The GCAS Review Journal
Vol. I, Issue 2/2021

Available online at https://www.gcasreview.com/publications-technology-and-society

The GCAS
REVIEW

Human-robots Interaction: A Philosophical Framework for Social and Political Assessment

VALERIA MARTINO

University of Turin, Italy

valeria.martino@unito.it

Abstract: *The European Parliament's Resolution of 16 February 2017 about Robotics, paragraph 50 deals with the possibility of joint actions made by human beings and robots. Dealing with joint actions, however, entails speaking of sharing goals, values, norms, plans, etc. and, as a consequence, it seems to assume only people being involved. The paper is intended to explore what it means to attribute joint actions to human-robot interaction. It starts from the description of a general account of joint actions as interpersonal actions and takes into consideration the possibility for the human-robot couple to be the subject of such a kind of action. In order to better explain this point, the paper takes the interaction with a social robot as PARO as an example. In this way, we can elucidate to what extent and in which sense such an interaction can be defined as social and, thus, give birth to genuine joint actions. Indeed, such an analysis seems necessary in order to deal with the possibility of attributing responsibility to robots – i.e., another important point highlighted by the Resolution itself, which wonders if it is necessary to create a specific legal status for robots, i.e., that of 'electronic*

person'. Indeed, sociality and responsibility seem to be very related concepts; is it possible to attribute the latter to someone who can only simulate the first? A theoretical framework seems necessary in order to understand social and political implications of some assumptions we risk to take for granted.

Keywords: Joint action; Human-robot interaction; Sociality; Responsibility.

The GCAS Review Journal is a Publication of GCAS College Dublin, Ltd.

I. Introduction

The European Parliament's Resolution of 16th February 2017 about Robotics in paragraph 50 deals with the possibility of joint actions made by human beings and robots. In particular, it states:

> Notes that development of robotics technology will require more understanding for the common ground needed around joint human-robot activity, which should be based on two core interdependent relationships, namely predictability and directability; points out that these two interdependent relationships are crucial for determining what information need to be shared between humans and robots and how a common basis between humans and robots can be achieved in order to enable smooth human-robot joint action.[1]

What is interesting in these statements is the explicit reference to 'joint action' and 'joint activity' applied to the human-robot interaction. Indeed, it presupposes the possibility of having such a kind of action with a hybrid subject, which is made of both human and not-human beings or intelligences. Dealing with joint actions, however, traditionally entails speaking of sharing goals, values, norms, plans, etc. and, therefore, it seems to assume that only people are involved. In which sense, thus, the roles that give birth to the joint action can be filled not only by human beings but also by a different kind of intelligence? What should be asked is if this is just a metaphorical language or if we can properly deal with joint actions even in these cases, maybe developing a more inclusive account or a less demanding one. Indeed, we can state that even in usual interactions, using a joint action account based on collective intentionality, mental representation, and possession of reason could lead to theoretical issues or gaps. Thus, for example, we can rightly wonder what happens when one of the subjects of those actions is a young child or a disabled person: Does the account apply too? Does it need changes or adjustments? Should we exclude the possibility of having these people as subjects of joint actions?

The paper tries to explore the above mentioned topics starting from a general account of joint action which can be considered less demanding, as it uses the paradigm of 'interpersonal actions' and thus gives a

[1] The document is available at the following link: https://eur-lex.europa.eu/legal-content/IT/ALL/?uri=CELEX%3A52017IP0051.

more comprehensive account of rationality. In this way, we think it is possible to use 'joint action' properly, giving it a meaning useful for further inquiring on responsibility and sociality and their close relation. Once we have investigated such an account, we will be able to apply it specifically to the human-robot interaction, also using a practical example represented by one of the so-called social robots, i.e., PARO Robot. Indeed, many studies have already been done on this specific robot, showing that it can really improve mood and reduce stress and anxiety, especially in the case of elderly people, people with dementia, and pediatric patients. Consequently, we believe that it is a good example in order to highlight the features we need to attribute to a joint action account, if we want to understand such interactions. Hopefully, this will give us clues for better understanding our social engagement with robots and artificial intelligences in general, also surveying the meanings of 'sociality' if applied to our daily interactions with them.

I. A general account: Joint actions as interpersonal actions

'Joint action' is the usual way in which philosophy and social ontology refers to a complex phenomenon, instantiated by actions made by more than one individual. Traditional examples of joint actions are carrying a piano,[2] playing a symphony or a duet, playing a passage.[3] These are complex actions as they require each of the agents to perform their own role, a role that does not coincide with that of the others – that is the reason why the action itself can be said to be carried out by a group of people, although this statement has been intended with very different meanings. Indeed, the person who plays the piano does not play a duet, just as the person who plays the violin does not. Only the two musicians together, each according to their role, can perform the action. Moreover, traditionally, dealing with joint actions implies the reference to collective intentionality. With 'collective intentionality' we mean the capacity of several people's minds to be oriented towards the same content, project, idea, and so on. When several subjects must act as a group, coordinating to complete a task, it seems essential to understand *how* this is possible, i.e., how the coordination is achievable. One of the

[2] Raimo Tuomela and Kaarlo Miller, "We-Intentions", *An International Journal for Philosophy in the Analytic Tradition* 53, no. 3 (May, 1988): 367–389.

[3] John R. Searle, *The construction of social reality*(New York: Free Press, 1995).

answers to this question consists precisely in referring to collective intentionality. But this reference can bring with it several consequences, as it can be formulated in very different ways.

