J. Seibt, M. Damholdt, C. Vestergaard (2020). Integrative Social Robotics, value-driven design and transdisciplinarity. *Interaction Studies* 21:1, pp. 111–144.

Integrative social robotics, value-driven design, and transdisciplinarity

Johanna Seibt, Malene Flensborg Damholdt, and Christina Vestergaard
Aarhus University

"Integrative Social Robotics" (ISR) is a new approach or general method for generating social robotics applications in a responsible and "culturally sustainable" fashion. Currently social robotics is caught in a basic difficulty we call the "triple gridlock of description, evaluation, and regulation". We briefly recapitulate this problem and then present the core ideas of ISR in the form of five principles that should guide the development of applications in social robotics. Characteristic of ISR is to intertwine a mixed method approach (i.e., conducting experimental, quantitative, qualitative, and phenomenological research for the same envisaged application) with conceptual and axiological analysis as required in professional studies in applied ethics; moreover, ISR is value-*driven* and abides by the "Non-Replacement Principle": *Social robots may only do what humans should but cannot do*. We briefly compare ISR to other value-sensitive or value-directed design models, with a view to the task of overcoming the triple gridlock. Finally, working from an advanced classification of pluridiscplinary research, we argue that ISR establishes a research format that can turn social robotics into a new transdiscipline.

Keywords: Integrative Social Robotics, Collingridge dilemma, ontology of asymmetric sociality, responsible robotics, participatory design, value-sensitive design, design for values, care-centered value-sensitive design, technomoral change, transdisciplinarity

1. Introduction

The aim of this paper is to motivate, present, and discuss "Integrative Social Robotics", a new approach to the RDD process (research, design, and development process) for social robotics applications. Introduced in 2015 (Seibt, 2016b, c) the approach was developed to promote a systematic and methodologically

reflected interaction with the new *theoretical and practical* tasks arising with the vision of putting robots into the physical and symbolic spaces of human social interaction. ISR is a generic model for how to implement RDD processes that fulfill recent demands for a "responsible" approach to social robotics, a potentially harmful technology given our current epistemic situation. Social robotics applications that are generated with ISR can be particularly easily justified relative to requirements of ethical-practical rationality of risk minimization and norms of cultural sustainability, since ISR developer teams include *professional* expertise about such normative requirements and produce applications that *anticipate* justificatory tasks.

Even though the ISR is primarily ethically-practically motivated, it carries important implications, we believe, at the theoretical level. In tandem with introducing ISR we present here for discussion three methodological aims that ISR can help to achieve.

First, ISR calls for an extension of the interdisciplinary composition of the two research fields the approach is drawing on and aims to contribute to, viz., "Social Robotics" (SR) and "Human-Robot Interaction Studies" (HRI) focused on the study of social robots (hereafter: sHRI). SR and HRI/sHRI are comparatively young fields; given that their research focus is a moving target (Dautenhahn, 2013), they still are, and perhaps will remain, under development. General assertions about these fields thus should be properly hedged as describing majority practices with exceptions. With this proviso in place, we perceive a growing awareness in the fields of SR and sHRI that the humanities provide important tools for the analysis of human social interactions with robots. This is reflected, for example, in the increased attention to anthropological research in HRI interdisciplinary teams, following the longstanding lead of the work of Selma Šabanović (see e.g. Šabanović 2007, 2010, 2016) and documented in the increased participation of anthropologists in HRI conferences, and partly in the turn to social robotics by anthropologists (see e.g., Robertson, 2017; Hasse, 2015, 2019a, b), as well as in the rise of new research exchange venues devoted to "Humanities research in and on social robotics" such as the *Robophilosophy* conference series (Seibt et al. 2014a, 2016a; Coeckelbergh et al. 2018). We offer here arguments in support of this development of an opening of SR and sHRI towards the full range of humanities disciplines, as relevant in the particular application context.

The full scope of the phenomena arising with human social interactions with robots only comes into view, we submit, once we trace the dynamics of "meaning-making processes" at and across the individual and socio-cultural level of human self- understanding, using the analytical categories of the research disciplines in the humanities that are devised for the purpose, e.g., anthropology, social phenome- nology, social ontology, philosophy of technology, history of ideas, art.

Importantly, the phenomena that can be investigated once the humanities are more fully included into the interdisciplinary scope of SR/sHRI, are relevant both for the *theoretical understanding* of our social, cultural and ethical practices, as well as for *directing these practices* themselves when it comes to developing social robotics applications. In other words, since (changes in) categories of human self-understanding matter centrally for ethical decisions, and since our experiences with social robots challenge these categories, the theoretical and practical motivations of ISR intertwine. Here in particular philosophical research on the normative implications of descriptive vocabulary becomes relevant. Thus, the first methodological aim we wish to highlight in this paper is the aim of producing research results on human social interactions with artificial agents that are *scientifically complete (i.e., include all relevant descriptive dimensions) and practically relevant*. ISR is a suitable 'paradigm' or guiding idea in order to arrive at this aim, since it calls for a well-organized mixed-method approach with wide interdisciplinary scope that includes the methods and analytical categories of all those humanities disciplines that are relevant for a targeted application. In short, ISR helps us to achieve greater interdisciplinary scope in research, design, development, and placement of social robotics applications.

Second, ISR imposes a restriction on which research targets may be pursued. Even though, as we explain below, the restriction imposed is contextualized, this second methodological aim is bound to be controversial. Is it not good science to keep research and regulation as far apart as possible? We argue that it can be advisable to give up on this general rule in certain situations where descriptive and normative-regulatory tasks are in conflict with each other. Such conflicts are familiar from the history of technology and science, e.g., from the early stages of genetic engineering, and typically have been addressed by regulating research for a certain period of time. In line with the movement for "responsible robotics" and the IEEE "Global Initiative on Ethics in Autonomous and Intelligent Systems" ISR calls for a change in our current praxis where, roughly speaking, SR and sHRI research investigates what social robots *can* do, while robo-ethicists and policy-makers deliberate *afterwards* what social robots (*may* or) *should* (not) do, relative to the professional discussion in applied ethics. In contrast, the ISR approach calls upon us to ask do what social robotics applications *can and (may or) should* do. By including, from the very beginning, expertise in philosophical value-theoretic analyses and professional normative ethics into the RDD process, the ISR approach restricts the research focus in social robotics to *culturally sustainable, i.e., value-preserving or value-enhancing* applications.

Third, while there appears to be much leeway in the use of characterizations of forms of pluridisciplinarity, we argue—relative to a particularly demanding sense of "transdisciplinarity" used in recent science studies (Nersessian & New-

stetter, 2013) that SR and sHRI as currently conducted are not yet transdisciplines. We point out that there are indications that SR and sHRI are on their way to becoming one transdiscipline in the given sense, and that ISR plausibly will accel- erate and facilitate this process.

ISR is currently being applied and developed, in a learning-by-doing format, in the context of a five-year research project funded by the Carlsberg Foundation that involves 26 researchers from 11 disciplines (Seibt, 2016c). Even though many details of the ISR approach are not yet available for citation, for the purposes of this special issue we believe it may be productive for the research community to introduce ISR in broader strokes, as researchers in this group currently understand the approach, focusing on principles rather than on the description of concrete procedures that we ultimately aim for. We currently develop several applications using ISR; one of these, for example, uses teleoperated robots as "Fair Proxies" in job applications and conflict mediation in order to reduce gender and ethnic bias in assessment and conflict communications and thereby to enhance social justice and peace (Seibt & Vestergaard 2018, Skewes et al. 2019).

We proceed as follows. In Section 2 we briefly rehearse the motivation of the ISR approach. In Section 3 we present five principles that summarize the basic ideas of ISR for how to organize RDD processes for social robotics applications. In Section 4 we clarify in which ways ISR differs from other proposals for value-oriented research formats for RDD processes for social robotics applications. In Section 5 we consider the question of whether the study of human social interactions with robots should and can become a transdiscipline and offer first observations from research conducted with ISR. We conclude with a consideration of ongoing and future tasks in pursuit of ISR.

2. The motivation for ISR: A triple gridlock

In 1980 David Collingridge observed that throughout the history of technology we often have faced the following dilemma. A technology is being developed whose social impact cannot be predicted before it is widely used; once it is widely used and potential risks have become apparent, however, the technology cannot be easily extracted again, since it is too far entrenched in central socio-economic practices of our societies (Collingridge, 1980). For example, in 1975 the new technology of genetic engineering posed such a dilemma, which the Asilomar Meeting resolved in favor of safety, recommending tight regulations on transgenic products until further research could improve risk assessments. Currently, or so one might argue (Kudina & Verbeek, 2018), the dilemma arises with respect to smart phones applications that could harbor the risk of replacing common skills

(such as planning, writing, or product comparison) with web-based AI using deep learning algorithms. However, the socio-cultural impact of these technologies on our linguistic and cognitive skills or our epistemic autonomy can only be assessed over the time-span of a generation, by which time potential negative effects may be irreparable; *but despite or even because* of these effects societies at that time will be unable or unwilling to give up on web-based decision support.

At first glance it might seem that "social robotics" presents us with just another case of the Collingridge dilemma. (By "social robotics" we refer hereafter broadly to the research, development and deployment of 'social robots', i.e., robots with affordances for human social interactions that are to be used in the physical and symbolic space of human social interactions). Robo-ethicists have used this *topos* – risk of irreparable devaluations of the socio- cultural sphere and degradation of human well-being – on various occasions in their warnings against social robotics (Sparrow, 2016, Sparrow & Sparrow, 2006, Sharkey, 2014, Sharkey & Sharkey, 2012, Sharkey, 2008).

However, social robotics actually presents us with even more profound variety of gridlock for prudential decision making, namely, a *triple gridlock* of description, evaluation, and regulation tasks (D-E-R gridlock) (Seibt, 2016b; Seibt et al. 2018) that deserves special recognition. The Collingridge dilemma arises since (i) we cannot regulate what we have not evaluated but (ii) can no longer regulate once evaluations are possible. By contrast, the D-E-R gridlock begins one step earlier – (i) we cannot regulate what we have not evaluated, but (ii) we cannot evaluate what lacks sufficiently clear and detailed descriptions; and (iii) at a time when suitable descriptions have been developed and evaluations have been worked out, regulation of the technology may no longer be possible (see Figure 1).

