxis.

The following considerations, written from the perspective of pragmatist analytical ontology, are intended to offer clarifications and classificatory tools that should prove useful for the ethical debate about social robotics applications. In addition, the conceptual distinctions offered below may also have heuristic value for the design of empirical studies of human-robot interactions, especially when undertaken in the context of research on social cognition. I will outline a conceptual framework on the basis of which any extant and future type of human-robot (and human-human) interaction can be classified, so that a comprehensive and detailed taxonomy of human-robot interactions can be built up incrementally.

I begin in *Section 1* with a brief reflection on the predication contexts 'x treats y as z,' 'x interacts with y as if it were z,' and 'x takes y as z.' As I explain, these contexts not only govern different modes of constructionality (make-believe, fictionality, conventionality or 'social reality') they also behave differently with respect to precisely those ascriptive terms that social roboticists are wont to use in their descriptions of human-robot interactions. I argue that for conceptual reasons we cannot adopt the–temptingly easy–strategy of treating human-robot interactions as fictionalist analogues to human-human interactions. Thus we are saddled with what I call the 'soft problem in the ontology of social robotics,' namely, the problem of how to describe human-robot interactions, from a second-person and third-person point of view, given that our concepts for human social interactions as such are inapplicable. I suggest addressing the soft problem by developing a theory of simulated social interaction. In *Section 2* I set out on this task and define five notions of simulation, formulated as relationships between process systems: replicating, imitating, mimicking, displaying, and approximating. In *Section 3* I sketch how these distinctions can be used to define for each interaction concept $\mathfrak{I}$ a "simulatory expansion" of $\mathfrak{I}$ and show how the latter can be used to characterize human-robot interactions. In *Section 4* I address the question of which sorts of simulations of a social interaction $\mathfrak{I}$ can qualify as a social interaction. Given that we currently have not one but several competing notions of sociality, I suggest that we should abandon the idea of a dualist distinction between social and non-social interactions; rather, we should conceive of sociality as a matter of degree. I show how we can use the notion of a simulatory expansion of a social interaction $\mathfrak{I}$ to create a taxonomy for human-robot interactions. As I shall explain, this taxonomy promises to enable us to address two tasks in one go. On the one hand, it should be possible to integrate within this taxonomy competing accounts of sociality with more and less restrictive requirements. On the other hand, with the descriptive tools of the taxonomy we can create a comprehensive conceptualization of simulated sociality that allows for differentiated descriptions of human-robot interactions from a second-person and third-person point of view.

2 Against fictionalist deflations of robot sociality

Social robotics creates a practical space that we currently have great difficulty conceptualizing. While adults interacting with social robots display signs of conceptual disorientation, children tend to resolve the conflict pragmatically by using contextual categorizations of robots ('alive enough to be x'), or hybrid categorizations ('mechanical yet with feelings,' cf. Bernstein and Crowley 2008), or practically relate to robots in ways that conflict with the chosen categorization (cf. Clark, 2008; Turkle, 2011). While most researchers might agree that we need a new ontological category (cf. Kahn et al, 2004b) in order to conceptualize human interactions with 'social robots' so-called, it is currently far from clear how we should go about this task. Do we need a new concept for the item we are interacting with, the robot, as a new sort of artificial 'living thing'? Or, leaving the robot in the familiar category of machines, do we need a new concept for our responsive dispositions for interacting with such machines, e.g., as dispositions for machine sociality? Or should both relata, robot and human, retain their traditional conceptualizations and should we introduce a new sort of relation between them: fictional social relations?

These are three basic strategies that traditional substance-attribute ontology has to offer, as well as various combinations thereof. Once we abandon the traditional substance paradigm, however, new systematic options for categories come into view. In this paper I will suggest quite a different line of approach, however. It is not *one or two* new ontological categories that we need in order to conceptualize human-robot interactions, I shall argue, but an entire new classificatory framework for simulated social interactions. In order to motivate this approach I will begin with some general remarks on the idea that human-robot interactions could be *fictional* social interactions.

Humans interact with their environment not only (i) physically but also (ii) symbolically, i.e., representing physical features, and (iii) 'figuratively,' i.e., assigning to representations of physical features new interactive significances. Consider the following three types of 'figurative' interactions of a human with her or his environment (an object, another human, an event etc.),[2] which I introduce here together with their characteristic linguistic forms, i.e., linguistic expressions that are typically used in descriptions of the participant, who experiences the interaction from a second-person point of view, and descriptions of an observer from a third person point of view, respectively.

1. *Make-believe or pretend-play*. The participant describes the object she is interacting with according to the interpretational rules of the make-believe scenario (e.g., 'I tied my horse to a branch'), and the observer describes the interaction *as* make-believe scenario (e.g., 'she treated the stick as a horse').
2. *Fictional interaction*. The participant describes the interaction in accordance with the conventions of the fiction (e.g., 'I greeted the king and was admitted to the court'), the observer describes the participant's behavior together with its signif-

[2] For the sake of simplification I shall throughout this paper assume that an interaction has just two participants (i.e., two human participants, or a human and a robot).

icance relative to the conventions of the fiction (e.g., 'she bowed as if she were greeting a king and he moved his head and right hand as if he were admitting her to approach').

3. *Socially instituted interaction*. The participant describes the interaction in accordance with extant social conventions (e.g., 'I showed him the receipt') and the observer describes the participant's behavior and its social significance (e.g., 'she handed him a piece of paper that counts as proof of payment').

Observer descriptions of make-believe scenarios are linguistically typically signaled by the phrase '*x* treats (considers) *y* as *z* (or: as if it were *z*),' while fictional interactions are typically described by '*x* interacts with *y* as if it were *z*,' and socially instituted interactions are typically characerized by the phrase '*x* takes (object or interaction) *y* to count as *z*.'[3]

With these distinctions in mind, let us now consider the following formulations of the design goals of social robotics and descriptions of human-robot interactions, respectively:

> [1] We interact with [a sociable robot] as if it were a person, and ultimately as a friend (Breazeal, 2002, p. ix).
>
> [2] Ideally, people will treat Kismet as if it were a socially aware creature with thoughts, intents, desires, and feelings. Believability is the goal. Realism is not necessary (ibid.p. 52).
>
> [3] This also promotes natural interactions with the robot, making it easier for them to engage the robot as if it were a very young child or adored pet (ibid. p. 100).
>
> [4] I find people willing to seriously consider robots not only as pets but as potential friends, confidants, and even romantic partners (Turkle, 2011, p.26).
>
> [5] "...social robots–the class of robots that people anthropomorphize in order to interact with them" (Breazeal, 2003, p. 167).

Formulations [2] and [4] describe human-robot interactions as make-believe scenarios where an *object* is treated as something else, or as if it were something else. Humans mainly engage in make-believe projections during childhood (pretend play), but many people allow themselves to entertain the special type of *anthropomorphistic* make-believe projections also in adult life, as a form of conscious, self-ironic sentimentality. Formulation [5] explicitly clarifies that the design of social robotics consciously targets our capacities and dispositions for anthropomorphizing make-believe projections.

Very little indeed is needed to anchor such projections and to allow us to treat an object as a human being or as a companion. You may treat the tree in front of your window as the master of the garden just because it overshadows all other plants of the garden, and you may treat your car as your companion or adversary just because its start-up performance in difficult weather resembles reliable or malicious actions.[4]

---

[3] For the sake of the argument in this section I operate here with a simplified version of Searle's definition of social reality: 'For all kinds Z, instances of kind Z is part of social reality iff there are X, Y, and C: X takes Y to count as a Z in circumstances C,' cf. Searle 2009.

[4] Here I bracket the question whether 'anthropomorphizing' is the right label for make-believe projections of this kind. Treating something as companion or foe does not necessarily imply treating it as human being. Especially if one applies the 'non-exceptionalist' notions of sociality I discuss

Projections of make-believe ('treating *x* as if it were *y*) are based on physical or functional similarities or analogies between, on the one hand, features or doings of an intentional agent, and, on the other hand, static or dynamic features of a natural or artificial item (Walton, 1990). Fictional interactions are also based on similarities or analogies, but there is an important difference between 'treating some *x* as if it were to do *y*' and 'interacting with *x* as if it were *y*.' When I treat my car as companion and greet it, the car does not perform any distinctive behavior in return; in contrast, a NAO-robot reacts to a human greeting with greeting behavior, or even autonomously displays greeting behavior to elicit a human greeting. In other words, make-believe scenarios typically are one-sided analogical projections where only the human agent involved executes the actions in the template of actions and reactions that define the interaction. In contrast, in fictional interactions both agents[5] behave in ways that resemble the actions and reactions prescribed by the interaction template. Typically we connect fictional interactions with role-play, where the relevant fictional conventions are understood by all agents involved. But a fictional interaction can also be said to take place even if one of the agents is not conscious or is not aware of any convention of fictionality being in place. For example, as long as the behavior of my dog resembles a greeting or as response to my greeting, one can say that I interacted with him as if we were greeting each other; and as long as the behavior of a hypochondriac resembles pain relief after medical treatment, one can say that by administering a placebo I interacted with him as if he were in pain. In short, whether a scenario is make-believe or a fictional interaction depends on whether the agents engender occurrences that resemble the actions and reactions of an interaction template.