Thus, the various conceptions of collective intentionality can be classified into three different kinds: instrumentalism, the summative account, and the non-summative account.[4] The first, which is also the road less traveled, is an instrumentalist position, since it accepts collective intentionality precisely as a purely instrumental element, a metaphysical fiction that helps us in the explanation of collective action and that only for this reason we can accept. According to this account, collective intentionality does not exist; we can just use the word in order to make the explanation of our social actions easier. The second option, namely the summative one, is strictly individualistic, since it argues that collective intentionality cannot be anything other than the sum of the individual intentionality of all – or at least most of – those who participate in the joint action itself. Even this road is not the most practiced, although there are several formulations, which for the sake of brevity can be summarized as the simple and the complex accounts. Both, however, have raised several problems, highlighted from several points of view.[5] In particular, the simple version – according to which the intentionality of most individuals is enough to obtain collective intentionality – has been deemed insufficient, since it excludes common knowledge as a significant element.[6] While it seems natural that intentionality is a private factor, this account lacks the correlation between individual elements which instead seems indispensable for there to be collective intentionality. Indeed, there could be cases in which each of the members of a group has a certain intentionality, but for some reason this happens in secret, and it remains unspoken for example for fear of others' opinion. It is difficult to argue that this is a case of collective intentionality, or even that it may give rise to some joint action, since the latter presupposes that the agents are

[4] See Deborah Tollefsen, "Collective Intentionality," in *Internet Encyclopedia of Philosophy*, 2004, available at the following link: https://www.iep.utm.edu/coll-int/.

[5] For example, both Gilbert and Searle point out its weak elements. Cf. Margaret Gilbert, *On Social Facts* (London: Routledge, 1989), in which there is a detailed criticism of the summative account and John R. Searle, "Collective Intentions and Actions," in *Intentions in Communication*, eds. Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (Cambridge, MA: Bradford Books, MIT Press, 1990), 401–415, in which you can find the well-known example of the difference between the case in which several people run simultaneously towards a shelter when it starts to rain and the case in which this same collective action is carried out as part of a performance.

[6] In this context, 'common knowledge' means the fact that all the individuals involved are aware of a certain fact and, in this sense, share a mental content, knowing at the same time, to share it, that is, also being aware of the fact that the other individuals involved possess that same mental content, in the form of knowledge.

aware of the fact that other individuals are participating and sharing intentions. Michael Bratman, for example, has shown how the summative account excludes the possibility of distinguishing between the case in which an action is carried out jointly and the case in which the same action is carried out together, but independently.[7] The complex summative account, on the other hand, is too weak because it does not allow for the possibility that two or more groups are formed by the same identical members, that is, it does not explain the functioning of coextensive groups. Indeed, in this version, common knowledge is added as an indispensable element, but this same element is not further specified, so we could have the paradoxical situation of a group that has common knowledge about intentions that are completely beyond the reason why the group was assembled.[8] The features of the complex summative account are not sufficient to obtain a good description of joint actions, since one group's beliefs cannot be transferred to those of another, even when the individuals involved are the same; for example, it makes no sense to say that the chess club believes that green is the color of the year, even if the members of the chess club are identical to those of the fashion club. A counterintuitive position is thus created, and we are led to reject it.

The most adopted position is therefore the third, that is the non-summative account, in which in fact the most well-known authors connected to the notion of collective intentionality can be included. They are John Searle, Michael Bratman, Raimo Tuomela, and Margaret Gilbert.[9] Although these four authors have very different positions, more or less individualistic, or even more or less dependent on the notion of acceptance or belief, they all reject the idea that in order to understand what collective intentionality is it is sufficient to sum individual intentionalities. However, this does not mean that they accept the paradoxical idea of the existence

[7] See Michael Bratman, "Shared Cooperative Activity," *The Philosophical Review* 10, no. 2 (Apr. 1992): 327–341; Michael Bratman, "Shared Intention," *Ethics* 104 (1993): 97–113.

[8] See Margaret Gilbert, "Modelling Collective Belief," *Synthese* 73, no. 1 (Oct. 1987): 185–204.

[9] These four authors, due to their relevance in the debate, that took place also and above all on the criticism and defence of their theories, have also been defined as the 'big four'. See Sarah R. Chant, Frank Hindriks, Gerhard Preyer, "Beyond the Big Four and the Big Five," in *From Individual to Collective Intentionality*, eds. Idd. (New York: Oxford University Press, 2014), 1–6. Here, we can also see how the debate is taking on a further connotation in recent years, in particular in the so-called phenomenological approach. Indeed, there is an attempt to identify the specific functioning of collective intentionality, starting from the way in which the collective experience is lived by the individual, from psychology and cognitive philosophy's perspective. See for example, Deborah Tollefsen, "A Dynamic Theory of Shared Intention and the Phenomenology of Joint Action," in *From Individual to Collective Intentionality*, 13–33; Kate Crone, "Collective Attitudes and the Sense of Us: Feeling of Commitment and Limits of Plural Self-Awareness," *Journal of Social Philosophy* 49, no. 1 (2018): 76–90; Elisabeth Pacherie, "The Phenomenology of Joint Action: Self-Agency vs. Joint-Agency," in *Joint Attention: New Developments*, ed. Axel Seemann (Cambridge, MA: MIT Press, 2012), 343–89.

of a collective, unique, and common mind, to attribute a collective intentionality to.[10] The challenge of the theme lies precisely in this: to resolve a tension between two apparently opposite propositions, namely the so-called 'Irreducibility Claim' and 'Individual Ownership Claim'.[11] The theories on collective intentionality, in fact, on the one hand affirm that it cannot be reduced to the sum of the individual ones (Irreducibility Claim), but on the other hand they affirm that it cannot be found anywhere, except in the minds of individuals (Individual Ownership Claim). The answers to this paradox have been different and depend on a different conception of what is collective in collective intentionality: the content of the intention for Bratman, the mode of intention for both Tuomela and Searle, or the subject who thinks the intention for Gilbert.[12] Despite these differences, which are far from being minimal or subtle, we can affirm that collective intentionality, in general, constitutes a theoretical proposal that has been considered convincing, even outside the field of the philosophy of action, because it allows us to shed light on how individuals act within society. For instance, studies of psychology and theory of mind highlight how the asymmetry of roles can lead those with more power, or more agency, to feel greater commitment to the joint goal or action.[13] They also highlight how the sense of authorship is an experience necessarily connected to our daily joint actions.[14] Again, the perception of being part of a larger group who acts towards a common goal, pushes us to be more attentive to that goal, especially if achieving it alone is impossible. Indeed, there are dynamics that take place in very young children too that cause human beings to feel a commitment to achieve common goals or to help others achieve a certain purpose which, for this reason, becomes common.[15] At the basis of such tendencies or attitudes, we should presuppose the capacity of being joint committed as the background of collective intentionality and