Ethical and legal regulations for the domestic and institutional use of social robotics applications should be based on research results that document the short-term and long-term effects of human-robot interactions on individuals and communities, that is, research results that evaluate these effects in (neuro-) psychological, anthropological, sociological, and ethical regards (where the latter are-taken to include not only moral but also existential and aesthetic aspects as these matter for what philosophers call "the good life"). Currently, however, such research is not yet available – we do not yet have an integrated, systematic description, let alone a theory, for the phenomena involved in human social interactions with robots, which range from pre-conscious psychological mechanisms of social cognition to reflected cooperation.[1] Since the "triple gridlock" differs

1. In the following we shall use the terms "description", "descriptive theory", and "theory" in a wider sense that does not contrast with but includes explanatory theories and models; the sense of description we are after here contrasts with the normative dimension, i.e., evaluation relative to norms and values, and normative regulation. We thank an anonymous reviewer for a related comment.

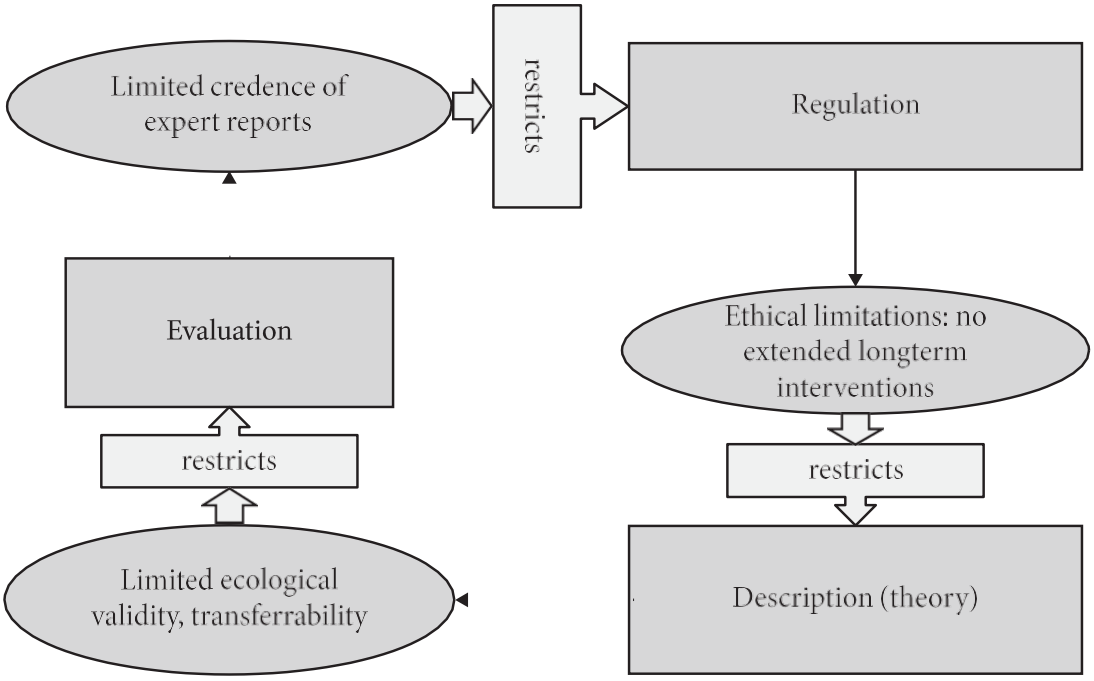

Figure 1. The triple gridlock of description, evaluation, and regulation as it currently arises for "social robotics" (i.e., research, development and deployment of 'social robots', i.e., robots with affordances for human social interactions that are to be used in the physical and symbolic space of human social interactions). Thin arrows represent implications, wide arrows represent inhibiting factors or negative feedback

from the Collingridge dilemma especially due to shortcomings in the descriptive dimension, we will devote most of this section to explaining in greater detail what we perceive as the current *description problem* in social robotics.

2.1 The description problem in social robotics

To begin with a fairly straightforward point, research on social robotics application is typically conducted in the form of (within subject) short-term studies – exceptional "long-term" studies investigate interactions for the duration of 3–6 months, see e.g., (Sabelli et al. 2011; Huttenrauch & Eklundh, 2002; Sung et al. 2009; Leite, 2015). These temporal limitations partly may have merely practical reasons; partly, however, they reflect the regulations of ethics boards and review committees. For instance, long-term studies – and in particular extended long-term studies for more than 6 months, involving children and other vulnerable populations – on emotional attachment to robots, on changes in human-human social interaction styles, or on changes in capacities for empathy towards other humans, would have great significance for the evaluation and regulation of social robotics applications (for results from a 10-day study on emotional attachment

see Dziergwa et al. 2018). But such studies are ethically not permissible – at least not in Denmark – as long as the reversibility of effects is unclear and potential benefits cannot be gauged.

Our second point about descriptive shortcomings in SR and sHRI is less straightforward and we need to begin with a general observation about the-disciplinary composition of SR and sHRI. SR and HRI have different historical origins and at the outset differed in focus and methodology. "Social robotics" was to pursue the construction of robots informed by social cognition research so that the new class of "social" or "sociable" robots could invite humans to react to them using the interactive templates of social interactions (Dautenhahn & Billard, 1999; Duffy et al. 1999; Sekiyama, 1999; Fong et al. 2003; Breazeal, 2002), while HRI defined itself several years later as "a field of study dedicated to understanding, designing, and evaluating robotic systems for use by or with humans", where it was to be understood that "interaction, by definition, requires communication between robots and humans" (Goodrich & Schultz, 2007), 204. While early social robotics sought interdisciplinary contact especially with developmental psychology, early HRI in many ways continued and expanded interdisciplinary research traditions of HCI (Human-Computer Interaction), combining engineering with communication and media studies, sociology, psychology, and design studies. In the meantime, however, there appears to be considerable overlap in the research communities of SR and sHRI, with some of the initial differences remaining. While in other contexts it may be necessary to differentiate more carefully between the two research lines, the following observation holds – in the sense of a characterization of the general interdisciplinary orientations of these fields – for both SR and sHRI, as they have–been conducted so far.

In this sense of a claim about general trends in past and current interdisciplinary composition of research teams, with acknowledged exceptions or even minorities, it seems correct to say that in SR and sHRI it is not yet *standard practice* to include the methods and analytical categories of all humanities disciplines that are relevant for our understanding of the phenomena human social interactions with robots.[2] As we mentioned in the introduction, there is an increasing

2. We cannot provide here a fuller empirical documentation for this assessment of *default* disciplinary compositions of research teams, but to offer a bit of backup we have conducted a simple word search in core publication venues for SR and HRI, ocusing on anthropology and philosophy as candidates for omitted humanities disciplines that focus on the analysis of social practices, experiences, and relations and thus are relevant for most social robotics applications. Of the 1236 papers presented at *ACM/IEEE Conferences on Human-Robot Interaction*, we found, searching keywords, author affiliations, and full texts, the following results for disciplinary distributions: 202 paper relate to "psychology"; 13 papers relate to "sociology"; 4 papers relate to "anthropology" (and are authored by researchers from anthropology); "ethics"

number of contributions bringing the familiar investigative tools of the humanities to the two fields, but also exploring new analytical concepts from philosophical phenomenology (see e.g Parviainen, 2016), from philosophical social ontology (see e.g., Hakli et al. 2017) and even new methods of qualitative research (e.g., Cheon et al. 2018). However, this is a more recent development and SR and sHRI do not yet by default, as a matter of established methodology, operate with interdisciplinary research teams that include researchers who are professionally trained in anthropology, philosophy, and other 'classical' humanities disciplines.

With this general observation in place, we can now begin to introduce our second point concerning the description problem in social robotics as follows. The descriptions of human social interactions with robots are currently problematically incomplete, since the research community in social robotics does not yet, at least not by default, involve researchers that are professionally trained to address questions of, e.g., individual and social ethics, value theory, social ontology (concepts of cooperation, collective intentionality, group actions), the interplay of cultural and personal meaning-making, and, in particular, changes in phenomenological, conceptual, and cultural content. One might object that researchers trained in communication and media science, design studies, and sociology, who often participate in sHRI multidisciplinary research teams, do introduce 'humanities angles'. We do not wish to contest this claim but rather proceed by putting the focus on the aspects that philosophers, anthropologists, historians of ideas, culture and art theorists could contribute and – in view of the evaluation and regulation tasks in social robotics – arguably *should* contribute to the description of human social interactions with robots.

is mentioned in 26 papers (but of these only 3 papers are authored by researcher with professional training in philosophy), "culture" is mentioned in 19 papers (author affiliations vary but are not in anthropology or philosophy); (source: IEEExplore Digital Archive). Of the 101 papers published since 2012 in the *Journal of Human Robot Interaction*, 14 papers relate to "psychology", 1 paper relates to "sociology"; 0 relate to "anthropology"; 2 mention "ethics" (but only 1 is authored by a professionally trained ethicist); 1 paper mentions "cultural" (but the authors are not from the humanities); source: journal website. In the journal *Interaction Stud- ies*, of the 164 papers published (since 2004) that contain the word "robot", 109 papers relate to "psychology", 16 to "sociology", 6 to "anthropology" (authors affiliated to anthropology), 24 mention "ethics" (but only 2 are by authors trained in philosophy), 59 mention "culture" (author affiliations vary; only 1 paper is by an author with affiliation in anthropology); source: ebesco Communication and Mass Media Data Base). In the *International Journal for Social Robotics*, of 509 papers 232 relate to "psychology", 29 to "sociology", 11 relate to "anthropology", 93 mention "ethics" (but only 49 include authors from philosophy), and 157 mention "culture" (of which 6 involve authors from anthropology, 17 authors from sociology, and 8 authors from philososophy); source: journal website on Springer Link.

In order to appreciate that full-blown humanities expertise needs to be included in the analysis of the phenomena of human 'social' interactions with robots, we need to acknowledge that these phenomena have dimensions – e.g., existential, ethical, and cultural dimensions – that are not within the scope of psychological or sociological research. That these dimensions are non-contingent, i.e., that they affect the assessments of psychological and sociological research, is most easily seen for the cultural dimension. By 'cultural dimension' we do not primarily mean here the interpretatory embedding of human interactions with social robots into our communal cultural understanding, i.e., the question of how communities or societies interpret, implement, evaluate, and regulate these new forms of interactions. This aspect of the 'cultural dimension', namely, the analysis of our communal cultural 'meaning-making' of human social interactions with robots, may *prima facie* appear to be separable from the analysis of individual behavioral and verbal responses that can be investigated with the methods and analytical categories of psychology and sociology. By contrast, the processes of *individual meaning-making*, which are also aspects of the cultural dimension of human social interactions with robots, are not in this fashion separable from the dimensions of human experience that are accessed by psychology and sociology; yet they are not accessible with the methods and, in particular, the analytical categories of psychology and sociology.