Applying these considerations to quotations [1] through [5] above, human-robot interactions are described as make-believe in [2], [4] and [5] (here explicitly), and as fictional interaction in [1] and [3] (in the context of this passage Breazeal explains that her robot Kismet is programmed to produce behavioral patterns that make it more "natural" for people to interact with it "as if it were a very young child or adored pet"). None of the quoted announcements of the design goals of social robotics use formulations that are the characteristic indicators of social actions, i.e., the movements of the robot are not supposed to be 'taken as' or to 'count as' certain actions, nor are the participating humans said to exhibit behavior that 'counts as' an action with social significance. In other words, the quoted passages do not describe human interactions with robots as scenarios where a social action *de facto* occurs.

In comparison with other descriptions of the design goals of social robotics in terms of intentionalist vocabulary, where robots are said to "perceive" their envi-

---

below, one might argue that even though human beings are the primary instances of social actors, our long-standing practice of projecting social roles onto natural things and artifacts is a way to *'socialize'* the world, not to 'anthropomorphize' it.

[5] Throughout this paper I use the term 'agent' in the broad disjunctive sense where it refers either to agents proper, i.e., conscious living things that can act on intentions, *or* to inanimate or living items that are causal origins of events that resemble (and thus can be treated as) actions, i.e., the doings of agents proper.

ronment or even are said to be "aware" of it,[6] the strategy of using linguistic forms that express the fictional irrealis of sociality rather than the realis may seem a very useful device of professional caution. But upon a closer look any attempt to deflate the question of robot sociality by using social vocabulary in fictionalist embeddings are bound to be unsuccessful, as I shall argue now.

Consider again formulation [1]. Could we ever interact with anything *'as if it were a person'*? As just mentioned, both make-believe and fictional interaction are based on resemblances or analogies between descriptive aspects of entities and interactions. But the predicate 'person' is not a descriptive predicate. When we call an entity a person, we thereby make certain commitments in the performance of that very utterance–we are not describing features but announce that certain commitments are undertaken. The performative-ascriptive use of language is not limited to promises and explicit declarations–it pervades our vocabulary for social interactions. Importantly, performative-ascriptive predicates cannot be embedded in contexts with fictionality markers. One cannot perform a linguistic-pragmatically and conceptually coherent speechact by uttering 'It is as if I hereby promise you...,' nor 'what I will say now is a bit like a promise...' Similarly, if we treat some *x* as a person we are committed to taking *x* to count as a person–that is, assuming that we wish to abide by linguistic norms and the actions they entail, we *must* interact with *x as* a person.

This is due to two facts of social reality.[7] The first fact is that commitments are strictly 'bivalent'–they are either undertaken or they are not undertaken; pretending to undertake a commitment is simply *to fail* to undertake it. But if we cannot make fictional commitments, we cannot make fictional promises, nor can we interact with someone as if she were a person. At best we can commit ourselves to treat an occurrence as having a certain special significance within a fictional context, but this is a real commitment to abide by special interpretational conventions.[8] Making a promise or treating someone as a person are real social interactions *by virtue* of engaging in a certain declarative practice (e.g., by uttering a sentence) which *is* the process of taking on the relevant commitments. If we address an infant as a person,

---

[6] Cf. Fong et al 2003, p.145

[7] In the context of this paper I take it that these two facts are self-evident elements of the 'logic' of social practices; a more detailed discussion of the semantics of fictional discourse in application to the performative-ascriptive predicates for social and moral roles is in preparation.

[8] A promise given by a fictional character of a stage play–e.g., Romeo's promise to Juliet to return the following night–is a behavior that counts as promise with respect to our actual social conventions. But due to the referential shifts introduced by the context of the stage play, the actor playing Romeo makes no commitment at all beyond the commitment to playing Romeo–the commitment to return is not the actor's fictional commitment but the commitment of the fictional character Romeo. It is an act that in the context of the fiction counts as the promise of the fictional character *x* that *p*–and in this the only coherent sense, I think, we can give to the idea of a 'fictional commitment.' My views here benefitted from many discussions with Stefan Larsen, in the course of his PhD project on the status of robot sociality, see Larsen (2016); Larsen offers a detailed investigation of the fictionalist conventions of the theatre and uses them as heuristic metaphor for the description of human-robot interactions.

for example, we are making a full commitment to certain actions and omissions, in accordance with rights and obligations.[9]

The second, more general reason for why one cannot make fictional commitments are the performance conditions of social interactions. It is a constitutive feature of social interactions that the agents involved can only react to the enactment of behavior since the other agent's intentions as such are not directly accessible. The agents of a social interaction $\Im$ take certain behaviors (their own and that of others) *to count as* social actions of a certain kind (e.g., a greeting, answering, helping) and thus to *imply* that the agents involved have the specific intentions associated with $\Im$. But the factual occurrence of these intentions is not part of performance conditions for $\Im$. To be sure, the action of greeting and the action of pretending to greet are two different actions–but in the context of a social interaction the behaviors involved in each of these actions *count as* the same *social* action, since they both fulfill the relevant performance conditions for the *social* action of greeting. To emphasize, I am not prejudging here the issue of how to define sociality, i.e., which sorts of joint or reciprocal acts of intentionality should occur for a behavior to count as social interaction $\Im$ in context *C*. I am making the pedestrian point that it is an obvious requirement of social praxis that the *performance conditions* of a social interaction–as opposed to its definition–must relate only to behavioral criteria and cannot take intentional states into account.[10] Often we criticize that a social interaction was performed without the intentional state that according to the concept of the interaction should have been part of it–we complain that someone merely 'went through the motions' in his greeting or offered an insincere apology, etc. Such charges about someone's performing a social action inauthentically or insincerely relate to the *quality* of the social action, not to the latter's *occurrence*–we do not question that the social interaction has been performed, but claim that it has been performed badly.[11]

In combination the mentioned two facts of social reality–i.e., that commitments are binary and that the performance conditions for social interactions relate to norm-governed behavior and not to agentive intentions–preclude the possibility of fictional social interactions. A fortiori, they rule out that declarative-performative terms can be used in fictional contexts.

---

[9] Significantly, describing a colleague's interaction with his infant daughter Breazeal keeps the fictional scope correctly focused on the *descriptive* predicate socially fully aware and responsive [agent]': "Simply stated, he treats his infant as if she is already fully socially aware and responsive–with thoughts, wishes, intents, desires, and feelings that she is trying to communicate to him as any other person would" (Breazeal, 2002, p.30).

[10] Unless one champions a purely behaviorist account of social interactions, the definition of a social interaction will include behavioral performance conditions, but also state additional conditions relating to the agent's intentions and understanding of norms.

[11] Since the performance conditions for social interactions relate only to behavior and not to the intentions of the agents involved, behaving as if one were to perform a certain social action A that is part of a social interaction $\Im$ is tantamount to exhibiting behavior that counts as the relevant part of social action $\Im$. The agent might intend to perform another action, e.g., she might intend to do B = [pretend doing A], but this does not detract from the fact that the performance conditions for A have been fulfilled.