[10] Instead, for example, Philip Pettit deals with the existence of collective mind in trying to account for particular cases, i.e., all those in which a group, such as an organization, has common intentions or beliefs, but these cannot in any way be reduced to those proper to single individuals. Cf. for example, Philip Pettit, "Groups with Minds of Their Own," in *Socializing Metaphysics. The Nature of Social Reality*, ed. Frederick F. Schmitt (Lanham, MD: Rowman & Littlefield, 2003), 167–94.

[11] David P. Schweikard and Hans Bernhard Schmid, "Collective Intentionality," *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2013/entries/collective-intentionality/>.

[12] David P. Schweikard and Hans Bernhard Schmid, "Collective Intentionality".

[13] Pacherie, "The Phenomenology of Joint Action: Self-Agency vs. Joint-Agency".

[14] Tim Bayne and Neil Levy, "The Feeling of Doing: Deconstructing the Phenomenology of Agency," in *Disorders of Volition* eds. Natalie Sebanz and Wolfgang Prinz (Cambridge: MIT Press, 2006), 53–57.

[15]. Bayne and Levy, 49–68.

joint actions themselves. It is therefore a question of attitudes ascribable to human beings, which cognitive psychology helps recognize and conceptualize, and which philosophy can make its own in the explanation of some phenomena. Nonetheless, the fact that other sciences are highlighting the importance of collective intentionality does not mean that it has to be understood through a non-summative account or that collective intentionality should necessarily lead to a specific kind of joint action such as sharing values and norms through collective intentionality itself.

Although usual accounts of joint actions refer to collective intentionality, in the terms just explained, there are other ways to conceptualize such phenomena. In particular, a less considered account dealing with joint action is Seumas Miller's one.[16] Through its teleological form, it does not constitutively distinguish between individual and joint action and, by limiting the morality of the joint action itself, it can also be applied to phenomena on a large scale, both in space and time, such as the construction of the Great Wall of China, or those long-term and particularly interesting actions, such as the protection of our environment. In this way, it is possible to foster a theory which considers joint actions as interpersonal actions performed to realise a shared end or, alternatively, considers 'collectivity' as a feature of the actions' content[17] – rather than of the modalities of thought or of the subject of the action. As we mentioned before, this is the case of Bratman's account, which indeed has been used as a starting point for the development of the BDI model (Beliefs-Desires-Intentions), i.e., the explanation of Artificial Intelligence behaviour.[18] Indeed, his account deals with shared intentions as interrelations between agents. According to him, there are individual intentions that can be correlated with each other in the formation of common projects, which can be negotiated, to produce a meshed plan, accepted by all the parties involved. Consequently, the study of intentionality partly coincides with that of design.[19] Shared intention is not an attitude in someone's mind. Rather it is a state of

[16] Seumas Miller, *Social Action. A Teleological Account* (Cambridge: Cambridge University Press, 2001).

[17] As in Michael Bratman, *Intention, Plans, and Practical Reason* (Cambridge: Harvard University Press, 1987).

[18] See Krister Segerberg, John-Jules Meyer, and Marcus Kracht, "The Logic of Action," *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/logic-action/>.

[19] See Michael Bratman, "Shared Intention," *Ethics* 104 (1993): 106. This specific aspect is also evident in Michael Bratman, "What is Intention?," in *Intentions in Communication*, eds. Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (Cambridge, MA: Bradford Books, MIT Press, 1990), 15–31. Here, in fact, the author specifies how 'intention' means 'intention to do something' and, therefore, how it has to do with future actions and the possibility of formulating plans and strategies aimed at achieving those same intentions. This also implies the ability of the individual to formulate

affairs, which consists primarily of attitudes and interrelationships. According to Bratman we can understand shared intentions as follows:

> We intend to *J* if and only if
> 1. (*a*) I intend that we *J* and (*b*) you intend that we *J*.
> 2. I intend that we *J* in accordance with and because of 1*a*, 1*b*, and meshing subplans of 1*a* and 1*b*; you intend that we *J* in accordance with and because of 1*a*, 1*b*, and meshing subplans of 1*a* and 1*b*.
> 3. 1 and 2 are common knowledge between us.[20]

This is a first point we should keep in mind, in order to have a general account of joint actions. Can this less demanding definition of shared intention be applied to robots too? Can a robot 'intend that we J' and have common knowledge with us about it? To answer those questions, it seems necessary to ask what rationality should be. Indeed, dealing with intentions, knowledge or even representation entails talking about reason and rationality. Is it possible to define a robot rational?

Consequently, the second point to take into consideration is the reference to Seumas Miller's account. Following it, we can question whether the level or type of intelligence is a necessary element to describe the agent-type we need. In the words of Miller himself:

> But why should a difference in the level of intelligence – whether manifested in the propositional character of mental states or in the fact that they are higher versus lower order mental states – of those involved in conflict make a difference to the sociality of their interaction?[21]

In his analysis, Miller supports a conception of social action, which can be appropriate for our purpose. Indeed, as it can be seen from the quotation above, he does not connect the feature of sociality to the possession of a certain type of intelligence. Again, reporting Miller's words: «I suggest that the level of agency required to have such attitudes [to act jointly] is something less than rational agency. I will refer to such agents

coherent intentions. In Bratman's example, if I intend to go to Monterey and to leave the car at home for Susan, in going through the process that will take me from the simple intention to its realization, I cannot express the plan to go to Monterey by car. The intentions are therefore subject to the criterion of coherence between means and ends.