People interacting with social robots draw on cultural conceptions of robots, humans, technology, and nature (see e.g., (Robertson, 2017; Nomura et al. 2008; Payr, 2018) to make sense of a novel experience. 'Social' robots are items that we do not (yet) understand. They are apparently 'social' agents that are not alive and as such challenge of our current perceptual categories and current cultural conceptions of social agents and social relations (Kahn et al., 2011; Seibt, 2017). People interacting with 'social' robots use different strategies to cognitively accommodate this experience, including the formation of new conceptual constructions (Turkle, 2011; Kahn et al., 2011). As long as the phenomenological and conceptual aspects of human experiences in interactions with social robots have not 'settled' into standardized new categories, the dynamics of individual meaning-making in these novel experiences should be carefully traced (Smedegaard, 2019). In view of neuropsychological research on implicit processes of social cognition (Wykowska et al. 2016; Wiese et al. 2017) and the uncanny valley (Złotowski et al., 2018), it is currently an open question, however, whether human experience of social robots will at all 'settle' in the envisaged fashion. Currently we cannot predict that with increased technological literacy we will develop a common cultural understanding of this new interaction type and individual meaning-making will be reduced to applying established new concepts. As long as people struggle to come to terms in articulating their inter- actions with social robots, as long as we do not yet have a comprehensive account

of which conditions trigger ascriptions of agency, aliveness, moral status etc., we should not, we submit, set aside analytical tools that have been developed for an in-depth analysis of individual meaning-making in situations of cultural change.

In sum, then, our second point with regard to the description problem in social robotics is that *the descriptions of human social interactions with robots are still 'dimensionally incomplete'*. The phenomena of human social interactions involve dimensions of human experience – e.g., ethical, existential, and cultural dimen- sions – that are best explored with the methods and categories of the humanities. In particular, the question of precisely how people make sense of their experiences in interacting with a social robot is, currently at least, a constitutive phase of the phenomenon that sHRI aims to investigate, and the tools needed to study such meaning making processes – e.g., conceptual analysis, phenomenological analy- sis, narrative analysis (Cheon & Su, 2018) – are at the core of humanities research. Our third point concerning the description problem concerns not the description of human experience but the descriptions of robotic capacities. While the descriptions of the human side in human-robot interactions are dimensionally incomplete, the descriptions of the robot side are outright misleading. Engineers are often found to describe what social robots 'do' without clear distinctions between functional and so-called 'intentionalist' vocabulary (Seibt, 2014b). Most of the characterizations of the capacities of social robots, if taken literally, falsely impute to the robot the sophisticated mental capacities that we associate with human intentionality or, in traditional philosophical terminology, with subjectivity – e.g., the capacity to represent, recognize, communicate, choose, move and act in accordance with agentive volitions, or apply normative judgement (Dennett, 1989). Social robots are said to "act as" care-givers or companions, but they are also directly described using intention-implying action verbs such as "understand", "disagree," "greet", "perform rehabilitation therapy", "work in retire- ment homes", "give directions", "recognize," "respond," "cooperate," "communi- cate", etc. In fact, the inadvertent use of intentionalist vocabulary even occurs in classifications of social robots (Fong et al. 2003; for a more detailed discussion see Seibt, 2014b, 2017; Zawieska & Stańczyk, 2015 also note that roboticists often use "anthropomorphic language" but without supplying further analysis of this term).

More precisely, in the literature there are currently four main strategies for the description of a robot's agentive capacities (in interactions with humans):

1. *Uncommented metaphorical extension:* Robotic capacities are described with verbs for intentional actions – robots are said to 'answer', 'greet', 'guide', 'advise', 'remind', etc.
2. *Design for deception:* As in (1) robotic capacities are described with verbs for intentional actions, but explicitly from the point of view of the person interacting with the robot: "We refer to this class of autonomous robots as social robots, i.e., those that

people apply a social model to in order to interact with and to understand. This definition is based on the human observer's perspective" (Breazeal, 2003:168). The person who interacts with the robot is supposed to *misunderstand* the robotic motions and acoustic productions as actions of answering, greeting, guiding, reminding, advising etc. The deceptive display of intentional actions is the expressed aim of the design of the robot: "Believability is the goal. Realism is not necessary" (Breazeal, 2002:52).

3. *Behaviorist reduction*
   Robotic capacities are characterized with intentionalist vocabulary on the explicit assumption that the common conceptual norms do not apply for the robot. Here authors expressly are committed to Behaviorism in philosophy of mind and hold that in their scientific use these verbs do not imply that the agent has intentions and beliefs but merely describe parts of behavioral patterns (in a sense of 'behavior' that applies to lower animals, i.e., agents without intentions), see e.g., (Arkin & Arkin, 1998).
4. *Fictionalist interpretation*
   Robotic capacities are characterized with intentionalist vocabulary as part of a description of a make-belief scenario. The person interacting with the robot is described as having accepted the conventions of a fictional context and understand robotic motions and acoustic productions *as if* these were answers, greetings, beckoning, reminders, etc.; see e.g. (op den Akker & Bruijnes, 2012).

Each one of these strategies for describing robotic capacities is highly problematic.

Ad (1): Uncommented metaphorical extensions can only be admitted in scientific contexts when the domain clearly implies non-literal use – atoms can be said to 'donate' electrons since they are clearly not the kind of entity that could have or simulate human intentions or other human mental states. But in scientific contexts such as AI and social robotics that pertain to the analysis, modelling, or simulation of human cognition such uncommented metaphorical extensions violate common norms of responsible use of scientific terminology.

Ad (4): While strategy (1) makes false statements about the robot, the *fictionalist interpretation* falsely describes the interactions between robot and human – neither from the first person perspective nor from the third person perspective can human robot interactions be understood as fictional social interactions. The phenomenological content in the first person perspective – i.e., the 'what it is like' – of an interaction under the explicit mutual convention of fictionality *contains* that aspect of fictionality—I experience the death of Romeo (in Shakespeare's play) *as* a fictional death. But this phenomenological element is not confirmed by empirical (qualitative) research on how social robots are experienced where, to the contrary, the genuine or authentic involvement of human

interactors is emphasized (see e.g., Weiss et al. 2009; Kahn et al. 2010; Turkle, 2011; Darling et al. 2015).

Ad (2) and (3): The second and third strategy for characterizing robotic capacities do not involve falsehoods. Strategy (3), however, offers only very coarse-grained descriptions. If, for example, a greeting is characterized merely in terms of those behavioral patterns that are shared in many individual greetings performed by humans across contexts, it is questionable whether any communal pattern will be found rather than a large number of patterns related only by so-called family resemblance relations. The same problem arises for strategy (2), design for deception. Here the attributions to robots are interpreted as what philosophers call *response-dependent predicates* – in essence, the sentence 'X is F' should be read as 'persons have the tendency to respond to X with the predicate 'F' (or: tends to be appeared to F-ly by X'). For example, 'the robot answers' should be interpreted as 'the robot behaves in ways that people tend to interpret as an answer', just like 'this surface is red' is to be read as 'this surface tends to elicit in people the response 'it is red' or an experience of *red*'. However, response-dependent descriptions of robotic capacities are very generic – there are many ways in which a robot can elicit in a person the interpretational model of a social interaction. Thus, a description of robotic capacities in terms of response-dependent predicates would only be useful if the degree of variation in individual responses were sufficiently small. But currently it is still unclear, for example, to what extent personality traits may affect such responses (see e.g. Damholdt et al., 2015). Since the stability of responses across time, personalities, and cultures is still an open task for HRI research, it should not be presupposed by the terminological framework that is used to conduct such research.

In sum, then, the description problem in social robotics and HRI research consists in the fact that (i) we currently lack long-term studies that would allow us to describe effects that are most relevant for identifying potential benefits and harms, and that would allow us clearly to identify what is, and what is not, due to the novelty effect in our interacting with social robots (Smedegaard, 2019); (ii) the descriptive and interdisciplinary scope in SR and sHRI is kept too narrow, leaving out the agent's meaning-making processes, i.e., the agent's self-understanding, which renders the descriptions dimensionally incomplete; and (iii) that none of the extant common strategies for the description of robotic capacities yield descriptions that are both true and sufficiently precise.

2.2 The evaluation and regulation problem; closing the negative feedback loop

Without true, precise, and relevantly complete descriptions of robotic capacities, however, we cannot evaluate the effects of human-robot interactions that truly

matter for the regulation of social robotics applications. In order to prepare such evaluations, we would need to have answers to questions like the following five, of which the first four are based on the methods and categories of humanities research. (Note the following abbreviations: for the term "interaction" read "interactions that occur ubiquitously and over extended long-term period (>6 months, preferably several years)" and for the term "social robots" read "social robots (differentiated for different types)").

1. How do interactions with social robots affect individual and social human well-being, including the ethical and existential dimensions of such well-being?
2. How will interactions with social robots affect human-human interactions *de facto* (e.g., will we come to interact with each other in more schematic ways, importing the social signaling used in human-robot interaction, will we have less empathy or will human empathy no longer rely on direct perception (Gallagher & Varga, 2014)?)?
3. How will interactions with social robots affect our cultural conceptions of a social interaction, e.g., will the notion of sincerity become obsolete?
4. How will interaction with social robots affect our value system, e.g., will we tie human dignity more closely to autonomy, which social robotics can enhance, instead of keeping it linked to interpersonal recognition?
5. How will our interactions with social robots affect our emotional, cognitive, and physical capacities, differentiated for different developmental phases of a human life?

As long as we do not have research-based answers to these and similar questions, we cannot draft research-based evaluations of the potential harms and benefits of social robotics applications. And as long as policy-makers and legislators are not furnished with research-based evaluations along these lines, they have little reason to issue any regulations beyond those pertaining to safety concerns.

Given increasing market pressures, the international community of researchers has taken efforts to get involved in lobbying, research communication, and training, as witnessed by the "Foundation Responsible Robotics", or to attempt "regulation from within", as witnessed by the IEEE "Global Initiative on Ethics in Autonomous and Intelligent Systems". While these initiatives are extremely important and valuable, they cannot, as such, remove the negative feedback from current regulations to description.

In sum, while interdisciplinary extensions of SR and sHRI research can improve on the dimensional completeness of description (theories) or human social interactions with robots, the temporal restrictions on empirical research must remain in place. Since there are complex feedbacks between the short-term

dynamics of individual meaning-making and the long-term dynamics of social adaptation and integration into social practices and conceptions, the temporal restrictions hamper both description and evaluation. We arrive at an exacerbated version of the Collingridge Dilemma, a triple gridlock: both description and evaluation are fully possible only at a time when regulation is impossible, i.e., at time when the technology is socially entrenched and no longer extractable.