In other words, for social interactions *there is no fictionality gap*. Fictionality requires that the real and the fictional domain are in some fashion separated–that there is a 'gap' between the two domains that is bridged by imaginative analogical reasoning, using the real as stand-in for the fictional (Walton 1990, 2005). We may treat a stick as a horse or interact with a robot as if it were a "young child" and yet remain fully aware of the difference in descriptive features between the real stick and the imagined horse, or the robot and the imagined child, respectively. But no such fictionality gap can be kept open for social actions and interactions. This is most palpable when performative-ascriptive predicates are used or implied–whoever addresses some x as a person, by that very utterance undertakes the relevant commitments and thus takes x to count as a person, i.e., engages in a real social interaction. It is important to see, however, that the fictionality gap collapses not only for performative-ascriptive predicates but for all predicates for social actions and thus obviates a fictionalist interpretation of human-robot sociality in general.[12] From a philosophical viewpoint it is thus a category mistake to assume that we can interact with anything "as if it were a person." The fictionalist interpretation of human-robot interactions overlooks that we cannot drag social (inter)actions into the scope of fictionality. We can create contexts of social interaction for fictional interactions, i.e., we can commit ourselves to respect the reinterpretations of a fictional domain, but we cannot fictionalize our commitments. In short, there is social fictionality but no fictional sociality.[13]

To summarize the observations in this sections, I have argued here that we cannot treat human-robot interactions as fictional social interactions, i.e., as interactions where humans engage in fictional commitments to a social interaction. While for physical interactions there is a clear difference between the real and the fictional interaction (e.g., pushing someone and behaving as if one were pushing someone) due to physical dissimilarities, for social interactions the fictionality gap collapses. A human's actions towards the robot cannot be considered as a pretended social action, since the performance conditions of social actions, i.e., the conditions for the performance of a social action from a third-person point of view, are insensitive

[12] In Seibt (2014b) I use the contrast 'friend' vs. 'person' to highlight the difference between descriptive and performative-ascriptive terms–in hindsight, this was infelicitous, since even though the term 'friend' is partly descriptive, the performative-declarative elements of the meaning of friendship arguably are dominant. The predicates 'friend' and 'person' belong on the same side and should be contrasted with (predominantly) descriptive predicates such as 'child' or 'woman.'

[13] That we cannot uphold a fictionality gap for social interactions is also implicitly reflected in concerns about the expectable cultural change effected by social robotics. Authors who warn against the pervasive use of social robots are not worried about humans losing themselves in realms of fictionality. Rather, they fear a *degradation* of social interactions, due to an increased functionalization of social relations (cf. Sharkey and Sharkey 2012) where the performance of connection seems connection enough" (cf. Turkle, 2011, p.26). The worry is that we increasingly will reduce to definition of a social action to its performance conditions, i.e., that we will abandon our current understanding that the concept of a social action comprises both conditions relating to behavior and conditions relating to intentional states. While we currently criticize each other for performing bad social actions by going through the motions, by not living up to the concept of the social action in question, such criticism will vanish–or so the argument goes–and with it the social standards for performing social actions well, i.e., sincerely.

to the difference between authentic and pretended action. As convenient it might seem to describe human interactions with a robot as fictional social actions, upon a closer look such descriptions are conceptually incoherent.

This leaves us with what I call the 'soft problem in robo-ontology.' Once we realize that we cannot use fictionalizations of familiar predicates for human social interactions in order to describe human-robot interactions, we face the following juncture. Extant philosophical definitions of social interactions determine that for a social interaction to occur, all agents involved must have the capacities required for normative agency, e.g., intentionality, consciousness, normative understanding.[14] Thus, going by our standard philosophical definitions, we need to treat human interactions with so-called 'social' robots as non-social interactions. Alternatively, we need to develop new conceptual tools for forms of non-reciprocal or asymmetric sociality, i.e., for social interactions where one agent lacks the capacities required for normative agency.

In the following sections I will pursue the second route and outline a solution to the 'soft problem' by devising a conceptual framework for simulated social interactions. By analyzing forms of simulations and classifying types of interactions parts of which are simulated, we can investigate in a differentiated fashion which, if any, human-robot interactions can qualify as social interactions. As will become apparent, I hope, discussing robot sociality in the terms of the suggested framework can also advance our understanding of sociality in general.

3 Five notions of simulation

The 'soft problem of robo-ontology' introduced in the previous section is best addressed, in my view, by devising a conceptual framework that is rich and discriminative enough to describe human-robot interactions at all stages of research development in robotics. The first step towards developing such a framework is to clarify the notion of simulation.

As we commonsensically understand the term, the predicate 'simulate' states a relationship between occurrences. In order to describe the denotation of the predicate more precisely, I suggest using a process ontology, i.e., an ontology that countenances processes among the basic categories. The following definitions are drafted on the background of a mono-categoreal process ontology called *General Process Theory*, but in the context of this paper it can remain open whether the embedding ontological framework has other basic categories, and how the explanatory predicates of the ontological theory (e.g., 'emerges,' 'is part of,' or 'is an instance of')

[14] For an overview of these requirements of standard accounts of social interaction see Hakli (2014). Hakli discusses what I call 'the soft problem' in a differentiated fashion that also involves larger epistemological perspectives on conceptual change and the conventionality of conceptual contents.

can be defined more precisely.[15] Since in General Process Theory processes are non-particulars, I shall use 'process' and 'process type' interchangeably.

Let 'C' and 'C^*' stand for processes or collections of processes. A collection of processes can occur in such a way that (i) other processes 'emerge' (in the period during which the collection occurs) or else (ii) without such 'emergent products.' Thus we can distinguish two sorts of co-occurrent collection of processes, (i) *complex processes* and (ii) *collections of processes*. Both complex processes and collections of processes may have themselves complex processes (or collections of processes) as parts. If a process system Σ is a collection of processes, the occurrence of C and C^* is merely a mereological sum, i.e., $\Sigma = \text{sum}(C, C^*)$. In contrast, if a process system Σ is a complex process, then there is a basic process, the interaction I of C and C^* that emerges from the interference of C and C^*, i.e., $\Sigma = \text{I}(C, C^*)$.[16] Both complex processes and collections of processes may be parts of complex processes or of collections of processes.

The distinction between a complex of processes and a collection of processes will become relevant below when we consider alternative classifications of processes. For example, if Kim works on a production line next to an industrial robot, she might take herself to work *alongside* of the robot, treating the robot as a feature of the production line, or she might take herself to work *with* the robot in a collaborative situation. This difference will matter when we discuss differences in classifications of social interactions from first, second, and third-person points of view.

In order to allow for such further distinctions the following definitions for modes of simulations are formulated in such a way that they can be applied to process collections with arbitrary dynamic organization. That is, the following relationships may hold for process systems Σ and Σ^* if these are parts of a mere collection of processes, occurring 'alongside,' but also if Σ and Σ^* are parts of a complex process, i.e., parts of an interaction.

The five relationships of simulation I will distinguish here are five ways of deviating from a common baseline, which I call the realization of an action[17]:

Definition 0 (Realization). For any process type A, process system Σ realizes A iff Σ generates an instance of A.

For present purposes, let us say that a process system Σ *generates* a process α just in case α is a (possibly emergent) part of Σ, and that α *is an instance of* process type A just in case α fulfills all (functional and non-functional) occurrence conditions F_i

[15] Relevant technical details, especially also on the part-relation for processes I shall use in the following definitions, can be found in Seibt 2005, 2009, 2014a.

[16] For details cf. Seibt 2014a.

[17] The following five definitions are simplified–and, I hope, thereby improved–versions of those presented in Seibt 2014b. – In the following I use capital letters as variables for process types and Greek letters as variables for instances of a process type. Note, though, that in *General Process Theory* there are, strictly speaking, no 'instances' or 'tokens', since I consider the type-token (kind-instance) distinction to mark extremal regions on a gradient scale of specificity; in order to simplify the exposition here I stay with the traditional idiom and speak of highly specific, localized processes as instances (tokens) of kinds (types) of processes.

for A.[18] For example, two people reading next to each other each realize instances of the action type *reading*.

Any two instances of a process type A differ with respect to at least one aspect that is not among the occurrence conditions F_i for A, e.g., its location. If a process α fulfills all functionally relevant occurrence conditions of A but fails to fulfill at least one the non-functional occurrence conditions of A (e.g., conditions of *A's* biochemical realization), α is no longer an instance of A but an instance of a process type that is functionally equivalent to A. Deviations from non-functional occurrence conditions are the hallmark of the first and strongest mode of simulation.

Definition 1 (*SIM-1*, Functional replication). For any process type A_P realizable by process system Σ, process system Σ^* functionally replicates A_P iff Σ^* realizes a process type C that is distributively functionally equivalent to A_P, or Σ^* realizes subprocesses $D_1...D_n$ of C that are distributively functionally equivalent to all subprocesses $B_1...B_n$ of A_P.[19]

Let me unpack this definition. First, the disjunctive clause in the righthand side of *SIM-1* is to ensure that the definition covers simulations of and by (i) simple processes, (ii) complex processes, as well as (iii) collections of processes. Second, processes X and Y are *distributively functionally equivalent* iff for each part of X there is an input-output equivalent process that is part of Y. Third, the parthood relation on processes is not the transitive parthood relation that holds for spatiotemporal regions, i.e., it is not the case that every part of the spatiotemporal region occupied by a process X is a part of X. Rather, parthood on processes is a non-transitive and context-dependent relation (Seibt 2014a). Claims about parts of a process A are thus always to be understood in relation to a finite partition P that states the mereological structure of A. In order to highlight that claims about functional replication–and other modes of simulation–are relative to a given finite partition of the simulated process type, a subscript 'P' is added to the variable of the process type, i.e, 'A_P'. For example, relative to a fine-grained partition for the process type A = *teaching math*, only an imaginary close-to-perfect android such as *Star Trek's* Mr. Data could functionally replicate A = *teaching math*; on the other hand, relative to a coarse-grained partition of A = *initiating a conversation* one can claim that even present-day educational robots such as Aldebaran's NAO-robot functionally replicate A.