[20] Bratman, "Shared Intention".

[21] Miller, *Social Action*, 50.

as basic agents or simply agents».[22] This means that to speak of sociality and to attribute it to the subjects of a joint action they need not have shown a certain level of intelligence or understanding, set as a limit – as if for participating in a social interaction were required to first pass a test. This seems to be a hard requirement for humans too. Therefore, we can define 'agent' anyone (or anything) able to use information to achieve a purpose, according to a minimum definition of agency, regardless of the degree of competence or understanding achieved.[23] Thus, we have on the one hand the description of a joint action as the interrelation between the individual actions of two or more individuals who have individual (but shared) intentions and, on the other hand, the possibility of performing them without a necessary level of rationality. In this context 'shared' simply means that people involved are conscious that all other agents (robots included) are willing to collaborate in order to achieve a joint goal. As a result, we can state that a joint action is an action done by two or more individuals who necessarily play their role and, in doing this, for this very reason, form a temporary group – whose duration depends on the achievement of that goal. Their collaboration is then limited to such an achievement and led by individual intentions meant as the result of the capacity to include other beings as a necessary partner in order to obtain that same achievement. It should be noted that even in the traditional cases of joint actions between human beings there may be examples of short-term actions limited to the achievement of a specific purpose. In this sense, there would be no real difference, but rather we would have a more inclusive paradigm that can be applied to one case as much as to the other. Bearing this in mind, we can wonder in which sense the roles that give birth to the joint action can be filled not only by human beings but also by a different kind of intelligence, wondering more specifically to which extent this paradigm can be applied to the human-robot interaction; it seems to depend on how we can define 'sociality'.

II. The human-robot couple as the subject of joint actions: Inquiring sociality

[22] Miller, *Social Action*, 76.

[23] See Lawrence H. Davis, "What It Is like to Be an Agent," *Erkenntnis* 18, no. 2 (Sep., 1982): 195–213; Andy Clark and Josefa Torobio, "Doing without Representing?," *Synthese* 101, no. 3 (Dec., 1994): 401–431; Christian List, "What is it Like to be a Group Agent?," *Nous* 52, no. 2 (2018): 295–319; Colin Allen, "Artificial life, artificial agents, virtual realities: technologies of autonomous agency," in *The Cambridge Handbook of Information and Computer Ethics*, ed. Luciano Floridi (Cambridge: Cambridge University Press, 2010), 219–233.

According to what has been said so far, it could be quite easy to speak of robots as subjects of joint actions. But joint actions are also meant to represent the expression of sociality, as far as it is precisely through joint actions that we can express our being social animals, engaging in common goals and working together in order to realize them. Is it possible for non-humans too to be engaged in such interactions? A way to understand it is to analyze the meaning of sociality. Indeed, we have seen that when a joint action takes place, at the same time, a group comes into being. But being a group can mean different things. We mentioned before the possibility of representing groups as 'realizations of structures', that is the fulfilling of roles connected by different relationships. A group, however, can be understood also as a random set of people and/or things, or as a proper sociological group, i.e., a social group whose members recognize themselves as part of the proper group.[24] It is according to this meaning, that it seems necessary to survey the concept of sociality. As, what we are trying to grasp is the understanding of what it is to act as a group. But if it seems difficult to say that robots feel as part of a group, what does 'social' mean in this context? Are Miller's words reported above still appropriate or do they require an adjustment?

Addressing the issue of defining the term 'social' is anything but simple. The question, in fact, takes on a triple complexity: first of all, it has a very long history behind it, which already begins with ancient philosophy and which cannot in itself be retraced in its entirety;[25] secondly, it is a term that can refer to things that are very different from each other, such as objects, phenomena, relationships; thirdly, it takes on different meanings depending on the term it is opposed to, whether it is *natural*, *mind-independent*, *private*, *psychological* or *individual*. As a general statement, however, we grasp that the term must take on a specific meaning; it cannot be used in the general meaning of 'what has to do with society'. This, in fact, would involve a tautology that is

[24] Thus, for example, Henri Tajfel notes that there are two distinct senses of the term 'group': "(1) objective collections of similar individuals as defined by outside observers […] (2) groups defined as such by their members through patterns of interaction and shared representations, that is, a dynamic social process in which the capacity of people to represent themselves as members of social categories is part of the process by which sociological categories may become meaningful social groups". Cf. Henri Tajfel, "Social psychology of intergroup relations," *Annual Review of Psychology* 33 (1982): 1-39.

[25] Analysis of the existence and nature of social entities can be found throughout the ancient tradition, and then continued in modernity – with well-known examples in Hobbes, Pufendorf, Rousseau, and Locke – as well as in the Marxist tradition, to make only some of the innumerable possible examples.

difficult to avoid, as well as a trivial meaning, which would add nothing to our attempt to understand groups and social interactions *as social*. At the same time, we will see how it seems difficult to separate its meaning from a biological background which helps recognize us as social beings.

Concerning the first point, we can just give a hint of this long history, referring to the existence of two main strands that deal with the relationship between the individual and society. Indeed, it is precisely on this dyad that philosophical analysis has been most focused: the first strand sees in society – to use a generic term that has been declined in several ways – a constitutive element of the individual (a classic example is represented by the well-known Aristotelian definition of the human being as *politikon zoon*); the second one – which has mainly been developed since the 1600s – gives predominance to the individual – to his or her will and desires – in order to explain things greater than the individual, such as the birth of society and therefore of the national states. An illustrative author in this sense is of course Thomas Hobbes. According to this first step, it could be interesting to notice the existence of groups and the phenomenon of grouping as typical of human beings (among other animals).[26] Indeed, we will see how this human feature is essential in order to understand how groups can come into being even in the human-robot interaction.