3. Five principles of ISR

How could we resolve the current description-evaluation-regulation gridlock (D-E-R gridlock) in social robotics? We cannot regulate yet, at least not on rational grounds, but it seems we must, especially to protect vulnerable sections of society, i.e., in regards of robotics applications in childcare and education (see e.g., Turkle, 2011) and healthcare (see e.g. Sharkey, 2014). In response to the risks of unregulated robotics several efforts have been made in the research community to create practices for "responsible robotics" and standards for "ethically aligned design".[3] To date, however, ISR is the only approach that has been developed (i) in response to a more detailed analysis of our current conditions for rational, value-oriented decision-making about social robotics applications, and (ii) as a set of principles that are geared to addressing theoretical goals of descriptive integration and practical goals in tandem. We will further explain and substantiate this announcement in the remainder of this paper.

ISR calls for a 'Gestalt switch' in the research, design, and development (RDD) paradigm of social robotics. To restate, by "social robotics" we do not refer to SR or sHRI but more broadly to research, design, development (including deployment) of 'social robots', i.e., robots with affordances for human social interactions that are to be used in the physical and symbolic space of human social interactions. In the previous section we argued that humanities researchers are not yet generally included in the interdisciplinary scope of SR and sHRI. This holds in particular for humanities researchers with professional training in ethics and value-theory. That is, in the current paradigm for the development of social robotics applications the interdisciplinary scope typically comprises engineering, psychology, communication and media science, design studies, and occasionally sociology, but not, by default, researchers trained to treat ethical questions based on an in-depth understanding of the normative-conceptual foundations of the community and embedding society at issue. Such researchers currently sit in governmental ethics councils and advisory boards for policy and legislation on public applications of social robotics. Thus, in the current paradigm, the professional

3. See www.responsiblerobotics.org and the website of the IEEE "Global Initiative for Ethics in Digital and Automated Systems."

expertise on normative issues comes too late in the process, — as the evaluation of a ready-made product (see. Figure 2a).

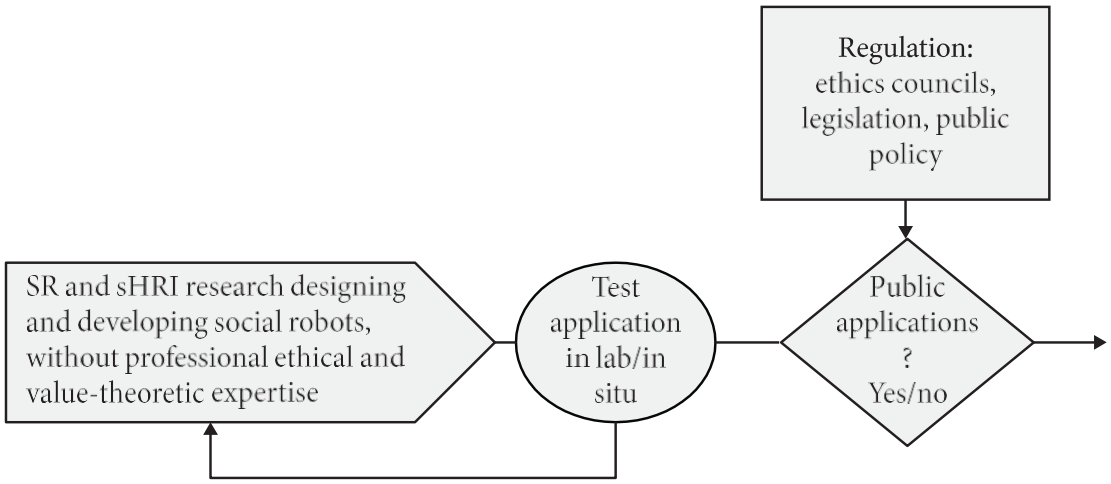

Figure 2a. Current research and decision framework concerning social robotics applications

By contrast, the approach of ISR prescribes wide interdisciplinary scope and interdisciplinary collaboration throughout the RDD process, with value-theoretical and ethical considerations driving the entire process. This changes the decision structure in ways that protect developers against deselection of applications on ethical grounds but also make the dynamic adjustments of normative frameworks possible, in the sense of a co-shaping of technology applications and ethical values (Kudina & Verbeek, 2018; Coeckelbergh, 2012); (see Figure 2b).

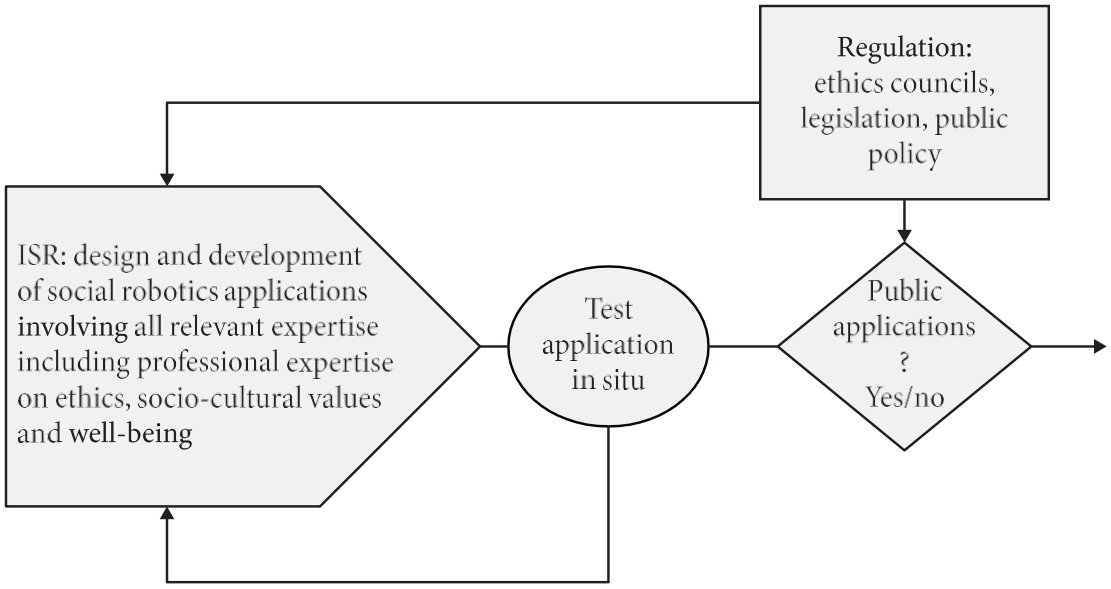

Figure 2b. "Integrative Social Robotics" – a method paradigm for culturally sustainable human-robot interactions based on wide-scope interdisciplinary integration of robotics research

We offer these simple graphics in an effort to convey the Gestalt switch from ethical expertise applied *after* social robotics RDD processes to ethical research applied *within* social robotics RDD processes. Other core ideas of ISR are better supplied in explicit formulation, which we state in terms of five guiding principles, with additional commentary. As mentioned in the introduction, the approach is currently worked out learning-by-doing in the context of a large research project; a more detailed version with flow-diagram and case studies is in preparation.

The first principle, the so-called "Process Principle", prescribes an ontological reorientation away from objects or things towards a focus on processes or, more precisely, interactions. (To formulate these principles compactly, let 'social* interactions' refer to interactions that a future theory of asymmetrical sociality – i.e., where one interaction partner does not have the capacities of normal human social agents – *may* classify as social interactions on the basis of a future theory of asymmetrical sociality (Seibt, 2017); moreover, let 'social' ('social*') interactions here refer broadly to social, cultural, or socio-cultural interactions.)

(P1) *The Process Principle*: The product of a RDD process in social robotics are not objects (robots) but social* interactions.

The main reason for why RDD processes should be guided by the *Process Principle* is that so-called 'social robots' simply *are not* things or instruments that people use – robots are treated as social agents, albeit most peculiar ones. Ontologically speaking, it is preferable, for reasons unrelated to social robotics, to adopt a process-ontology (i.e., an ontology that countenances not object and persons as basic categories but only processes (interactivities, interactions, activities, developments etc.). In a theoretical framework for social robotics that uses a process ontology both robots and people are categorized as process systems and their properties are "affordances for interactions", among which are affordances for social* interactions.

There are many theoretical advantages in adopting a process-ontology in the philosophy of mind and social ontology (see for example Bickhard, 2009; Campbell, 2015; Seibt, 2016d; Bickhard, 2017; Seibt, 2017), but we propose the *Process Principle* in particular also for its heuristic advantages. Once we recategorize social robots as systems of affordances for social* interactions, the interactive dimensions of all design aspects (physical appearance and materiality, kinematic design, functionalities) come more clearly into view. Once we understand the products of social robotics engineering not as things but as affordances for social* interactions, we can approach the evaluation and regulation task with much greater precision. For example, the question: 'Should we regulate the use of service robots of type X in public institutions of type Y?' becomes the much more specific and manageable question: 'Should we regulate the occurrence of the interaction of quasi-assisting of type X (or: quasi-teaching, quasi-counseling,

quasi-advising, etc.) in public institutions of type Y?' The ontological redirection from things to social* interactions facilitates regulatory tasks in two ways. On the one hand, discussions about whether or not to introduce a thing into a human interaction context immediately face the problem of the "functional multi-stability of artifacts" (D. Ihde) – things can be used and abused in so many unforeseeable ways. By contrast, the affordances of social interactions can be more easily calculated since the interaction type can be precisely described (Seibt, 2017, 2020). On the other hand, and more importantly, the focus on interactions immediately opens up a search space for a *how*. *How* can we generate the desired interactions in the given context? While introducing a robot effects a *determinate* change in the physical affordances of an interaction context, envisaging an interaction is tantamount to raising questions about alternative designs (e.g., 'how should the quasi-guiding be realized, by a floating or a rolling device, non-verbally or verbally?' etc.), promising more reflected and adequate changes in the physical affordance space.

If we admit that social robotics is concerned with producing new (affordances for) social* interactions, i.e., social, cultural, or socio-cultural interactions, it is also easier to determine which disciplines should be involved to clarify the relevant affordance for any particular application. Social robotics aims to intervene in the world of human socio-cultural interactions, the most complex ontological domain we are familiar with. Thus, we need to widen the scope of possible disciplinary input in this area of technology production and bring as much as expertise to bear on the envisaged interventions as is justifiably relevant.