[18] For present purposes I must rely on an intuitive understanding of the theoretical predicates 'functional' and 'non-functional'; a more precise statement of the envisaged distinction is quite involved, especially from within a naturalist metaphysics where all properties, even *qualia*, are in some sense 'functionalized.'– Here I am also neglecting specific issues of realizations that arise when the process in question is an action and intentional scope comes into play. In Seibt (2014b) I formulate the right-hand side of the biconditional as a disjunction: action A is realized by a process system Σ if the system generates an instance of A *or* if Σ realizes instances of all subprocesses $B_1...B_n$ of A. The disjunctive formulation is to account for variations in intentional scope, e.g., the difference between the holistic performance of an action by an expert versus the summative performance of the same action by a novice.

[19] More precisely the definiendum should be formulated as: 'process system Σ^* functionally replicates A_P as realizable in Σ'; here and in the following definitions I take this restriction to be understood.

The definition of functional replication demands maximal similarity of functional structure (to reemphasize: relative to a given partition of the simulated process type). The second mode of simulation, here called ‘imitation,’ loosens this constraint by reducing the requirement of distributed input-output equivalence, i.e., input-output equivalence along all parts of the simulated process, to a requirement of input-output equivalence for most parts. Let us say that processes X and Y are *functional analogues* iff all non-ultimate parts of X have functional equivalents in Y (where a ‘non-ultimate’ part of X is any part that has a part in the partition of X). In other words, processes X and Y are functional analogues just in case they are input-output equivalent and have similar (though not maximally similar) functional structure. Aiming again, as in *SIM-1*, for a definition that can cover simple and complex processes as well as collections (complexes) of processes, we can define a second mode of simulation as follows:

Definition 2 (*SIM-2*, Imitation). For any process type A_P that has non-ultimate parts and is realizable by process system Σ, process system Σ^* *imitates* A_P iff Σ^* realizes a process type C that is a functional analogue of A_P, or Σ^* realizes subprocesses $D_1 ... D_n$ of C that are functional analogues of all immediate subprocesses $B_1 ... B_n$ of A_P.

For example, let us assume that the partition for the process A = *fetching new linen when commanded* lists among the parts of A the process B_1 = *reacting to a command* which has as non-ultimate parts B_2 = *classifying type of speech act* and B_3 = *analyzing semantic content*. Relative to this partition, Aethon’s TUG robot, a mobile, self-navigating delivery device for hospitals that is activated by touchscreen input, cannot be said to functionally replicate process *fetching new linen when commanded*, since TUG does not generate processes that are functionally equivalent to B_1, B_2, and B_3. But TUG generate a process that is functionally analogous to B_1, B_2, and B_3, and thus can be said to imitate A. Or again, to illustrate the second disjunct in *SIM-2*, a multi-robot system of Kilobots imitates the group process A = *forming a V-shape* since the immediate subprocesses of A as realized by, e.g., a swarm of birds, i.e., the local interactions between neighboring birds, have functional analogues in the programmed interactions between the kilobots.

The third mode of simulation, which I call ‘mimicking,’ further relaxes the required input-output equivalence of simulating and simulated process. For the purposes of social robotics, one might argue, the first two modes of simulation are not really relevant since it does not matter whether two processes are input-output equivalent in all regards–all that matters is input-output equivalence relative to the capacities of human observation. Let us say that process Y is an *empirical proxy* of X iff for any observable part of X there is a part of Y that is observably input-output equivalent. A process may then be said to be an empirical proxy of another even if there are considerable deviations in functional structure.

Definition 3 (*SIM-3*, Mimicking). For any process type A_P that has observable parts and is realizable by process system Σ, process system Σ^* *mimicks* A_P iff Σ^* realizes a process type C that is an empirical proxy of A_P, or Σ^* realizes subprocesses $D_1 ... D_n$ of C that are empirical proxies of all immediate parts $B_1 ... B_n$ of A_P.

For example, Hiroshi Ishiguro's *Geminoid* robots, which are physical copies of the bodies of particular people, can be said to mimick some human facial movements, e.g., rolling one's eyes or smiling, since–or so we may assume for the sake of the illustration–these artifical facial movements are empirically indistinguable, down to their smallest observable phases, from their real counterparts in a human face.

By ways of contrast, consider Cynthia Breazeal's robotic head *Kismet*. Kismet also moves some of its components (called Kismet's 'ears,' 'lips,' 'eyebrows,' and 'eyelids') in ways designed to be interpretable as expressions of surprise, delight, disgust, sadness, etc., but here the observable similarity with human expressions of these emotions is reduced to some typical elements. Kismet performs the facial movements familiar from cartoons and animations, and like these it exploits our tendency to 'anthropomorphize' our environment.[20] When Kismet and similar social robotics applications are designed in ways that capitalize on the human tendency to 'anthropomorphize,' the following mode of simulation is targeted.

Definition 4 (*SIM-4*, Displaying). For any process type A_P that has parts and is realizable by process system Σ, process system Σ* *displays* A_P iff Σ* realizes subprocesses $D_1 ... D_n$ of C that are empirical proxies of all and only typical immediate parts $B_1 ... B_n$ of A_P.

Which parts $B_1 ... B_n$ of a process A are 'typical' for A in the sense that a human being will take the occurrence of empirical proxies of these parts as an indication for the occurrence of A I take to be context-dependent; but note that some relevant aspects of Gestalt perception may be innate and some typical cues for behavior that steer the dynamics of social interactions, such as pointing and gaze-following, may be culturally universal (cf. Enfield and Levinson (2006)). Thus claims about the *displaying* of a process A in the sense of *SIM-4* are relative to a partition P whose nodes are annotated for typicality.

Displaying is the target mode of simulation for social robotics applications that belong into the genre of entertainment technology, where smooth social interactions are instrumental for the primary design goal of engaging the user in a game. For example, Sony's robotic dog *AIBO* can be said to display a dogs greeting or a dog's invitation to play.

---

[20] As noted above, fn. 4, instead of 'anthropomorphizing' we should rather speak of a human tendency to 'socialize' the environment. Above I pointed out that the performance conditions for human social behavior cannot take intentional states into account. The phenomenon we commonly call 'anthropomorphizing' indicates, I think, that the performance conditions for social behavior operate with observable criteria that are schematic, involving generic Gestalts. As it appears, judging from our practices of interpretation, the upward position of the corners of a mouth does not need to be an *intentional* smile to count as a smile, nor does it need to resemble a human smile in all regards–a mouse can smile, and so can a sponge, a car, or a tree. That we use such general observable indicators of socially relevant actions and emotional states could be explained in evolutionary terms as follows. Surely it is preferable for humans to risk erroneous inclusions into the space of social interactions rather than erroneous exclusions; if we mistake something for a social agent, the error can be corrected without incurring social repercussions, but if we fail to recognize a social agent as such, this would amount to a social offense.

Other applications, especially in the area of assistive technology where special cooperative efforts on the side of the human can be expected, may settle for the fifth and 'poorest' mode of simulation:

Definition 5 (*SIM-5*, Approximating). For any process type A_P that has parts and is realizable by process system Σ, process system Σ^* *approximates* A_P iff Σ^* realizes subprocesses $D_1 ... D_n$ of C that are empirical proxies of only some typical immediate parts $B_1 ... B_n$ of A_P.

For example, the robotic seal PARO approximates process type A = *a cat's purring* by creating empirical proxies of some typical subprocesses of A such as physical vibrations and sounds of the right wave lengths (but fails to create empirical proxies for other typical subprocesses of a cat's purring, such as the relaxation of muscle tone).

This concludes the distinctions in modes of simulations I wish to suggest here; further distinctions could be added if necessary. As the reader will have noticed, the definitions provided are not exact–they rely on a number of vague predicates ('functional,' 'observable,' 'typical'). In addition, they contain a context-relative parameter: the partition of a process. But, I submit, they capture five relevant intuitive notions of simulation: simulation as 'a process with the same structure occurring in a different medium' (*functional replication*), simulation as 'a process with similar structure that yields the same results as the original process' (*imitation*), simulation as 'a process that yields the same observable results as the original process' (*mimicking*), simulation as 'a process that yields the same relevant observable results as the original process' (*displaying*), and simulation as 'a process that we can treat as observably similar to the original process' (*approximating*).