Regarding the second point, we can say that the concept of social reality is the main object of investigation of social ontology, i.e., the field of inquiry of joint actions and collective intentionality, which then declines the most different aspects of sociality.[27] It remains true, however, that even within this specific context, what can be defined as social are very different things and the possibilities of declining it are very numerous – so that this road appears fruitless for our purpose as saying that interactions as relationships are social seems to add nothing if not further specified.

Coming to the third point, 'social' must not be opposed exclusively to 'natural', a contrast that among other things creates many problems, but must provide a specific meaning that makes the expression

[26] A relevant position in this regard is Frans De Waal's one. He explicitly criticizes the concept of human being inspired by Hobbes' philosophy and elaborates a conception that takes into account sociality and the emotional sphere as a basis for morality already at the biological level, as an expression of the development of the human species, from the specific point of view of biology. Cf. Frans De Waal, *Primates and Philosophers* (Princeton: Princeton University Press, 2006).

[27] Brian Epstein, "Social Ontology," *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/social-ontology/>.

substantial. Another classic contrast, indeed, consists of the social-psychological dyad, often treated in an anti-reductionist key. The latter also brings with it the distinction between what depends on the mind and what is independent from it – where society is often taken as the expression of what depends on our will and nature as what is (human) mind-independent. This theme is connected to the so-called self-fulfilling prophecies, that is the fact that in some special cases it seems that what apparently does not depend on us can anyway adapt to our ideas or theories about it. This distinction was used to distinguish the social sciences from the natural ones, but also questioned as valid.[28] Then, in our case the adjective 'social' must be opposed to 'psychological' – as far as a social interaction is something that happens outside of our minds – and to 'private' as well, as sociality implies the expression of the intention to be engaged in a joint action, as we have seen before, even though this same expression could be implicit, for example in our behavior or gestures, without any verbal formulation.[29]

Another important point to take into consideration is that traditionally, in the literature dealing with sociality and human interactions, we find definitions such as the following: «A phenomenon is a social phenomenon if and only if it involves one person's being connected either mentally or in some causal way with another person or persons».[30] Is it enough to do without the term 'person' to readjust the definition without losing its applicability? A first answer seems to be negative, as 'being mentally connected' implies not only that I as a human being have a representation of the machine in my mind, but also that the machine has a representation of me – which we are not inclined to concede. Certainly, the reference to representation[31] is not without implications as the apparent simplicity of the statement might suggest. There are different ways of conceptualizing it, on which the possibility of attributing its capacity to artificial intelligences also depends. Rather, what matters is that if we refer to intellectual abilities, feelings, or reason – although such concepts

[28] See Francesco Guala, *Understanding Institutions: The Science and Philosophy of Living Together* (Princeton: Princeton University Press, 2016), 153–163.

[29] About this possibility meant as 'readiness to act', cf. in general Margert Gilbert theory, for example in Margaret Gilbert, *On Social Facts* (London: Routledge, 1989).

[30] M. Gilbert, "Concerning Sociality: The Plural Subject as Paradigm," in *The Mark of the Social: Discovery or Invention?*, ed. John D. Greenwood (Lanham: Rowman & Littlefield 1997), 7–36.

[31] On the complexity of the debate on the notion of mental representation, cf. David Pitt, "Mental Representation," *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2020/entries/mental-representation/.

may be theorized in completely different ways – we will never be able to find a solution to the debate. The latter should not focus on the differences or analogies that exist between human and machine, but rather grasp the potential of their interaction. Referring to mental abilities in turn seems to assume that, in order to attribute sociality in the full sense to a subject, the latter must possess both reason and the ability to represent,[32] at least in a minimal sense. It is certainly possible to speak of degrees of sociality, but the one just reported above seems to be a minimal definition, not subject to further minimization – under penalty of loss of sense of the very concept of sociality.

However, this kind of argument seems problematic: if on the one hand, it supports our perception that sociality must characterize us as human beings, on the other, it implies theoretical problems related for example to the issues of agency. Even in that context, as we have seen, the possibility to act depends on the definition of rationality and the ability to represent ends that can be minimized.

Another possibility to grasp the meaning of sociality is thus to focus on the element of reciprocity – understood in a more generic way than mental representation. Indeed, it is the absence of a real exchange between human and machine that makes us unwilling to recognize an element of sociality in this type of relationship. The machine, however developed, remains a tool that humans use to achieve a purpose.[33] The idea is that, as much as we may be willing to have friendly relations with machines, we are still capable of discriminating between a machine and a human, in our ability to relate with them[34] – at least according to the technological innovations we are witnessing now. In particular, this impossibility seems to be due to an obvious deficiency of the machines, namely their inability to feel emotions and, consequently, to empathize with the humans they interact with. It is true that we are not interested in considering the real intentions of

[32] Raul Hakli and Johanna Seibt (eds.), *Sociality and Normativity for Robots* (Springer International, 2017), 17– 24.

[33] Even when the goal in question seems to be something that characterizes the human being, in its highest purposes, as in the case of caring relationships – which interest the so-called social robots – or in couple relationships – as in the case of Azumi Hikari, the Japanese hologram that acts as a virtual assistant for singles. At the following link you can find the advertisement for the hologram in question: https://www.youtube.com/watch?v=nkcKaNqfykg.