Such an opening of the interdisciplinary scope is a matter of general methodological soundness in scientific research *and* a matter of practical rationality – relevant knowledge must not be left out. Therefore, as argued in the previous section, we need to approach social robotics as an interdisciplinary research area which centrally involves the very disciplines that specialize in knowledge production about the socio-cultural domain, namely, not only the social sciences but also the humanities. The second principle of ISR formulates this requirement:

(P2) *The Quality Principle*: The RDD process must involve, from the very beginning and throughout the entire process, expertise of all disciplines that are relevant for the description (theory) and evaluation of the social* interaction(s) involved in the envisaged application.

As formulated (P2) is a general principle of research quality and as such of course pursued in SR and sHRI research, albeit so far without including humanities research by default. More recent methodological reflections of SR and HRI research on the standardization of test methods and evaluation frameworks (Bethel & Murphy, 2010; Jost et al. 2020) provide important input for the for-

mulation of the ISR research algorithm that realizes (P2), as does the USUS evaluation framework (Weiss et al. 2011). The USUS evaluation framework also recommends the systematic methodological diversification envisaged in (P2), but, unlike ISR, it prescribes research methods for fixed evaluation factors. By contrast, ISR integrates the selection of research methods and evaluation factors into the RDD process for a particular application. Which types of expertise are 'relevant' in a given projected application context – for instance, the use of a guiding robot at an airport *versus* the use of a guiding robot in a nursing home – is to be determined by a heuristics using various methods: e.g., conceptual, value-theoretic, phenomenological analysis, ethnographic field, participant observation, focus groups, interviews work at the context and intake workshops with all stakeholders, grouped and in full assembly. Thus the interdisciplinary composition of research teams for any specific application comprises certain default disciplines (e.g., robotics, (neuro)psychology, sociology, anthropology, philosophy, linguistics, design, art) while others (e.g., geriatric neurology, education, nursing, sociology, museum pedagogy, history etc.) are context-dependent extensions of the interdisciplinary scope.

(P2) promotes in particular the standardized inclusion of the humanities into the RDD process. This serves the overall goal of loosening the triple D-E-R gridlock by working on the D-E gridlock and the E-R gridlock at the same time, as follows.

First, we can approach the description problem by using humanities research in (social) ontology to set up a sufficiently precise terminological framework for the description of human-robot interactions (for extant attempts see e.g., Seibt, 2014b, Fiebich et al. 2015; Misselhorn 2015; Fiebich 2017; Hakli et al. 2017; Seibt, 2017, 2018, 2020). On the basis of such a unified descriptive framework we can integrate the research results from the human and the social sciences as well as of the human- ities to devise a comprehensive theory of asymmetrical social interactions and explore how and where to set the boundaries of the domain of sociality. To re- emphasize, such attempts at descriptive (if not yet theoretical) integration will greatly benefit from recent and current proposals for unified test method and evaluation frameworks as discussed in sHRI research.

Second, if humanities research on norms, values, and cultural practices etc. is involved *throughout* the RDD process in social robotics, we can address the evaluation-regulation gridlock in a processual fashion, using short term regulatory feedback based on continuous evaluation instead of the prevalent 'after-the-fact' attempts at ethical-cultural evaluation and regulation. SR and sHRI research does make reference to ethics – see footnote 2 – and important research has been undertaken to bring ethical issues into view, especially in connection with the issue of empathy (Bartneck et al. 2009; Darling et al. 2015; Złotowski et al., 2016) or robot nudging (Borenstein & Arkin, 2017). But non-philosophers may under-

estimate the difference between, on the one hand, 'folk-ethical' considerations that can be brought into the research scope by user-centered designs and, on the other hand, professional engagements of ethical and value-theoretic concerns. While the 'folk-ethical' intuitions and observations of users (and researchers!) play a central role in the ISR approach and are interwoven, by regular feedback with stakeholders, into the RDD process, the particular asset of the ISR approach is that professional ethicists are involved from the very beginning of the RDD process until and including the phase of the placement-cum-entrenchment of the application in the concrete social location. Applications developed with the ISR approach thus will be professionally prepped for evaluations by governmental ethics councils and, just as importantly, inform such councils.

The philosophical foundations for this processual approach to the E-R gridlock, to norms and values, can be found in the early pragmatist tradition, especially in the work of J. Dewey, but even if one does not share the (anti-)metaphysical convictions of the early pragmatists, it should be clear, we submit, that method pluralism and the opening of the interdisciplinary scope in the investigation of human-robot interaction can only improve the quality of research in this area and thus the reliability of the evaluations that are to inform regulations.

The following two principles of ISR formulate substantive assertions about the nature of social and socio-cultural interactions.

(P3) *Ontological Complexity Principle*: Any social interaction I is a composite of (at least) three social interactions $<I_1, I_2, I_3>$, namely, the interaction as understood from (at least) two interacting agents and an external observer. In addition, the descriptions of each interagent (i.e. I_1 and I_2) contain a first person, second person, and third person view of the agent's contribution, while the external observer only takes a third person perspective. The RDD process in social robotics must envisage, discuss, and develop social* interactions on the basis of this ontological conception, taking them to be perspectival composites.

We speak here of ontological 'complexity' primarily in order to convey the philosophical point that the way in which social interactions exist should not be articulated using the model of being that we use to assert the existence of things – while things may be taken to exist 'simpliciter' as being of this or that kind, social interactions have an internal perspectival complexity (Seibt 2018).[4] Social interactions such as a greeting involve what interaction partners do and experience, but each interaction partner experiences the interaction taking her or his own point

4. Whether such ontological complexity also should be taken to generate 'complexity' in the technical sense used in complex systems theory we wish to leave open at this point; we thank an anonymous reviewer for this observation.

of view ('what do I need to do in order to fulfill the action norm?'), the other person's point of view ('is what I am doing perceived as fulfilling the action norm?'), as well as the point of view of an imaginary third person ('are what I and the other one are doing in agreement with the action norm as externally perceived?'). Social interactions – understood as normative interpersonal practice – can only be established with a minimum of three people involved, i.e., the third-person perspective must exist *de facto*, in addition to be imaginatively rehearsed. Social interactions are thus intrinsically complex entities involving descriptions from (at least) seven perspectives. In designing and developing a robot, engineers need to conceive of the robot's social agency from all three perspectives of an interagent – they need to consider the robot's simulated action from the 'first-person' perspective, how it is perceived from the second-person point of view of the human interagent, and how an external observer (society) perceives the interaction from a third-person perspective.

Such perspective-taking becomes complicated once normative issues enter (e.g., 'what does it look like to a bystander if the robot lifts the patient like this?'). In the ISR paradigm, due to (P2), engineers are not alone in this complicated task but collaborate with those that have the required expertise in the normative systems that our socio-cultural practices implement. In combination with (P2), (P3) thus can facilitate the processual dissolution of the E-R gridlock, i.e., the selection of culturally sustainable human-robot interactions, since evaluations from a third-person perspective of evaluation by the normative community at large are already anticipated in the development of the application.

The fourth principle accounts for the very fine-grained identity conditions of social interactions.

(P4) *Context Principle*: The identity of a social interaction is relative to its (spatial, temporal, institutional etc.) context. The RDD process in social robotics must be conducted with continuous and comprehensive short-term regulatory feedback loops (participatory design) so that the new social* interaction is integrated with all relevant contextual factors.

Since social and socio-cultural significances depend on the context of interpretation, it is not possible in social robotics to operate with general projections of the workings of the (physical, kinematic, and functional) design of a social robot. This is nothing new – it is a particular case of a well-known general problem, the problem of the context-dependency as it arises for many technological innovations. Since the 1970s, with first proposals originating in Scandinavia, the problem of context-dependency has given rise to dynamic design formats called 'cooperative design', 'co-design', 'user-driven design', or 'participatory design'.

So at first glance (P4) seems to assimilate ISR to these formats. There are several characteristics of ISR, however, that set it apart, with the most obvious one

to appear with principle (P5) below. But also (P4) formulates crucial differences. First, unlike the mentioned approaches, in ISR the degree, scope, and temporal extent of the 'participation of all stakeholders' is not left open. In ISR stakeholders are involved throughout the RDD process, at specified intervals, not only at the beginning. Second, provisions are made that the pool of stakeholders may vary with time – relevant (especially: temporal) factors of a context may appear only in the course of the development of an application. Third, developer teams that use ISR not only shall create applications that are tailored to the specific conditions of particular interaction contexts (schools, nursing homes, public places etc.) but they shall also accompany social robotics applications into the 'placement' phase, i.e., *beyond* the introduction period to the first phase of the 'new normal', i.e., to a time when new normative practices for the handling of exceptions and malfunctions are in place (Nickelsen, 2018).

The fifth principle of ISR is by far the most important one, not only in practical regards but also in terms of the farther-reaching value-theoretic commitments incorporated.

(P5) *Values First Principle*: Target applications of social robotics must comply with a (deliberatively established specification of) the *Non-Replacement Maxim*: social robots may only do what humans should but cannot do. (More precisely: robots may only afford social* interactions that humans should do, relative to value V, but cannot do, relative to constraint C). Throughout all stages the RDD process in social robotics should be governed by axiological analysis and evaluation.

To begin with the practical rationale for (P5), given the D-E-R gridlock, given that we are currently operating in a compounded situation of epistemic uncertainty (where we cannot describe the interactions we are creating nor know about their effects), the most rational procedure is to select only those application projects that fulfill the non-replacement maxim relative to a value V that in our current axiological system (system of ethical, aesthetic, cognitive, practical etc. values) ranks highest in the given context. Currently ethical (moral) values still rank highest in our practical reasoning, and the non-replacement maxim operates as a filter that forces innovative energies to deprioritize economic considerations of brute force productivity gain. Instead of asking 'How we can actualize the "automation potential" of our current professions?' (McKinsey Global Institute, Manyika, 2017), i.e., instead of asking 'How can we replace human social interactions with human-robot social* interactions?', the non-replacement maxim forces developers to ask 'Is there a high-ranking (moral) value that, in the given context and given constraints C, *cannot* be realized by human-human interaction but *can* be realized by means of a human-robot social* interaction?'.

It is important to note, however, that non-replacement maxim is formulated with parameters V and C which need to be determined in a process of ethical deliberation within the RDD process, by all members of the research team and including all stakeholders. The constraints C may relate to material aspects (e.g., humans cannot be exposed to radioactive radiation or cannot run on solar energy) or to more subtle features of kinematics and appearance (e.g., humans typically cannot repeat actions precisely and indefinitely, or fail to have gender). The fact that the non-replacement maxim refers to constraints by means of a variable ('C') renders the maxim as open as it needs to be in order to engender joint reflections, by all stakeholders, about the necessity of the envisaged replacement of human social interaction.