To illustrate how one might use these distinctions to clarify the capacities of robots, consider the robot *Leonardo*, created by Brian Scassellati (Scassellati, 2002). Leonardo is programmed to produce, for a restricted interaction context, a representation of the beliefs and intentions of the person it is interacting with, to the extent that it can pass an empirical test for the capacity of 'false belief-attribution,' the so-called 'Sally-Ann test.' In the philosophical discussion about sociality, false-belief-attribution counts as a prime example for social cognition. Given that Leonardo passes the Sally-Ann test, precisely what has been simulated: the Sally-Ann test, false-belief attribution, or even social cognition?

To bring the above definitions to bear on this question, assume that we operate with a partition Q for the process A = *attributing-to-X-the-false-believe-that-p* that lists the process B = *passing-the-Sally-Ann-test* as one of the typical parts of A, and as another typical part the process C = *reacting-emotionally-to-the-belief-that-p*, belonging to the affective dimension of the theory of mind (cf. Kalbe et al 2010). Then we should say that Leonardo does not replicate, imitate, mimick, or display, but merely approximates false-belief attribution A–if conceptualized as A_Q–since Leonardo only realizes an empirical proxy for one of the processes that are typical for A. We cannot conclude from this that Leonardo thereby also approximates social cognition, since even if false-believe attribution were a typical part of social cognition, the empirical proxy of a typical part of X which is a typical part of Y is not an approximation of a typical part of Y. But assuming that we conceptualize B, the passing of the Sally-Ann test, in terms of a partition $P*$ that lists as typical parts of B_{P*} the processes D = *uttering 'p' in response to queries about X's belief*, E = *applying certain inference schemata to p, F = uttering 'r' upon queries about an object's location in r*, and G = *tracing the locations of an object*, we can say that Leonardo mimicks the process of passing-the-Sally-Ann-test, since the robot realizes processes that are input-output equivalent to all the typical parts of B_{P*}.

In sum, the five definitions highlight that there are different degrees of simulation relative to structural and observable similarities. They also highlight that claims about one process X simulating another process Y depends on how we describe (partition) process Y. In the case of robotic simulations of human actions the following inverse relationship holds. If action A is described in terms of a coarse-grained partition that predominantly uses predicates for natural and functional process types (e.g., *uttering 'p'* or *inferring*) a robot can be be said to simulate action A to a high degree–it might be able to mimick, imitate, or even functionally replicate the action. *Vice versa*, if a fine-grained partition is used to describe action A and if the partition includes predicates for mental and reflectively normative process types (e.g., *judging* or *noticing*), a robot will likely achieve only low degrees of simulation–at least relative to the current state of technology, it will only be able to display or approximate actions under such a description. The five definitions of modes of simulations thus accommodate the description-dependency of claims about simulation while at the same time allowing us to formulate simulation claims more precisely.[21]

4 Simulatory expansions of interactions

The next step in developing a conceptual framework for simulated social interactions is to analyze the notion of interaction and relate the latter to the five modes of simulation as previously defined. Since there is no ready-made theory of concepts of interaction to draw on, let us set out with some preparatory reflections.[22]

The notion of interaction always signals a contrast to the mere co-occurrence of a plurality of processes. In the terminology introduced in the previous section, an interaction is a complex process as opposed to a collection of processes. However, in common-sense and scientific discourse we use the concept of an interaction in two senses. In its wide sense, the term denotes a complex *non-agentive occurrence*

---

[21] The suggested classification of forms of simulation also should prove useful in the discussion of design issues in social robotics. For example, we can use it to compare and evaluate robotic simulations of an action A_P in terms of degrees of simulation; here one might refine *SIM-5* by introducing degrees of approximation; or it may be used to plan design goals, e.g., in order to decide whether there are any ethical reasons to aim for higher degrees of simulation than mimicking.

[22] Apart from Aristotle's conceptual analysis of interactions (cf. Gill 2004), the most important source and resource for a future ontology of interactions is the work of Mark Bickhard, who combines empirical and conceptual research to promote "interactivism" both as a paradigm for empirical research and as a comprehensive theoretical stance or metaphysical view; cf. e.g. Bickhard 2009a,b and his chapter in this volume.

consisting of two or more non-agentive occurrences (often we use terms for objects or forces to refer to occurrences involved in an interaction). In contrast, in the term's narrow sense, 'interaction' denotes an complex *action* consisting of two or more actions. For example, we use the wide sense when we describe corrosion (the formation of iron oxides) as an interaction between the physical processes that constitute iron atoms and those that constitute oxigene atoms, or define cell metabolism as an interaction between catabolic and anabolic processes. In contrast, a joint action that two or more agents intend to perform as such, such as playing football, discussing, carrying a table, or choire singing, is an interaction in the narrow, agentive, sense.

How do non-agentive interactions (wide sense) relate to agentive interactions (narrow sense)? This is largely unexplored territory in ontology. The question has been touched upon occasionally, e.g., when environmental ethicists consider whether the agents of a collection of actions X_n are responsible for the unintended interactive product of the processes that are, or result from, X_n (cf. e.g. Björnsson 2011; Petersson 2013). But currently we still lack a sufficiently comprehensive theory of multi-agent actions that would relate complex actions, collections of actions, complex non-agentive occurrences, and collections of non-agentive occurrences in all possible combinations.

In particular, we lack theoretical concepts for interactions that cannot be subsumed under either the wide or the narrow sense of interaction. Besides complex non-agentive occurrences whose parts are non-agentive occurrences (the first, wide sense), and complex actions whose parts are actions (the second, narrow sense), there are complex processes which have both actions and non-agentive occurrences as parts, and which may be either intended by an agent (third sense) or not intended (fourth sense).

The third sense of interaction is the primary target in our present context. In accordance with the four different senses of interaction identified, let us speak of *four kinds* of interaction–which each may be further differentiated into subordinate *types* of interaction–and let subscripts ('interaction$_1$' etc.) indicate which of the four kinds of interaction is at issue. We engage in an interaction$_3$ whenever we act in response to natural processes that are generated or shaped by our actions. For example, a basketball player doing solo dribbling drills engages herself in an interaction$_3$–the ball bounces in response to her actions and *vice versa*. Similarly, playing a single-player computer game can qualify as an interaction$_3$ (at least if we consider the events on the screen as natural processes rather than as the actions of the programmer). It is important to note that interactions$_3$ do not need to be initiated by the agent–consider a dynamic flood response, by building dams here and opening channels over there, performed dynamically in response to the spread and flow of the collecting waters. However, whether initiated or not, it is a hallmark of an interaction$_3$ that the agent intends to perform it by performing the actions that are among its parts. Interactions$_3$ thus resemble complex actions (interaction$_2$), and I shall call them 'asymmetric' complex actions.

With these preliminary considerations in place, let us now turn to defining the two notions of interaction that are most relevant for present purposes.

Definition 6 (Interaction$_2$). A process system Σ is an *interaction*$_2$ iff Σ is a complex action A_P and all parts of A_P are (simple, complex, or collective) actions B_i.

Definition 7 (Interaction$_3$). A process system Σ is an *interaction*$_3$ iff Σ is a complex action A_P, some of the parts of A_P are (simple, complex, or collective) actions B_i, and some of the parts of A_P are non-agentive (simple, complex, or collective) occurrences C_i.[23]

Central for these two definitions is the distinction between actions and non-agentive occurrences, which is notoriously difficult to define. In the present context we can circumnavigate this difficulty by operating with the following simple condition. A process A_P is an action iff an agent can intend to do A_P.

To illustrate that this condition is sufficiently discriminative, consider a prototypical interaction$_2$, a strategic boardgame such as chess. Each ‘move’ of the game is a process A_P that one of the players intended to do (under the description A_P) and each player intends to participate in the game *as* an interaction$_2$, i.e., each player considers the processes executed by the other player as actions–as something that an agent can intend to do. In contrast, when two musicians perform a duo, each musician will react to sounds and phrases understood as non-agentive acoustic occurrences (one can intend to do what it takes to produce a certain sound on the instrument, but one cannot intend to sound); nevertheless each musician intends to bring about the interaction as complex process where the musician’s actions respond to and influence sounds from another source. Thus a duo is the simultaneous occurrence of two interactions$_3$, generated by two musicians.