[34] Although it is possible to at least question such a claim, if we consider the spread of an ethical-legal debate based on the need to recognize, and consequently to assert, the right to know the nature – whether human or artificial – of one's own interlocutor. About this debate, cf. Statement of Artificial Intelligence, Robotics and 'Autonomous Systems' by EGE, 11; World Commission on the Ethics of Scientific Knowledge and Technologies (COMEST), Unesco, *Report on Robotics Ethics*, 2017.

individuals, since we can judge actions and their results without referring to hidden intentions, but the instrumentality of the relationship is so evident that leaves little room to other considerations.

If in general it makes sense to speak of sociality, it is because this concept allows us to understand some of our specific features and, at the same time, to pass from the attribution of sociality to that of responsibility – where moral responsibility must be well distinguished from the legal or the social ones. Sociality can be understood in biological terms, and, in this sense, it appears meaningful as long as it involves living beings.

Consequently, if it is possible to minimize the definition of 'agent' to include also machines that are sufficiently sophisticated to be able to act in function of a specific objective, it seems not so necessary to modify the concept of sociality to the point of attributing it to every form of interaction between two elements of whatever nature they are. Therefore, it seems possible to distinguish two different uses of 'sociality'. On the one hand, it is a feature that humans have as animals and, in this sense, it is not applicable to robots too, but on the other hand it could be applied for analogy to all human interactions as far as humans are willing to do it, in accordance with their proper attitudes.

We can therefore affirm that there are certainly cases of social groups that include machines, and this happens either when the machine, although sophisticated, is seen as a simple tool but the group is composed of several human beings or when the machine is strongly anthropomorphized up to be treated by those who use it *as if* it were a human being. While the first case is not really interesting, as it relapses into a traditional social interaction, in the second case, sociality will be grounded in the feeling of being part of a social group that can share the aim of carrying out a certain action – a feeling felt only by the human beings involved but sufficient to create a sense of unity. In this sense, if it seems difficult to attribute a real sociality to this type of groups, we can nonetheless acknowledge a possible change with respect to imaginable future scenarios, as well as the possibility of recognizing the existence of something that is very close to a social group, in those cases – certainly not paradigmatic nor common – in which human beings develop a sense of belonging to a group, even if the other individual involved does not belong to our species. However, this does not imply that a joint

action could not come into being; what seems to happen is merely that such actions do not necessarily lead to a proper social group's birth.

Being able to reproduce human actions and being capable of adapting to the external environment are key features to improve robots and humanoids in order to ascribe them a form of sociality. As, even if they are able to do things very hard for humans, they seem to show an inadequate development concerning interaction skills.[35] In which sense, then, can we call them 'social robots'? We can better understand it by using an example.

III. An example: PARO, the social robot

First, let's briefly describe this robot. It is a Japanese baby seal shaped robot designed by Takanori Shibata in order to relieve depression and anxiety and therefore to have a calming effect on patients in hospitals, nursing homes, and retirement homes. It is therefore a therapeutic robot in all respects,[36] which allows the application of *pet therapy*, with a not insignificant difference; according to its creator, the use of an animal-like shaped robot instead of animals, would have all the advantages of this kind of therapy, without the burdens of caring for the animal or their risking disease and death. What PARO does, therefore, is to simulate the relationships that usually are established between a patient and the person who takes care of him or her, being designed to receive affection and react docilely, thus being able to create a general situation of calm and relaxation.

As we have tried to show, the reason why we believe it is interesting to consider this kind of social or therapeutic robots, is precisely the fact that the relationships established seem to be the same as those that

[35] According to the so-called Moravec paradox, which states: "it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility." Hans Moravec, *Mind Children* (Cambridge, MA: Harvard University Press, 1988), 15.

[36] A description of this robot can be found on the official website: http://www.parorobots.com/. The reason why we have chosen this example is also due to the fact that its use has been discussed by various scholars. For example, Sherry Turkle expressed doubts about the use of these means, especially by those who find human relationships difficult, who could therefore withdraw into themselves, interacting only with robots that simulate relationships, but cannot really establish them. Cf. Sherry Turkle, *Alone together: Why we expect more from technology and less from each other* (New York: Basic Books, 2011). On the same subject, cf. also Ingar Brinck and Christian Balkenius, "Mutual Recognition in Human-Robot Interaction: A Deflationary Account," *Philosophy & Technology* 33 (2020): 53–70.

would be formed between two people with the same roles. The issue we are considering is precisely whether the PARO-elderly couple can be conceived as a group exactly like the nurse-elderly couple, since the individuals of the second couple seem to have the same type of structure of the first. Our answer, however, was negative. Indeed, this same end is only metaphorically attributable to PARO – although the same must be applied to therapy animals – since it seems that the only way we have to refer to the notion of end in a similar context is identifying it with the function for which the robot in question was designed. Furthermore, the relationships that are established seem to be necessarily univocal, that is to say they only go in the human-machine direction, but not vice versa and, in this sense, they cannot be properly defined as social. Therefore, this is an example of a hybrid group in the proper sense, which can barely be called a social group. At the same time, however, we were able to point out how there are some elements that can make this couple a group, in particular when the human involved feels part of it. Different would be the case of, let's say, older people interacting with a PARO robot used for their care. In this second case, however, once again, sociality would be given above all by the relationships between the human individuals involved. Consequently, we would have a sophisticated robot, which reacts to different situations in an appropriate way, but the sociality and purpose of the group would once again be assigned only to human beings and not to hybrid subjects themselves, who would continue to have the same role as other objects. Nevertheless, a difference can be seen. Think for example of a simple electrical appliance as an oven. In baking a cake, we use the oven in its proper function – apparently as in the PARO case. But we are willing to notice a difference as an oven has very simple functions and the interactions it can instantiate are few, well established, and completely passive. On the contrary, in the case of a social robot, we have a machine that can act and moreover react to our actions. In this way, even if intentions, believes, and desires can be attributed to the robot only metaphorically, exactly as only metaphorically it can be defined social, notwithstanding it can engage in social interactions as joint actions, mostly because of our proper nature of social beings capable of interacting socially even with objects (if we are prone to attribute them some capacities as that of reacting, adjusting targets, engaging in means-ends relations). Indeed, social robots are and should be designed taking into consideration our social attitudes, in

order to respond to our needs and proper feelings of being part of groups and communities.[37] In this sense, we can remark what we were supposing at the beginning: if we deal with joint actions, putting aside the whole complex set of concepts usually related to them, namely collective intentionality, rationality, and sociality, we can refer to the human being-robot couple as their subject, without losing a meaningful sense of the expression. Joint action remains a specific kind of interaction meant as the basis for achieving common tasks and forming short-term groups who can act in accordance with the same specific end and humans can be engaged in them with robots.