While (P5) prescribes that the RDD processes of ISR are value-driven, it should be noted that the relationship between values and their realization in social reality is not conceived on the model of the instantiation of eternal abstract entities, so-called 'Platonic ideals' – instead, values here are taken to be continuously 'in the making', to exist only in their realizations in interactions and in the heuristic deliberations that guide towards these realizations. We also would like to emphasize that the axiological or value-theoretic deliberation that is to inform RDD processes does not exclude considerations of utility – it just means that utility and economic gain is carefully evaluated relative to ethical, existential, aesthetic values and factors of well-being broadly conceived. Here again ISR extends a trajectory that can already be found in multifactorial evaluation frameworks proposed within HRI research (Weiss et al., 2011).

This process-philosophical account of values, which rearticulates the position of the early pragmatists (especially J. Dewey), points to the larger philosophical trajectories that buttress the construction of ISR. As such, this we wish to stress, ISR is practically motivated in terms of the D-E-R gridlock, but it is an approach that at the same time also aspires to be defensible in wider philosophical perspectives. Alas, we need to leave the elucidation of the philosophical embedding of ISR for another occasion. But just to mention one of these trajectories, it appears, following Heidegger's line of thought in the *Question of Technology*, that we have reached that 'moment' or period in human cultural history where for ontological reasons the challenge that technology poses to humanity reaches an extremal phase. The method paradigm of ISR is grounded in a process-geared metaphilosophy, a 'post-post humanism' (neo-humanism) that responds to the 'grand narrative criticisms' of 20th century analytical and continental philosophy (constructivism, poststructuralism, postmodernist, posthumanism etc.) with the pragmatist dodge of turning contested entities or relations into well-defined heuristic procedures: truth lies in truth-producing procedures; our humanity lies in the way in which we respond to situations that challenge it, etc. Such a process-philosophical 'neo-humanism' rearticulates a non-relativist but ineradicably dia logical conception of our world.

4. Value-driven versus value-sensitive design

As sketched in the previous section, ISR addresses the triple gridlock in a double movement: (i) the D-E gridlock is slowly dissolved by working towards descriptions of human-robot interactions that are relevantly complete and precise; (ii) the E-R gridlock is bypassed by selecting only applications that are (in the given context) uncontroversially valuable. In fact, here (P2) through (P5) play together to ensure that we focus on what can justifiably be called 'uncontroversially' valuable applications. While (P2) ensures the inclusion of relevant expertise on well-being and values throughout the RDD process, (P3) warrants that the stakeholders in ISR include not only the persons and institutions of the immediate application context but also society at large; (P4) establishes that the RDD process is guided by normative considerations that result from continuous mediations of normative demands of the stakeholders relative to the context; and (P5) specifies a rather restrictive filter for the selection of applications, so that the discussion of these within the project group ensures that an application is not driven by economic utility considerations but by a context-adequate understanding of values.

To be sure, the attention to values in technology design is as such not new. At least since the late 1990s members of the research community in technology design have been promoting the inclusion of value considerations at the beginning of technology design. As a general design strategy for technology development, proposals for "value-sensitive design" (Friedman et al. 1997; Friedman & Bainbridge, 2004) and "design for values" (Van den Hoven, 2013; Van den Hoven, 2005) have been gaining momentum during the last two decades. There are, however, some important differences between these approaches and ISR, as well as the "care centered value-sensitive design" approach by A. van Wynsberghe (Van Wynsberghe, 2013), and it will be instructive to undertake a more extensive comparison between ISR and these recommendations for how to include value considerations into RDD processes. Here we mention a few tentative contrastive points; a more extensive discussion is in preparation.

First, "value-sensitive design" (VSD) as originally introduced and "design for values" (DV) proceed from the assumption that the primary motivation for the introduction of new technology should be utility – values come into view as factors that can be considered next to others and may be preferred since they enhance or stabilize the instrumental value of the technology (Friedman et al., 1997), p.15) (van de Poel, 2015), p.116. By contrast, ISR is strictly value-*driven* in the sense that, given our current situation of descriptive and evaluative uncertainty, the realization of utilities would be sacrificed to the realization of values. "Care centered value-sensitive design" (CCVSD), where the principles of care ethics are used to identify ethical and moral values as relevant in the application

context, shares with ISR this switch in priorities from utilities to values and the switch from instrumental to ethical engineering.

Second, in DV values enter early in the design process but it does not appear to be an explicit requirement that they are continuously re-engaged throughout the RDD process. By contrast, in CCVSD at least one ethical evaluation loop during the RDD process is envisaged (Van Wynsberghe, 2013:315); VSD explicitly envisages "conceptual [i.e., including ethical], empirical, and technical investigations, applied iteratively and integratively" (Friedman et al., 1997:13); and in ISR ethical (phenomenological, axiological), psychological, and anthropological field studies accompany the entire process, up until the first of the "new normal" engendered by the application (see P4).

Third, as has been observed (Manders-Huits, 2011), VSD and DV are not grounded in a philosophical theory of values – here engineers are invited to study eclectically various philosophical definitions of values and to explore intuitively how they can be realized in a given application context; it remains unclear what values are,[5] how they relate to actions and interactions, and, in particular, how we are to decide on the ranking of values. CCVSD improves on this situation by working from substantive ethical commitments to care ethics and an understanding of care practices as "responses to needs", where needs are "broadly construed" and to be contextually determined (Van Wynsberghe, 2013, p.316). By contrast, RDD processes in ISR draw on *all* relevant types of expertise provided by researchers that are fully trained in the respective discipline, i.e., also research on the ontology of interactions and their relationships to values, as well as axiological research on the context-relative rankings of values of all kinds (moral, ethical, aesthetic, cognitive etc.). In addition, ISR is committed to a specific pragmatist process-ontology of values that takes values to be realized in (inter-) actions of dialogical search (which may have been Plato's original conception, see Wieland, 1999). The recent proposal of "ethics from within", exploring "technomoral change" (Kudina & Verbeek, 2018) seems to be in alignment with this interactional, dialogical conception of values.

Fourth, neither VSD, DV, nor CCVSD include a commitment to the value of human-human interaction or a criterion for when human social (socio-cultural) interactions *may* be replaced by social* interactions with artificial agents. By contrast, the non-replacement maxim of ISR formulates such a criterion and a (neo-humanist) commitment to replacing human interactions only when a sufficiently uncontroversial option for value-enhancement could be realized.

5. But note that (Friedman, et al. 1997) explicitly mention that VSD is an "interactional theory" and exclude certain understandings of values.

Fifth, the strategies of VSD and DV are recommended for domains affected by the Collingridge dilemma, but not for domains affected by the triple gridlock – user experiences are still taken to be projectable, and so values remain projectable. Also CCVSD does not seem to view the descriptive and evaluative uncertainty of the domain of social robotics quite as fundamental as this is done in ISR. For CCVSD envisages ethical evaluations to be undertaken by expert consultants, drawn in at decisive stages of RDD process to offer expert advice, while in ISR participating philosophers offer their expertise as guidance within a joint creative process – they stay with the entire RDD process to facilitate what they know will remain an *adventure* of value realization, given the great plasticity of the domain of socio-cultural interactions.

Finally, ISR differs from the mentioned value-geared design strategies also by being more than just a design strategy. As stated in the 'quality principle' (P2) above, ISR aims to reposition social robotics research within the landscape of scientific investigation and engineering, as the following section will elaborate.

5. Social robotics – interdiscipline or transdiscipline?

In Section 2 above we argued that the descriptions of human-robot interactions as currently furnished by social robotics engineering and sHRI research are problematically incomplete, since processes of agentive self-understanding – at the individual and social level – should be, but are currently not, considered to be part of these interactions. In line with this observation we want to suggest that the term 'social robotics' should be understood as denoting a field of research which *contains* (i) the domain of social robotics engineering and (ii) the domain of sHRI, but goes beyond these two areas as currently conducted by taking more fully into account the socio-cultural dimensions of human interactions with so-called 'social' robots. 'Social robotics', as this term is understood in ISR, investigates the phenomena of asymmetric social interactions with embodied artificial agents, and is conducted in pluridisciplinary collaborations with wide interdisciplinary scope.

This creates the question of which *format* of pluridisciplinary collaboration social robotics is likely to take in the long run. Is social robotics going to be a transdisciplinary research area? Answers to the question depend, in the first instance, on the definition of transdisciplinarity. While older characterizations of this term by B. Nicolescu and M. Gibbons suggest in transdisciplinary research areas a "consilience of knowledge" is used to solve practical problems (Hannibal & Lindner, 2018), we take our bearings from recent work in philosophy of science and science studies.

In contemporary science studies there appears to be general agreement that *different degrees of terminological and methodological integration* can be used to distinguish between three main types of pluridisciplinary collaboration – collaboration in the form of a "multidiscipline", "interdiscipline", or a "transdiscipline", respectively (Nersessian & Newstetter, 2013). In a multidiscipline, such as climate research, "participants from disciplines come together in response to a problem, create a local integration to solve that problem, and go back to their respective disciplines, with these largely unchanged by the transient interaction" (ibid., p. 717). An interdiscipline, such as biomedical engineering, generates a new understanding of the domain and new modeling resources by the "integration of concepts, methods, materials, models" (ibid., p. 719). Finally, in a transdiscipline, such as integrative systems biology, "each field in the adaptive …problem space will likely penetrate and change significant practices in regions of the collaborating field" (ibid., 723). In order words, the conceptual and methodological results gained in the new domain of 'adaptive interdisciplinary transactions' change the terminology and/or methods of the original disciplines, thereby engendering *transformative* repercussions.

Relative to these distinctions, sHRI and social robotics engineering partly seem to display the form of organization of a multidiscipline and partly – especially when affective computing and affective learning theories are involved – integrative strands typical of an interdiscipline. But if undertaken with the ISR approach, there are good prospects that social robotics will reach the degree of integration of an interdiscipline.

As explained above, principles (P2) and (P4) in combination create continuous interaction and exchange of a multiplicity of research results and methods focused on one application context. Moreover, it is an expressed aim of ISR to devise a comprehensive descriptive framework in terms of which an empirical theory of asymmetric social interactions can be formulated (Seibt, 2020).

Of course, these are prospects and aims, not yet facts. Thus one might suspect that the question of whether ISR will turn social robotics into a transdiscipline is even more a matter of speculation about future developments. We believe, however, that already now there are three indications that the criterial *transformative* transactions can be expected.