These illustrations show, first, that the suggested requirement for actions is subtle enough to dissociate the question ‘which parts of the interaction are actions?’ from the question ‘which kinds of entities, humans or non-humans, cause the processes that are parts of the interaction?’ Second, the examples also make plausible that interactions$_3$ are complex actions rather than complex occurrences–in this type of interaction the agent(s) who intend(s) to do the actions B_i that are part of A_P, also intend(s) to do A_P.

Precisely what is involved when we intend to do A_P has been much debated in the philosophy of action for the last half century it least. But even the least demanding conceptions of intentions take the latter to involve (at least) three abilities: (i) the ability to form beliefs about the process type A_P as a standard for action and interaction, i.e., as a template all parts of which *should* be realized if one wishes to realize A_P, possibly as an element within an interaction context, (ii) the ability to

[23] Since the two definitions are to be as generic as possible, no requirements for dynamic, temporal, or spatial relationships among the parts of an interaction have been added. That is, I am assuming here that the processes that are the parts of an interaction may occur all or partly simultaneously, with overlap, or sequentially in series. But we can easily introduce types and subtypes of interactions$_i$ by specifying which temporal, spatial, and dynamic relationships need to hold among the parts of the interaction in question. To simplify the exposition I omit here and in the following definitions specifications of the partition levels at which the parts of A_P are situated; as mentioned above, the embedding ontological framework of *General Process Theory* operates with a non-transitive part relation and parts are indexed to partition levels.

understand A_P as a process that underlies normative judgments (i.e., as an action that should, must, may etc. be done) and thus generates commitments, and (iii) the ability to initiate a process that is an instance of A_P and to pursue its realization despite some changes in the action environment.[24] At the current state of technology robots lack at least the first two of these abilities; thus they cannot intend to do A_P, and thus they cannot realize an action. But they may realize non-agentive occurrences that simulate an action–currently at least in the sense of approximating, displaying, on mimicking.[25] Thus we can define the class of human-robot interactions as follows.

Definition 8 (Human-robot interaction). A process system Σ is a *human-robot interaction* iff

1. Σ has two process subsystems H and R which satisfy the predicates 'human agent' and 'robotic agent,' respectively,
2. Σ is a complex action A_P and for any part X of A_P it holds that:
   a. X is either an action or the simulation of an action, and
   b. X is an action iff X is part of H, and
   c. X is the simulation of an action iff X is part of R.

This definition can be made more precise in a number of ways but the current version suffices, I trust, to convey the general idea that human-robot interactions are interactions$_3$.

In order to gain more specific descriptions of human-robot interactions we need to indicate the mode in which the robot simulates an action. For this purpose we first need to diversify the simulatory expansion of an interaction$_2$ as follows:

Definition 9 (Simulatory expansion matrix). For any process type A_P with n part processes, the *simulatory expansion matrix of* A_P is a $6^n \times n$ matrix of process types. Each row of this matrix consists of a sequence of n process types C_i so that each C_i is either identical with or a simulation of a process B_i just in case B_i is part of A_P.

To offer a toy example, let us assume that the process type A_P = *welcoming* has the following parts according to the partition P: <B_1 = *approaching target person,* B_2 = *establishing eye contact,* B_3 = *opening one's arms,* B_4 = *uttering in a friendly tone of voice "hello!"*>. A section (!) of the simulatory expansion matrix for A_P = *welcoming* would then look as follows:[26]

[24] For an overview over current accounts see e.g. (Setiya, 2007)

[25] To keep all options open, I will here also assume that it is conceptually possible to entertain the thesis that future robots may be able to imitate or functionally replicate *intending to do X.*

[26] Abbreviations for the modes of simulation are used as names for occurrences that simulate the action in the relevant column.

	approaching	eye-contact	open-arms	utterance
...	...	...	...	...
E_1:	Imit	Dis	Mim	App
E_2:	Dis	Mim	App	Real
E_3:	Mim	Mim	Mim	Mim
E_4:	Dis	Dis	Mim	Dis
E_5:	Dis	App	App	App
...	...	...	...	...

Each row in the simulatory expansion matrix of an action A_P describes one of the many ways to simulate A_P, stating a complex mode of simulation that can be formally represented as an ordered sequence of the modes in which each part of A_P is simulated. Using the expansion matrix of our simple example, Honda's ASIMO robot realizes simulatory expansion E_1, while H. Ishiguro's ROBOVIE robot, which is teleoperated and thus arguably realizes utterances, realizes the simulatory expansion E_2.

Simulatory expansions can be defined for any sort of process, simple, complex, and collective, and both for actions and non-agentive occurrences. For complex processes the array of simulatory expansions quickly becomes very large. In general, for any kind of interaction $\mathfrak{I}$ with m agents and n parts, the simulatory expansion matrix for $\mathfrak{I}$ states all combinatorial possibilities for the realization of all parts of the interaction or for their simulation in any of the six modes, for all agents. For the case of an interaction of two agents the simulatory expansion matrix of this interaction has the dimensions $(6^n \times n) \times (6^n \times n)$.

Let us here focus on two-agent interactions$_2$ (i.e., complex actions) and let 'A_P $[S1{:}B1_1...B1_k,\ S2{:}B2_1...B2_h]$' be shorthand for 'the interaction$_2$ A_P whose parts are actions $B1_1...B1_k$ performed by agent S1 and actions $B2_1...B2_h$ performed by agent S2, as structured by partition P.' The following definition introduces a special subtype of simulatory expansion matrices for interactions$_2$.

Definition 10. The *asymmetric simulatory expansion matrix* of an interaction$_2$ A_P $[S1{:}B1_1...B1_k,\ S2{:}B2_1...B2_h]$ is a $k \times (6^h \times h)$ matrix which correlates each of the actions $B1_1...B1_k$ of agent S1 with any of the possible $6^h - 1$ simulatory expansions of the actions $B2_1...B2_h$ of agent S2.

In other words, the asymmetric simulatory expansion matrices of two-agent interactions$_2$ is a list of all interactions$_3$ that can be generated by simulatory expansions of all (and only) the actions of one of the two agents. Let us call any row in this matrix an *asymmetric simulatory expansion* of the relevant interaction$_2$.

5 Simulated social interactions

On the basis of the definitions presented in the two preceding sections we can reformulate the familiar *topos* of what robots can and can't do by asking 'Which of the asymmetric simulatory expansions of one specific interaction$_2$ is a possible or desirable human-robot interaction?' This way of posing the question allows for differentiated answers.[27] More importantly, the core question of robot sociality can now be stated as follows:

> Q1: For any interaction$_2$ A_P that is a social interaction, which asymmetric simulatory expansions of A_P qualify as social interactions?

At first glance, it might seem that the answer to question (1) is quite straightforward: none. If we take a (again, for simplicity: two-agent) social interaction to be part of "social reality" and postulate, following Searle (Searle, 2009), that this requires "acceptance or recognition" (Searle, 2009, p. 8) by both agents for the "statuses" of the actions they perform, a social interaction implies that its parts are social actions. Thus, one might argue, since no simulation of an action can qualify as a social action, it follows that none of the asymmetric simulatory expansions of a social interaction is a social interaction; *a fortiori*, no human-robot interaction–which are asymmetric simulatory expansions of a social interaction at best–can qualify as a social interaction.

But matters are more complex. There are good reasons to question the philosophical standard model of social interactions, which still follows largely Hegelian lines and calls for joint "recognition" of normative statuses or for various forms of reflective collective intentionality (cf. e.g. Laitinen 2011; Tuomela 2013; Gilbert 2008). In fact, as Cerulo (2009) has pointed out, the intentionalist notions of sociality in philosophy and the social sciences that exclude nonhumans are hardly ever argued for but presumed from the outset. Moreover, any principled exclusion can be challenged on empirical grounds in two ways. On the one hand, social robotics research itself provides strong evidence for reviewing the principled exclusion of nonhumans, since people prefer to use social vocabulary to describe their dealings with robots.[28] On the other hand, and more importantly, the intentionalist standard model of sociality does not sit well with recent research on the conditions of sociality in cognitive science, anthropology, psychology, and linguistics. For in view of this research it is far from clear that human social interactions are in all cases interactions$_2$.

---

[27] For example, surveying the set of asymmetric simulatory expansions of an interaction$_2$ one could investigate whether there are dependencies among expansions ('if this partial action is merely displayed, then that must be mimicked', etc.) or, *vice versa*, one could try to identify certain clusters of asymmetric simulatory expansions ('this sort of simulation succeeds for all interactions$_2$ that are short term/in educational contexts/involve touch etc.'). Answers can be further tailored to technical possibility versus practical feasibility, or practical versus moral desirability.)