IV. The possibility of attributing responsibility to robots

In conclusion, we can add something to our considerations referring to another concept related to the topic analyzed, i.e., responsibility. Indeed, sociality and responsibility seem to be very related concepts: is it possible to attribute the latter to someone who can only simulate the first? Apparently not, or at least it is possible just in a metaphorical way. Of course, firstly, it is possible and even necessary to distinguish between moral and legal responsibility – which do not necessarily go together. When the Resolution quoted at the beginning refers to responsibility, it only deals with the latter. At the point 59, for instance, it states:

> Calls on the Commission, when carrying out an impact assessment of its future legislative instrument, to explore, analyse and consider the implications of all possible legal solutions, such as […] creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently.[38]

[37] A good example of this can be found in the ERC project wHISPER, https://whisperproject.eu/aims. Cf. for example, Sara Incao, Francesco Rea, and Alessandra Sciutti, "A Self for robots: core elements and ascription by humans," Human-Robot Interaction (HRI) 2021 Workshop on Robo-Identity: Artificial identity and multi-embodiment, Boulder, USA, March 8, 2021, https://doi.org/10.5281/zenodo.5645583 and Ana Tanevska, Francesco Rea, Giulio Sandini, Lola Cañamero, and Alessandra Sciutti, "A Socially Adaptable Framework for Human-Robot Interaction," *Front. Robot. AI* 7, no. 121 (Oct., 2020), doi: 10.3389/frobt.2020.00121.

[38] https://eur-lex.europa.eu/legal-content/IT/ALL/?uri=CELEX%3A52017IP0051

But it seems that also limiting its applicability, the concept of accountability lacks its meaning if applied to robots. Indeed, when we talk of accountability, responsibility, culpability, and so on, we are always referring to something else too. We mean that the concept goes always together with other concepts, i.e., that of morality, which we have excluded, or that of being rehabilitated to social life. This seems to be a nonsense if its subject is a robot as there is no sense in jailing a lion who killed a man; it could be just a matter of safety (for other humans) but not of real accountability or responsibility for lions. As a consequence, for sure, it is necessary to introduce regulations about the possible uses of robots and artificial intelligences, but just with a focus on possible harms or disadvantages for human beings. In this sense, it seems we could still refer to the famous Three Law of Robotics by Isaac Asimov.[39] Indeed, even though they come from a fiction, they are a reference point in studies on robotics. In this regard, they are useful as they show how, even in this paradigmatic case, the reference to responsibility is limited to the area of harm (as in the First Law). Moreover, the Second Law highlights the subordination of robots to the human will, pointing out at the same time the impossibility of dealing with proper sociality for human-robot interaction.

Another case would be that of attributing responsibility to the human-robot couple, as a hybrid subject,[40] but this would lead us too far. For our purpose, we could recognise the importance of avoiding dangerous interactions and that of limiting or forbidding them, or asking for damages compensation to their productors, for example, or their users, but not to robots themselves. This is a promising field of inquiry, which we have just mentioned to shed light on the complexity of the topics raised by our interaction with robots and artificial intelligences. If it seems something we could not avoid still it is something we are asked to conceptualize in order to avoid the risks related to the blind acceptance of something that can be changed.

V. Conclusion

[39] They can be found in the book *I, robot* by Isaac Asimov and they state: "First Law. A robot may not injure a human being or, through inaction, allow a human being to come to harm. Second Law. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. Third Law. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law".

[40] See Raul Hakli and Pekka Mäkelä, "Moral Responsibility of Robots and Hybrid Agents," *The Monist* 102, no. 2, (April 2019): 259-275, https://doi.org/10.1093/monist/onz009.

The paper started from the observation of a possible formation of joint actions between humans and robots. In order to test the validity of such an attribution, we first investigated the notion of joint action and its correlation with those of collective intentionality and rationality. We have thus come to affirm that it is possible to define a joint action as the performance of different roles by individuals to whom it is possible to attribute individual but shared intentions and a minimum definition of agency. Talking about joint actions, however, also implies talking about sociality. We therefore asked ourselves what 'sociality' means in this context and whether it is possible to attribute it to robots. Our response to the question was negative: robots can only simulate it unless we are willing to distort the meaning of 'social'. However, robots can engage in joint actions with humans – actions that will not form a real, long-lasting social group, but which will be limited to achieving some goals. We have exposed this possibility through the example of a social robot as PARO. We therefore asked ourselves if we can attribute responsibility to those who simply simulate sociality and we answered in negative terms. Indeed, the concept of responsibility brings with it a whole semantic area that acquires meaning only if connected to human beings. We can speak of responsibility as long as it has to do with the compensation for any damage suffered by a human being, but not in all the other meanings that 'responsibility' assumes, such as the reintegration into society of those responsible for a fault, re-education, repentance, in short, the whole sphere linked to emotions and sociality that we have excluded to be attributable to robots.

Bibliography

Allen, Colin. 2010. "Artificial life, artificial agents, virtual realities: technologies of autonomous agency." In *The Cambridge Handbook of Information and Computer Ethics*, ed. Luciano Floridi (Cambridge: Cambridge University Press), pp. 219-233.