First, the phenomena of human social interactions with robots call for a fundamental revision of the notions of subjectivity and sociality in philosophy, both of which have been resting on the traditional *res cogitans* model of the mind. According to the Cartesian conception of subjectivity, the capacities for rationality, normative competence, moral and agentive autonomy, feeling, conceptual experience, creativity or spontaneity etc. are a package deal. The fact that we readily accept artificial agents as social agents shows that these capacities can be pried

apart – social competences, rationality, and moral competences do not need to go hand in hand, and symbolic social interactions do not require the symmetric distribution of the relevant capacities for these competences. This goes against the reciprocity assumptions that is standardly presupposed of philosophical definitions of sociality.

Second, the exploration of human-robot interactions with the wide interdisciplinary scope required by ISR includes a fundamental revision of our conception of social action, which likely will have repercussions for a number of the disciplines involved (e.g., philosophy, anthropology, psychology, cognitive science, sociology). For so far these disciplines have treated social actions as though they exist in the same way as natural occurrences, but this does not seem to be the correct 'model of existence'. A natural occurrence such as the falling of a drop of water or an avalanche may be simple or very complex, but it is not perspectivally complex; what occurs may be described in many different ways, but all descriptions – if they are indeed definite descriptions of the same occurrence — have one and the same referent. By contrast, the ontological analysis of human-robot interactions has revealed that social actions – whether with humans or with robots – have irreducible perspectival complexity. If I greet the robot Asimo, who has the true description of what is occurring: me, Asimo, or you who are observing the two of us? The complexity principle (P3) of ISR expresses this insight; in essence, it says that what is going on in a social interaction is nothing simple but a complex of (at least) three occurrences from there are (at least) three viewpoints, and each of these has as much claim to truth as the others. Human-robot interactions throw this ontological complexity of social actions into strong relief.

Third, social robotics undertaken with ISR may have repercussions for the notion of 'anthropomorphization' as used in psychology. When robots are accepted as social interaction partners, HRI researchers so far have turned to psychological terminology to describe these phenomena in terms of the human 'tendency to anthropomorphize', i.e., "the tendency to attribute human characteristics to inanimate objects, animals and others with a view to helping us rationalize their actions. It is attributing cognitive or emotional states to something based on observation in order to rationalize an entity's behavior in a given social environment." (Duffy, 2003, p. 180). However, as we have argued elsewhere (Damholdt et al., forthcoming), once one operates with ISR and a richer set of methods as prescribed by (P2), there are good reasons to distinguish between 'tendencies to anthropomorphize' and 'tendencies to sociomorph'. As worked out by philosophers, there are at least ten different ways of understanding social agency, and not all of them involve the projection of characteristically "human characteristics". Thus, the projection of social agency may, but need not, involve anthropomorphizing. Especially our interactions with non-humanoid social robots seem

to involve sociomorphing rather than anthropomorphizing. Further research on the difference between anthropomorphizing and sociomorphing in the domain of human-robot interaction may lead to a transformation of the notion of anthropomorphization in psychology in general.

6. Conclusion

The aim of this paper was to offer a brief introduction to the motivations and basic ideas of the approach of Integrative Social Robotics (ISR), which proposes a new 'paradigm' for the organization of research, design, and development processes for social robotics applications. We have presented ISR, and some of the main arguments motivating the approach, in support of the current ongoing effort to open up the research field of SR and sHRI more fully to the research perspectives and investigative methods of the humanities. The particular elements of ISR are developed as a targeted response to the current constellation of theoretical and practical problems with social robotics applications, the "triple gridlock of description, evaluation, and regulation". While the details of the ISR approach are still being worked out learning-by-doing style in the context of a large research project, we have presented here the five guiding principles that summarize the core insights of ISR as conceived so far. ISR includes ideas from participatory and user-centered design, but, and even more than extant other value-geared design strategies, pursues a strictly value-driven approach that includes a substantive selection filter for legitimate social robotics applications ("Non-Replacement Principle"). Moreover, as we have tried to sketch, ISR is also intended to serve comprehensive theoretical goals. It shall provide the interdisciplinary integration needed to develop a unified framework for the description of social robotics application, and facilitate the institutional establishment of social robotics as a field where engineering, human and social sciences, and the humanities collaborate. Finally, we have suggested that there are reasons to expect that interdisciplinary research in this area will lead to fundamental revisions of concepts of the disciplines involved. This may be taken as an indication that social robotics, as undertaking with ISR, will constitute itself in the form of a new transdiscipline.

Funding

This research has been supported by a Semper Ardens Grant of the Carlsberg Foundation (CF16-0004).

References

Arkin, R.C., & Arkin, R.C. (1998). *Behavior-based robotics*. MIT press.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Bethel, C.L., & Murphy, R.R. (2010). Review of Human Studies Methods in HRI and Recommendations. *International Journal of Social Robotics*, *2*(4), 347–359. https://doi.org/10.1007/s12369-010-0064-9

Bickhard, M. H. (2009). The interactivist model. *Synthese*, *166*(3), 547–591. https://doi.org/10.1007/s11229-008-9375-x

Bickhard, M. H. (2017). Robot Sociality: Genuine or Simulation? In *Sociality and Normativity for Robots* (pp. 41–66). Cham: Springer. https://doi.org/10.1007/978-3-319-53133-5_3

Borenstein, J., & Arkin, R.C. (2017). Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior. *AI & SOCIETY*, *32*(4), 499–507. https://doi.org/10.1007/s00146-016-0684-1

Breazeal, C. (2002). *Designing Sociable Robots*. MIT Press.

Breazeal, C. (2003). Toward Sociable Robots. *Robotics and Autonomous Systems*, (*42*), 167–175.

Campbell, R. (2015). *The metaphysics of emergence*. Springer. https://doi.org/10.1057/9781137502384

Cheon, E., & Su, N.M. (2018). Futuristic Autobiographies: Weaving Participant Narratives to Elicit Values around Robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 388–397). ACM.

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. Palgrave Macmillan. https://doi.org/10.1057/9781137025968

Coeckelbergh, M., J. Loh, M. Funk, J. Seibt, M. Nørskov. (2018). *Envisioning Robots in Society – Proceedings of Robophilosophy 2018*. Amsterdam: IOS Press.

Collingridge, D. (1980). *The Social Control of Technology*. London: St. Martin's Press.

Damholdt, M., Nørskov, M., Yamazaki, R., Hakli, R., Hansen, C.V., LL, C., & NN, J. (2015). Attitudinal change in elderly citizens toward social robots: the role of personality traits and beliefs about robot functionality. *Frontiers in Psychology*, *6*, 1701. https://doi.org/10.3389/fpsyg.2015.01701

Damholdt, M., Vestergaard, C., Seibt, J. (2019). Testing for anthropomorphizations – a case for mixed methods. In Jost, C., Pedevic, B. & Grandgeorge, M. (Eds), *Methods in Human-Robot Interaction Research* (forthcoming). New York: Springer.

Darling, K., Nandy, P., & Breazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction (pp. 770–775). Presented at the *Robot and Human Interactive Communication (RO-MAN)*, 2015 24th IEEE International Symposium, IEEE Press.

Dautenhahn, K. (2013). Human-robot interaction. *The Encyclopedia of Human-Computer Interaction*, 2nd Ed. www.interaction-design.org

Dautenhahn, K., & Billard, A. (1999). Studying robot social cognition within a developmental psychology framework. In *Advanced Mobile Robots, 1999. (Eurobot'99) 1999 Third European Workshop on* (pp. 187–194). IEEE.

Dennett, Daniel C. (1989). *The Intentional Stance*. The MIT Press.

Duffy, B. R., Rooney, C., O'Hare, G. M., & O'Donoghue, R. (1999). What is a Social Robot? Presented at the *10th Irish Conference on Artificial Intelligence & Cognitive Science*, University College Cork, Ireland, 1–3 September, 1999.

Duffy, B.R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3–4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Dziergwa, M., Kaczmarek, M., Kaczmarek, P., Kędzierski, J., & Wadas-Szydłowska, K. (2018). Long-Term Cohabitation with a Social Robot: A Case Study of the Influence of Human Attachment Patterns. *International Journal of Social Robotics*, *10*(1), 163–176. https://doi.org/10.1007/s12369-017-0439-2

Fiebich, A. (2017). Social Cognition, Empathy and Agent-Specificities in Cooperation. *Topoi*, 1–10.

Fiebich, A., Nguyen, N., & Schwarzkopf, S. (2015). Cooperation with robots? A two-dimensional approach. In *Collective Agency and Cooperation in Natural and Artificial Systems* (pp. 25–43). Springer.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems*, (*42*), 143–166. https://doi.org/10.1016/S0921-8890(02)00372-X

Friedman, B., Kahn Jr, P.H., & Borning, A. (1997). *Value sensitive design and information systems*. In P. Zhang, & D. Galetta (Eds.), *Human-Computer Interaction in Management Information Systems* (pp. 348–372). New York: Routledge.

Gallagher, S., & Varga, S. (2014). Social constraints on the direct perception of emotions and intentions. *Topoi*, *33*(1), 185–199. https://doi.org/10.1007/s11245-013-9203-x

Goodrich, M. A., & Schultz, A. C. (2007). Human-Robot Interaction: A Survey. *Foundations and Trends® in Human-Computer Interaction*, *1*(3), 203–275. https://doi.org/10.1561/1100000005

Hakli, R. & Seibt, J. (Eds). (2017). *Sociality and normativity for robots – philosophical investigations*. Cham: Springer. https://doi.org/10.1007/978-3-319-53133-5

Hannibal, G., & Lindner, F. (2018). Transdisciplinary Reflections on Social Robotics in Academia and Beyond. In M. Coeckelbergh, J. Loh, M. Funk, J. Seibt, M. Nørskov (Eds.), *Envisioning Social Robots—Proceedings of Robophilosophy 2018*. Amsterdam: IOS Press.

Hasse, C. (2015). *Multistable roboethics. Technoscience and Postphenomenology: The Manhattan Papers*. Books, Lexington, 169–188.

Hasse, C. (2019a). The Vitruvian robot. *AI & Society*, *34*(1), 91–93.

Hasse, C., Trentemøller, S., & Sorenson, J. (2019). Special Issue on Ethnography in Human-Robot Interaction Research. *Journal of Behavioral Robotics*, *10*(1), 180–181.

Hasse, C., & D. M. Søndergaard (Eds.) (2019b), *Designing robots, designing humans*. New York: Routledge, 2019.

Huttenrauch, H., & Eklundh, K. S. (2002). *Fetch-and-carry with CERO: observations from a long-term user study with a service robot*. 158–163. IEEE.

Jost, C., Podevic, B., & Grandgeorge, M. (2020). *Methods in Human Robot Interaction*. New York: Springer.