[28] Cf. Turkle 2011. See Hakli 2014 for a discussion of possible methodological strategies of accommodating these phenomena for a theory of sociality.

As anthropologists and linguists have argued, sociality requires patterns of interactive practices of turn-taking organization and mutual orientation (cf. e.g. Schegloff 2006; Goodwin 2006), the elements of which are non-agentive occurrences–they are pre-conscious routines acquired by imitation learning. Similarly, cognitive scientists have identified pre-conscious "mechanisms" of social cognition, such as epistemic alignment (Samson et al, 2010), sensory contagion (Blakemore et al, 2005), and "implicit" learning and tracking processes (Frith and Frith, 2012). Finally, from the side of evolutionary biology it has been pointed out that sociality can be explained without presupposing mindreading, i.e., the inferential capacities of a theory of mind, but rather on the basis of practices of conditioning (Mameli, 2001).

All these results can be taken to suggest that all human social interactions are interactions$_3$ consisting of a mixture of actions and non-agentive occurrences.[29] In short, recent empirical research suggests that there is much more involved in human social interaction than what we do or can intend.

Matters are further complicated by the fact that, as noted above, the definition and demarcation of social actions should take into account differences in point of view–differences between the participant's second person point of view and the observer's or third-person point of view.[30] Furthermore, we need to distinguish the *observer's* third person point of view from the point of view of the *omniscient* third person presupposed in philosophical analyses of the *concepts* of social interaction.[31] If we reflect on normative content assuming an omniscient third person point of view, it makes sense to postulate requirements of mutual reciprocal acknowledgements of intentionality, while from the viewpoint of the empirical observer, who does not have epistemic access to the intentional states of another agent, definitions of social interactions can only refer to the internal states of others in terms of manifestations of such states. Arguably, one cannot decide on (Q1) without first clarifying the more foundational methodological question (Q0):

> (Q0) Which of these three viewpoints, two empirical and one normative, has the best claim to delivering the 'right' account of sociality? Or should we consider sociality as a complex phenomenon that involves the facts of all three perspectives?

---

[29] To call such preconscious, non-intended occurrences 'non-agentive' may be problematic–one might agree with my criterion for agency above that if an occurrence is an action then it must be possible to intend it, but deny that the implication also holds in the other direction; all processes that occur as parts of an intended action are agentive, one might say, even though they are not intended. I must bracket this issue here.

[30] Phenomenological analyses of forms of responsiveness long have drawn attention to the importance of this difference, but there is also increasing interest in "second person cognitive science," cf. e.g. Reddy 2008; Schilbach et al 2013.

[31] Elsewhere (see Seibt 2017) I argue that for the purposes of attributing responsibility in a sufficiently differentiated fashion we need to distinguish between: (i) the *second person* point of view of the human interactor with the robot; (ii) the *internal third person* point of view of the roboticist who designs the interaction; (iii) the *external third person* point of view of the observer of a human-robot interaction; and (iv) the *omniscient third person* point of view of the cultural community evaluating the human-robot interaction and its effects on the surrounding context relative to teh community's norms and values.

Whichever stance we take on (Q0), the theoretical tools of the "simulatory expansion matrix" of an action (in a given partition) as defined above should prove helpful for formulating the relevant claims more precisely, I submit. For example, compare the following two characterizations of social interactions. The first one, (A), is the omniscient third person account of social interactions requiring joint knowledge about the mutual acknowledgement of norms of action (Gilbert, 2013); the second one, (B), is an account from a second person point of view in terms of the disposition to respond to the encounter of a conspecific in ways that express certain narrative-practical competences concerning action explanation (Hutto, 2012). If we choose characterization (A) as the 'baseline' of normal social interaction, we could discuss (B) as an account of sociality that focuses only on those entries in the simulatory expansion matrices of an interaction A_P that realize simulation mode 4 (*displaying*) for all partial actions of A_P that are not bodily movements.

Once we have made headway on the question (Q0) and decided where to set the bar of the baseline of a 'real' social interaction and from which viewpoint, we can then address the demarcation question (Q1). To illustrate which form such demarcation questions will take, assume we operate with baseline (A) as just stated (the 'reciprocity account') and consider again the simplified example of an expansion matrix given above. Imagine I instantiate the simulatory expansion E1 of *welcoming*; that is, I happen to walk precisely into your direction, thereby imitating the intentional action of *approaching you*, I happen to look approximately at you, thereby displaying the intentional action of *seeking eye contact*, I happen to stretch my arms, thereby mimicking the intentional action of *opening my arms*, and finally, as I am almost bumping into you, produce purely reflectively a 'hello!', thereby approximating the intentional action of *uttering 'hello!'* – have I been welcoming you? Or again, consider simulatory expansion E6, where all the parts of *welcoming* are realized apart from the last, the meaningful and friendly utterance of a welcome, which is only mimicked–should we count this, from a second or third-person perspective as a social action? E6 is the behavior of a person in the autism spectrum.

In sum, there are sufficient grounds to reject the idea that demarcation questions have straightforward answers. Which robotic occurrences are acceptable contributions to a social interaction depends not only on which forms of simulations we find acceptable, but also on the viewpoint we adopt for the definition of a social interaction in the first place–what appears as a simulation from one viewpoint (and relative to one account) is just the normal perfomance of sociality from another viewpoint (and relative to another account). The best strategy for coming to grips with the demarcation question may be to (i) adopt a maximally demanding baseline (i.e., to adopt the most restrictive realization requirements for social interactions from the omniscient third person point of view), (ii) to state the symmetric and asymmetric simulatory expansions for some typical social interactions, and then (iii) to discuss which regions of the expansion matrix should count as social interactions, drawing on alternative and less restrictive definitions of sociality.

6 Conclusion

The aim of this paper was to offer in outline a conceptual framework for the classification of forms of human-robot sociality. I have tried to differentiate between make believe and fictional interactions ('treating as if'), social 'taking as,' and various forms of the 'as if' of simulation. Since social terms are performative-ascriptive, I argued, they cannot be used in fictionalizing contexts, as social roboticsts are wont to do when describing social robotics applications. Thus the difficult task of conceptualizing what happens in human-robot interactions, the 'soft problem for the ontology of social robotics,' cannot be addressed by a fictionalist 'quick fix.' In order to solve the 'soft problem' we need to develop an ontology of simulated social interaction (SISI), I suggested, and sketched some possible trajectories for how one can approach this task. On the background of an embedding process ontology I have offered five definitions to distinguish varieties of the simulatory 'as if' of human actions and interactions, in terms of five modes of simulations: *functionally replicating, imitating, mimicking, displaying*, and *approximating*.

According to sketch of SISI presented in this chapter, human-robot interactions are instantiations of asymmetric simulatory expansions of social interactions. The simple definitional tools of SISI allow us to develop classifications in terms of which we can formulate the questions of *robo-ontology* with greater precision. But the framework also holds out the promise, I believe, of facilitating the construction of a unified account of sociality.[32] For example, as I highlighted above, relative to standard 'intentionalist' accounts of sociality in philosophy, also human-human interactions instantiate asymmetric simulatory expansions of social interactions, since not all of what we socially do is the result of conscious norm-guided actions. SISI can be used as a conceptual interface between philosophical and empirical research on human sociality–it allows us to investigate which simulatory expansions of social interactions in the standard intentionalist sense, i.e., using notions of sociality championed in philosophy, should be included in the scope of social interactions, given empirical research results on human sociality undertaken on the basis of non-intentionalist notions of sociality. As cognitive scientists, evolutionary biologists, anthropologists, and linguists have argued, the phenomenon of sociality does not hinge on the capacity of norm-guided action but begins much earlier, with joint attention to basic patterns: response rhythms and turn taking, and other elements of our pre-conscious capacity of imitating a praxis. Greater attention to the different modes and ways in which an intentional action can be simulated can facilitate reflections on how much (or how little) mentality we should require for a definition of social action. A fully worked out ontology of simulated social interactions might be

---

[32] The conceptual tools of SISI are particularly basic (since it is grounded in a foundational ontology) yet precise and–due to its simple combinatorial strategies–highly expressive. It is therefore possible to translate into the classificatory framework of SISI other proposals of distinctions in capacities for moral agency (see e.g., Wallach and Allen (2008)), or in asymmetric forms of collective agency (see in particular the interesting contributions to Misselhorn (2015) the details of these embeddings are currently worked out, together with a critical discussion of the classificatory unification one may achieve in this way.

the key to a comprehensive account of sociality where we can map out the relationships between empirical and conceptual analyses of sociality more clearly.