Bayne, Tim and Neil Levy. 2006. "The Feeling of Doing: Deconstructing the Phenomenology of Agency." In *Disorders of Volition*, eds. Natalie Sebanz and Wolfgang Prinz (Cambridge: MIT Press), pp. 53-57.

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press.

—. 1990. "What is Intention?." In *Intentions in Communication*, eds. Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (Cambridge, MA: Bradford Books, MIT Press), pp. 15-31.

—. 1992. "Shared Cooperative Activity." *The Philosophical Review* 10, no. 2 (Apr.): 327-341.

—. 1993. "Shared Intention." *Ethics* 104: 97-113.

Brinck, Ingar and Christian Balkenius. 2020. "Mutual Recognition in Human-Robot Interaction: A Deflationary Account." *Philosophy & Technology* 33: 53-70.

Chant, Sarah R., Frank Hindriks, and Gerhard Preyer (eds.). 2014. *From Individual to Collective Intentionality*. New York: Oxford University Press.

Clark, Andy and Josefa Toribio. 1994. "Doing without Representing?." *Synthese* 101, no. 3 (Dec.): 401-431.

Crone, Kate. 2018. "Collective Attitudes and the Sense of Us: Feeling of Commitment and Limits of Plural Self-Awareness." *Journal of Social Philosophy*49, no. 1: 76-90.

Davis, Lawrence H. 1982. "What It Is like to Be an Agent." *Erkenntnis* 18, no. 2, (Sep.): 195-213.

De Waal, Frans. 2006. *Primates and Philosophers*. Princeton: Princeton University Press.

Epstein, Brian. 2018. "Social Ontology." *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/social-ontology/>.

European Group on Ethics in Science and New Technologies. 2018. *Statement on artificial intelligence, robotics and 'autonomous' systems : Brussels, 9 March 2018*, Publications Office, 2018, **https://data.europa.eu/doi/10.2777/786515**

Gilbert, Margaret. 1987. "Modelling Collective Belief." *Synthese* 73, no. 1 (Oct.): 185-204.

—. 1989. *On Social Facts*. London: Routledge.

—. 1997. "Concerning Sociality: The Plural Subject as Paradigm." In *The Mark of the Social: Discovery or Invention?*, ed. John D. Greenwood (Lanham: Rowman & Littlefield), pp. 17-36.

Guala, Francesco. 2016. *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton: Princeton University Press.

Hakli, Raul and Johanna Seibt (eds.). 2017. *Sociality and Normativity for Robots*. Cham: Springer International.

Hakli, Raul and Pekka Mäkelä. 2019. "Moral Responsibility of Robots and Hybrid Agents." *The Monist* 102, no. 2, (April): 259-275, https://doi.org/10.1093/monist/onz009.

Incao, Sara, Francesco Rea, and Alessandra Sciutti. 2021. "A Self for robots: core elements and ascription by humans." Human-Robot Interaction (HRI) 2021 Workshop on Robo-Identity: Artificial identity and multi-embodiment, Boulder, USA, March 8, 2021, https://doi.org/10.5281/zenodo.5645583.

List, Christian. 2018. "What is it Like to be a Group Agent?." *Nous* 52, no. 2: 295-319.

Miller, Seumas. 2001. *Social Action. A Teleological Account*. Cambridge: Cambridge University Press.

Moravec, Hans. 1988. *Mind Children*. Cambridge, MA: Harvard University Press.

Pacherie, Elisabeth. 2012. "The Phenomenology of Joint Action: Self-Agency vs. Joint-Agency." In *Joint Attention: New Developments*, ed. Axel Seemann (Cambridge, MA: MIT Press), pp. 343-89.

Pettit, Philip. 2003. "Groups with Minds of Their Own." In *Socializing Metaphysics. The Nature of Social Reality*, ed. Frederick F. Schmitt (Lanham, MD: Rowman & Littlefield), pp. 167-94.

Pitt, David. 2020. "Mental Representation." *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2020/entries/mental-representation/.

Schweikard, David P. and Hans Bernhard Schmid. 2013. "Collective Intentionality." *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2013/entries/collective-intentionality/>.

Searle, John R. 1990. "Collective Intentions and Actions." In *Intentions in Communication*, eds. Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (Cambridge, MA: Bradford Books, MIT Press), pp. 401-415.

—. 1995. *The construction of social reality.*New York: Free Press.

Segerberg, Krister, John-Jules Meyer, and Marcus Kracht. 2020. "The Logic of Action." *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/logic-action/>.

Tajfel, Henri. 1982. "Social psychology of intergroup relations." *Annual Review of Psychology*33: 1-39.

Tanevska, Ana, Francesco Rea, Giulio Sandini, Lola Cañamero, and Alessandra Sciutti. 2020. "A Socially Adaptable Framework for Human-Robot Interaction." *Front. Robot. AI* 7, no. 121 (Oct.). doi: 10.3389/frobt.2020.00121.

Tollefsen, Deborah. 2004. "Collective Intentionality." *Internet Encyclopedia of Philosophy*, https://www.iep.utm.edu/coll-int/.

—. 2014. "A Dynamic Theory of Shared Intention and the Phenomenology of Joint Action." In *From Individual to Collective* Intentionality, eds. Sarah R. Chant, Frank Hindriks, and Gerhard Preyer (New York: Oxford University Press), pp. 13-33.

Tuomela, Raimo and Kaarlo Miller. 1988. "We-Intentions." *An International Journal for Philosophy in the Analytic Tradition* 53, no. 3 (May): 367-389.

Turkle, Sherry. 2011. *Alone together: Why we expect more from technology and less from each other.* New York: Basic Books.

World Commission on the Ethics of Scientific Knowledge and Technologies (COMEST), Unesco. 2017. *Report on Robotics Ethics.*