Kahn Jr, P. H., Ruckert, J. H., Kanda, T., Ishiguro, H., Reichert, A., Gary, H., & Shen, S. (2010). Psychological intimacy with robots?: using interaction patterns to uncover depth of relation. Proceedings of the *5th ACM/IEEE International Conference on Human-Robot Interaction*, 123–124. IEEE Press.

Kahn, P.H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., … Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction – HRI '11* (p. 159). Lausanne, Switzerland: ACM Press. https://doi.org/10.1145/1957656.1957710

Kudina, O., & Verbeek, P.-P. (2018). Ethics from within: Google Glass, the Collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values*, 0162243918793711.

Leite, I. (2015). Long-term interactions with empathic social robots. *AI Matters*, *1*(3), 13–15. https://doi.org/10.1145/2735392.2735397

Manyika, J., Chui, M., Miremadi, M., Bughin, J., George, K., Willmott, K., & Dewhurst, M. (2017). Harnessing Automation for a Future that Works. Retrieved from https://www.mckinsey.com/mgi/overview/2017-in-review/automation-and-the-future-of-work/a-future-that-works-automation-employment-and-productivity

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

Nickelsen, N.C.M. (2018). Socio-Technical Imaginaries and Human-Robotics Proximity—The Case of Bestic. M. Coeckelbergh J. Loh, M. Funk, J. Seibt, M. Nørskov (Eds.). *Envisioning Robots in Society – Power, Politics, and Public Space*, 212–220. https://doi.org/10.3233/978-1-61499-931-7-212

Nomura, T., Suzuki, T., Kanda, T., Han, J., Shin, N., Burke, J., & Kato, K. (2008). What people assume about humanoid and animal-type robots: cross-cultural analysis between Japan, Korea, and the United States. *International Journal of Humanoid Robotics*, *5*(01), 25–46. https://doi.org/10.1142/S0219843608001297

op den Akker, R., & Bruijnes, M. (2012). Computational models of social and emotional turn-taking for embodied conversational agents: A review. *COMMIT Deliverable*.

Parviainen, J., Van Aerschot, L., Särkikoski, T., Pekkarinen, S., Melkas, H., & Hennala, L. (2016). Motions with emotions. A double body perspective and human-robot interaction in elderly care. In: J. Seibt, M. Nørskov, S. Schack Andersen *What Social Robots Can and Should Do – Proceedings of the Robophilosophy 2016 conference IOS*, Amsterdam, 210–219.

Payr, S. (2018). In Search of a Narrative for Human–Robot Relationships. *Cybernetics and Systems*, 1–19.

Robertson, J. (2017). *Robo Sapiens Japanicus: Robots, Gender, Family, and the Japanese Nation*. Univ of California Press. https://doi.org/10.1525/california/9780520283190.001.0001

Sabanovic, S. (2007). Making Friends: Building Social Robots through Interdisciplinary Collaboration. https://www.aaai.org/Papers/Symposia/Spring/2007/SS-07-07/SS07-07-016.pdf

Šabanović, S. (2010). Robots in society, society in robots. *International Journal of Social Robotics*, *2*(4), 439–450. https://doi.org/10.1007/s12369-010-0066-7

Šabanović, S., & Chang, W.-L. (2016). Socializing robots: constructing robotic sociality in the design and use of the assistive robot PARO. *AI & Society*, *31*(4), 537–551. https://doi.org/10.1007/s00146-015-0636-1

Sabelli, A. M., Kanda, T., & Hagita, N. (2011). *A conversational robot in an elderly care center: An ethnographic study*. 37–44. ACM.

Seibt, J., Hakli, R. & Nørskov, M. (Eds.) (2014a). *Sociable robots and the future of social relations – Proceedings of Robophilosophy 2014*, Amsterdam: IOS Press.

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*, pp. 97–105). IOS Press.

Seibt, J., Nørskov, M. & Schack Andersen, S. (2016a). *What Social Robots Can and Should Do – Proceedings of Robophilosophy/TRANSOR 2016* Amsterdam: /IOS Press.

Seibt, J. (2016b). Integrative Social Robotics – A new method paradigm to solve the description problem and the regulation problem? In Seibt, J., Nørskov, M., & Schack Andersen, S., *What social robots can and should do – Proceedings of Robophilosophy/TRANSOR 2016* (pp. 104–114). Amsterdam: IOS Press.

Seibt, J. (2016c). Integrative Social Robotics – Semper Ardens Project Carlsberg Foundation. Retrieved November 3, 2018, from http://www.carlsbergfondet.dk/en /Forskningsaktiviteter/Forskningsprojekter/Semper-Ardens-forskningsprojekter/NN _Integrative-Social-Robotics

Seibt, J. (2016d). How to naturalize intentionality and sensory consciousness within a process monism with gradient normativity. In O'Shea, J. (Ed.), *Sellars and His Legacy* (pp. 187–221). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198766872.003.0010

Seibt, J. (2017). Towards an Ontology of Simulated Social Interaction: Varieties of the "As If" for Robots and Humans. In Hakli, R. & Seibt, J. (Eds), *Sociality and Normativity for Robots* (pp. 11–39). Cham: Springer. https://doi.org/10.1007/978-3-319-53133-5_2

Seibt, J., Damholdt, M., Vestergaard, C. (2018). Five principles of intergrative social robotics. In Coeckelbergh, M., Loh, J., Funk, M., Seibt, J. & Nørskov, M. (Eds), *Envisioning Robots in Society – Proceedings of Robophilosophy 2018* (pp. 28–42). Amsterdam: IOS Press.

Seibt, J. (2018). Classifying Forms and Modes of Co-Working in the Ontology of Asymmetric Social Interactions (OASIS). In Coeckelbergh, M., Loh, J., Funk, M., Seibt, J. & Nørskov, M. (Eds), *Envisioning Robots in Society – Proceedings of Robophilosophy 2018* (pp. 133–147). Amsterdam: IOS Press.

Seibt, J. (2020). How to describe human 'social' interactions with robots – the ontology of simulated sociality (OASIS). In Seibt, J., Hakli, R. & Nørskov, M. (Eds), *Robophilosophy – Philosophy of, for, and by social robotics* (forthcoming). Cambridge, MA: MIT Press.

Sekiyama, K. (1999). Toward social robotics. *Applied Artificial Intelligence*, *13*(3), 213–238. https://doi.org/10.1080/088395199117405

Sharkey, N. (2008). The ethical frontiers of robotics. *Science*, *322*(5909), 1800–1801.

Sharkey, A. (2014). Robots and human dignity: A consideration of the effects of robot care on the dignity of older people. *Ethics and Information Technology*, *16*(1), 63–75.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*, *14*(1), 27–40.

Skewes, J., Amodio, D.M., & Seibt, J. (2019). Social robotics and the modulation of social perception and bias. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1771), 20180037.

Smedegaard, C. (2019). Reframing the role of novelty within social hri: from noise to information. In *14th annual ACM/IEEE International Conference on Human-Robot Interaction*.

Sparrow, R. (2016). Robots in aged care: a dystopian future?. *AI & society*, *31*(4), 445–454.

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, *16*(2), 141–161.

Sung, J., Christensen, H.I., & Grinter, R. E. (2009). Robots in the wild: understanding long-term use. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction – HRI '09* (p. *45*). La Jolla, California, USA: ACM Press. https://doi.org/10.1145/1514095.1514106

Turkle, S. (2011). *Alone Together*. New York: Basic Books.

Van den Hoven, J. (2013). Value sensitive design and responsible innovation. *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, 75–83.

van de Poel, I. (2015). Design for Values. In *Social Responsibility and Science Innovation Economy* (P. Kawalec, R.P. Wierzchoslawski, pp. 115–165). Lublin: Learned Soceity of KUL.

Van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics*, *19*(2), 407–433.

Weiss, A., Wurhofer, D., & Tscheligi, M. (2009). "I love this dog" – children's emotional attachment to the robotic dog AIBO. *International Journal of Social Robotics*, *1*(3), 243–248. https://doi.org/10.1007/s12369-009-0024-4

Weiss, A., Bernhaupt, R., & Tscheligi, M. (2011). The USUS evaluation framework for user-centered HRI. *New Frontiers in Human–Robot Interaction*, *2*, 89–110. https://doi.org/10.1075/ais.2.07wei

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01663

Wieland, W. (1999). *Platon und die Formen des Wissens*. Vandenhoeck & Ruprecht.

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Phil. Trans. R. Soc. B*, *371*(1693), 20150375. https://doi.org/10.1098/rstb.2015.0375

Zawieska, K., & Stańczyk, A. (2015). Anthropomorphic language in robotics. Presented at the *Workshop Bridging the Gap between HRI and Robot Ethics Research* at the *7th International Conference on Social Robotics* (ICSR2015).

Złotowski, J.A., Sumioka, H., Nishio, S., Glas, D.F., Bartneck, C., & Ishiguro, H. (2018). Persistence of the Uncanny Valley. *Geminoid Studies: Science and Technologies for Humanlike Teleoperated Androids*, 163–187. https://doi.org/10.1007/978-981-10-8702-8_10

Złotowski, J., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2016). Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn, Journal of Behavioral Robotics*, *7*, 55–66. https://doi.org/10.1515/pjbr-2016-0005

Address for correspondence

Johanna Seibt
Research Unit for Robophilosophy
School for Culture and Society
Aarhus University
Jens-Christian Skousvej 7
DK-8000 Aarhus C
Denmark
filseibt@cas.au.dk

Biographical notes

Johanna Seibt is Professor at the Department of Philosophy and the History of Ideas, at Aarhus University, and coordinator of the Research Unit for Robophilosophy. She works on the ontology of human-robot interactions and is currently the PI of the research project "What Social Robots Can and Should Do – Towards Integrative Social Robotics", supported by a Carlsberg Semper Ardens Grant, which involves 26 researchers from 11 disciplines. Since 2014 she coordinates the biennial Robophilosophy Conference Series.

Malene Flensborg Damholdt is an associate professor at the Department of Psychology & Behavioural Science, and at the Department of Clinical medicine. Her research focuses on the effect of individual differences in relation to social robots.
malenefd@psy.au.dk

Christina Vestergaard is a post.doc at the Department of Philosophy and the History of Ideas, University of Aarhus. She is an anthropologist with research interests is interdisciplinary methodology, anthropology of technology, and robo-philosophy.
etnocv@cas.au.dk

J. Seibt, M. Damholdt, C. Vestergaard (2020). Integrative Social Robotics, value-driven design and transdisciplinarity. *Interaction Studies* 21:1, pp. 111–144.

J. Seibt, M. Damholdt, C. Vestergaard (2020). Integrative Social Robotics, value-driven design and transdisciplinarity. *Interaction Studies* 21:1, pp. 111–144.