In short, then, the ‘soft problem’ of robo-ontology, i.e., the problem of how to describe human interactions with ‘social’ robots, motivates the development of an ontology of simulated social interactions and at once provides an important heuristics for an integrated account of sociality where empirical results and normative insights are combined.

Using the simple classificatory tools of SISI we receive new analytical perspectives on the ontology of the design process in robotics. As I argue elsewhere (Seibt 2017), in order to generate sufficiently differentiated ascriptions of responsibility we can model design processes as new forms of collective intentionality, with more or less emphasis on engendering human-robot interactions in the so-called ’we-mode’ or the ’I-mode.’

It must be stressed, however, that the suggested approach at best yields an empirically grounded taxonomy of asymmetric simulated social interactions, for humans and robots. Such a taxonomy cannot enlighten us with respect to the ‘hard problem’ of robo-ontology, namely, the question of what, if anything, is lost when a process is perfectly simulated (in the sense of functional replication as defined above). But discussion of taxonomic questions might help us to decide whether and where the hard problem matters at all.

References

Bernstein D, Crowley K (2008) Searching for signs of intelligent life: An investigation of young children’s beliefs about robot intelligence. The Journal of the Learning Sciences 17(2):225–247

Bickhard MH (2009a) Interactivism. Springer

Bickhard MH (2009b) The interactivist model. Synthese 166(3):547–591

Björnsson G (2011) Joint responsibility witwith individual control. In: Moral Responsibility Beyond Free Will and Determinism, Library of Ethics and Applied Philosophy, vol 27, Springer, pp 181–199

Blakemore SJ, Bristow D, Bird G, Frith C, Ward J (2005) Somatosensory activations during the observation of touch and a case of vision–touch synaesthesia. Brain 128(7):1571–1583

Breazeal C (2002) Designing Sociable Robots. MIT Press

Breazeal C (2003) Toward sociable robots. Robotics and Autonomous Systems (42):167–175

Cabibihan JJ, Javed H, Ang Jr M, Aljunied SM (2013) Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. International Journal of Social Robotics 5(4):593–618

Cerulo KA (2009) Nonhumans in social interaction. Annual Review of Sociology 35:531–552

Clark HH (2008) Talking as if. In: Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on, IEEE, pp 393–393

Coeckelbergh M (2010) Health care, capabilities, and ai assistive technologies. Ethical Theory and Moral Practice 13(2):181–190

Coeckelbergh M (2012) Growing moral relations: Critique of moral status ascription. Palgrave Macmillan

Danish Ethical Council (2010) Sociale robotter. Udtalelse fra det etiske råd. http://www.etiskraad.dk/ /media/Etisk-Raad/Etiske-Temaer/Optimering-af-mennesket/Publikationer/Udtalelse-om-sociale-robotter.pdf, accessed Sep. 13, 2016.

Dautenhahn K (2014) Human-robot interaction. In: Soegaard M, Dam RF (eds) The Encyclopedia of Human-Computer Interaction, 2nd edn, The Interaction Design Foundation, Aarhus, Denmark, available online: https://www.interaction-design.org/encyclopedia/human-robot_interaction.html

Enfield NJ, Levinson SC (eds) (2006) Roots of Human Sociality. Berg Publishers

Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robotics and Autonomous Systems 42:143–166

Frith C, Frith U (2012) Mechanisms of metacognition. Annual Review of Ps 63:287–313

Gilbert M (2008) Social convention revisited. Topoi 27(1-2):5–16

Gilbert M (2013) Joint Commitment: How We Make the Social World. Oxford University Press

Gill ML (2004) Aristotle's distinction between change and activity. Axiomathes 14(1):3–22

Goodwin C (2006) Human sociality as mutual orientation in a rich interactive environment: Multimodal utterances and pointing in aphasia. In: Enfield NJ, Levinson SC (eds) Roots of Human Sociality: Culture, Cognition and Interaction, Berg, Oxford and New York, pp 97–125

Gunkel D (2012) The Machine Question. MIT Press, cambridge, MA

Hakli R (2014) Social robots and social interaction. In: Seibt J, Hakli R, Nørskov M (eds) Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014, IOS Press, vol 273, pp 105–115

Hutto DD (2012) Folk psychological narratives: The sociocultural basis of understanding reasons. MIT Press

Kahn PH, Freier NG, Friedman B, Severson RL, Feldman EN (2004a) Social and moral relationships with robotic others? In: Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on, IEEE, pp 545–550

Kahn PH, Friedman B, Perez-Granados DR, Freier NG (2004b) Robotic pets in the lives of preschool children. In: CHI'04 Extended Abstracts on Human Factors in Computing Systems, ACM, pp 1449–1452

Kalbe E, Schlegel M, Sack AT, Nowak DA, Dafotakis M, Bangard C, Brand M, Shamay-Tsoory S, Onur OA, Kessler J (2010) Dissociating cognitive from affective theory of mind: a tms study. Cortex 46(6):769–780

Laitinen A (2011) Recognition, acknowledgement and acceptance. Recognition and social ontology pp 309–348

Larsen S (2016) How to build a robot and make it your friend. PhD thesis, Aarhus University, Denmark

Leyzberg D, Avrunin E, Liu J, Scassellati B (2011) Robots that express emotion elicit better human teaching. In: Proceedings of the 6th International Conference on Human-Robot Interaction, ACM, pp 347–354

Mameli M (2001) Mindreading, mindshaping, and evolution. Biology and Philosophy 16(5):595–626

Misselhorn C (2015) Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation, vol 122. Springer

Petersson B (2013) Co-responsibility and causal involvement. Philosophia 41(3):847–866

Reddy V (2008) How infants know minds. Harvard University Press

Samson D, Apperly IA, Braithwaite JJ, Andrews BJ, Bodley Scott SE (2010) Seeing it their way: evidence for rapid and involuntary computation of what other people see. Journal of Experimental Psychology: Human Perception and Performance 36(5):1255

Scassellati B (2002) Theory of mind for a humanoid robot. Autonomous Robots 12(1):13–24

Schegloff EA (2006) Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In: Enfield NJ, Levinson SC (eds) Roots of Human Sociality: Culture, Cognition and Interaction, Berg, Oxford and New York, pp 70–96

Schilbach L, Timmermans B, Reddy V, Costall A, Bente G, Schlicht T, Vogeley K (2013) Toward a second-person neuroscience. Behavioral and Brain Sciences 36(04):393–414

Searle J (2009) Making the Social World: The Structure of Human Civilization. Oxford University Press

Seibt J (2005) General Processes–A Study in Ontological Category Construction. Habilitationsthesis at the University of Konstanz

Seibt J (2009) Forms of emergent interaction in general process theory. Synthese 166(3):479–512

Seibt J (2014a) Non-transitive parthood, leveled mereology, and the representation of emergent parts of processes. Grazer Philosophische Studien 91:165–191

Seibt J (2014b) Varieties of the ‘as if’: Five ways to simulate an action. In: Seibt J, Hakli R, Nørskov M (eds) Sociable Robots and the Future of Social Relations: Proceedings of Robophilosophy 2014, vol 273, IOS Press, pp 97–105

Seibt J (2017) Robophilosophy–philosophy of, for, and by social robotics. MIT Press, chap An Ontology of Simulated Social Interaction, pp 1–32, forthcoming

Setiya K (2007) Reasons without Rationalism. Princeton University Press, Princeton

Sharkey A, Sharkey N (2012) Granny and the robots: Ethical issues in robot care for the elderly. Ethics and Information Technology 14:27–40

Sparrow L, Sparrow R (2006) In the hands of machines? The future of aged care. Minds and Machines (16):141–161

Sullins JP (2008) Friends by design: A design philosophy for personal robotics technology. In: Kroes P, Vermaas PE, Light A, Moore SA (eds) Philosophy and Design: From Engineering to Architecture, Springer, pp 143–157

Tuomela R (2013) Social Ontology: Collective Intentionality and Group Agents. Oxford University Press

Turkle S (2011) Alone Together. Basic Books, new York

Vallor S (2011) Carebots and caregivers: Sustaining the ethical ideal of care in the 21st century. Philosophy & Technology (24):251–268

Veruggio G (2006) The EURON roboethics roadmap. In: Humanoid Robots, 2006 6th IEEE-RAS International Conference on, IEEE, pp 612–617

Wallach W, Allen C (2008) Moral Machines: Teaching Robots Right from Wrong. Oxford University Press

Walton KL (1990) Mimesis as Make-Believe: On the Foundations of the Representational Arts. Harvard University Press

Walton KL (2005) Metaphor and prop oriented make-believe. In: Kalderon ME (ed) Fictionalism in Metaphysics, Oxford University Press, Oxford, p 65