to appear in: R. Hakli & J. Seibt (ed.), *Sociality and normativity in human-robot interaction*, Berlin, New York: Springer.

The Problem of Understanding Social Norms and What it Would Take for Robots to Solve it

Johannes L. Brandl & Frank Esken[1]
University of Salzburg, Institute of Philosophy, Austria

Abstract. In this paper, we argue that there is no evidence in sight warranting the conclusion that robots are social agents in some strong sense that requires an understanding of social norms. In support of this skepticism, we first consider an argument to the effect that a basic sensitivity to norms requires no mindreading abilities and may therefore also be found in non-human animals. In rebutting this view, we rely on arguments based on Searle's theory of institutional facts and Tomasello's theory of collective intentionality. We, then, extend these arguments to the case of robots and argue that robots' social behaviour does not reach the level at which an understanding of norms becomes crucial.

Acknowledgements We are grateful for very helpful and substantial comments by Johanna Seibt, Raul Hakli and two anonymous reviewers. Research for this paper and the cooperation of authors was financially supported by Austrian Science Fund Project I638-G15, ‘‘Rule-understanding, shared intentionality, and the evaluation by others,’’ as part of the ESF EUROCORES Programme EuroUnderstanding initiative.

Robots can do many things that are best described in social terms: they engage with others (i.e. with other robots or with natural beings like us) in various forms of joint action, often cooperating with others in solving problems, and they can guide and teach us humans much like a human instructor might do. Should we therefore treat robots as genuinely social agents? What matters here most is the question whether robots can understand social norms. In addressing this question, we propose to take an indirect route. We will first consider an argument for the claim that

[1] This paper is thoroughly co-authored. The order of the authors is purely alphabetical.

some non-human animals – notably chimpanzees – possess a primitive sense of social normativity (section 1). We then confront this claim with two prominent objections: one objection is based on Searle's theory of social facts, the other on Tomasello's claim that only humans possess the kind of collective intentionality that is necessary for understanding social norms (section 2). We then want to show why these considerations do not exhaust the problem of understanding social norms. As recent work on children's protest behavior has revealed, doubts remain whether the conditions specified by Searle and Tomasello are sufficient for a genuinely normative understanding of conventional rules (section 3). These discussions lead us to distinguish three stages of social intelligence and argue that genuine understanding of social norms occurs only on the third stage (section 4). Finally, we suggest that even on a generous interpretation of the available evidence about robots' social behavior, they reach only something in between the first and the second stage.

1. The case of animals

Not only humans, but many non-human animals are highly social creatures. From this fact alone, one might infer that understanding social norms is not a human-specific capacity. However, ascribing such an understanding to animals has turned out to be highly controversial. A prominent recent advocate of the claim that non-human animals have a basic sense of social normativity has been Kristin Andrews (see Andrews 2009 and 2013). We want to take her view therefore as a starting point for our discussion.

The case that Andrews builds for ascribing a primitive sense of social norms to animals is partly conceptual and partly empirical. On the one hand, Andrews relies on a conception of primitive normativity that takes its inspiration from Kant and Wittgenstein. According to this conception, we possess a primitive sense of normativity that requires neither an understanding of reasons for action nor an understanding of explicitly stated rules (see Ginsborg 2011). For Andrews, this suggests that one can have such a primitive sense of normativity also without yet possessing a theory of mind, i.e. without possessing the ability to attribute to oneself or to others explicit knowledge that one is following a rule or a norm. This is what opens up the door for ascribing a sense of normativity to animals that may qualify as an implicit form of understanding norms, in contrast to a reflective, explicit understanding

that is bound to language and to mindreading abilities (cf. Andrews 2009 and 2013).

Andrews appeals here to a venerable tradition in philosophy in support of her distinction between implicit and explicit understanding of norms. However, it is not philosophical reasoning *per se* that does the real work for her. Such a distinction remains empty as long as there are no criteria given for determining what kinds of behavior count as an implicit understanding of norms, with the additional requirement that such behavior cannot be explained in non-normative terms. Let us therefore turn now to the empirical evidence for ascribing such an implicit understanding of normativity to non-human animals. Andrews thinks that there is ample evidence, notably in the social behavior of chimpanzees. Here is one telling example that she cites as evidence for a norm violation:

> "Chimpanzee infants are given deferential treatment by other community members. Newborn chimpanzees are extremely interesting to other community members, and adults will watch intently but not try to approach the new member of their group. Juveniles and older infants, however, act on their interest by trying to approach or touch the new infant, which leads the mother to act hostilely or defensively; these young chimpanzees quickly learn not to approach newborn infants. Once infants are old enough to venture away from their mother, adults are extremely tolerant of infants climbing over them and even stealing their food or tools. Adults are also reported to self-handicap when playing with infants. A violation of the norms regarding infants, such as infanticide or other other aggressive acts towards infants, often results in an uproar of vocal protest from adult females, and can also cause third party intervention." (Andrews 2013, 194)

We do not deny that examples like this create a strong impression that chimpanzees possess something like a moral sense or at least understand something about social norms. But the mere fact that chimpanzees show emotions that have moral relevance *from our point of view*, does not necessarily mean that they have this relevance also from the apes' point of view. Andrews recognizes that there is still a line one can draw here. She therefore speaks cautiously only of "proto-social norms" manifest in behavioral regularities that "when violated, elicit bystanders reactions, including punishment and the expression of moral emotions" (Andrews 2013, 194; von Rohr, Burkart & Schaik 2011). Andrews argues that apes possess an implicit understanding of social norms and moral constraints in the sense that they track behavioral regularities in their society, which from a human point of view could be called social norms. However, she is less cautious when she claims that this tracking, together with a

motivation to follow the regularities they observe in their conspecifics, results "in a moral status that is not unlike that of many humans." (Andrews ibid., 194).

So the question remains: Are these proto-social norms really norms or only proto-norms, i.e. social regulations that precede the formation of genuinely social norms?

2. Two objections

In this section, we want to consider two lines of argument in support of a skeptical response to Andrews' view. One of them is based on the work of John Searle on social reality, the other can be found in Michael Tomasello's work on the origins of human thinking. Of course, we cannot provide here a comprehensive summary of their complex theories. We must confine ourselves to a raw sketch of the positions advocated by Searle and Tomasello.

A basic point in Searle's theory is that human society is a highly 'institutionalized' construct. As we understand his view, this makes human society fundamentally different from even the most sophisticated animal societies. Although this may sound like a strong metaphysical claim, Searle takes it as an obvious fac that does not commit him to an unbridgable gap between our species and the animal kingdom. When he contrasts humans and other animals, Searle says, "the point is not to make a plea for the superiority of our our species". (Searle 2010, p.7 fn.) Nothing in his theory of social facts rules out the possibility that some non-human creature might one day show all the features of a human society, including "income tax, presidential elections, divorce courts, and other institutional facts" (ibid.). However, it seems hard to deny that humans are the only ones to have created such facts to date. Therefore, Searle does not hesitate to describe the phenomena he is interested in, as "distinctive human phenomena" (ibid.).

Of course, it is no accident that as far as we know humans are the only ones to have created institutional facts. The reason for this can be found in the different functions that humans impose on objects, in contrast to the functions imposed by animals. Searle writes:

"[Non-human] animals can impose functions on natural phenomena. Consider, for example, the primates that use a stick as a tool to get bananas that are out of reach. And

some primates have even developed traditions of agentive functions that are transmitted from one generation to the next. Thus, most famously Imo, a Japanese macaque, used water to get the sand off her potatoes […]. Thanks to Imo, today potato-washing in salt water is an established tradition which infants learn from their mother as a natural adjunct of eating potatoes. […]. But the truly radical break with other forms of life comes when humans, through collective intentionality, impose functions on phenomena where the function cannot be achieved solely in virtue of physics and chemistry, but requires continued human cooperation in the specific forms of recognition, acceptance, and acknowledgment of a new *status* to which a *function* is assigned." (Searle 1995, p. 40)

How should we take this claim about a "radical break" in forms of life? One way to understand Searle's view here would go like this: Social facts are facts that result from the assignment of functions, and such assignments can be grounded in forms of collective intentionality that are part of animal life. Therefore, Searle is open to the view that social norms are not necessarily tied to institutional facts. If Imo learns from his mother how to properly wash the potatoes, he has learned a social norm without having learned an institutional fact.[2]

On this interpretation, the "radical break" would not seem to be very radical at all. First of all, it has not been shown that there can be social norms only in a society with an 'institutionalized' structure as Searle describes it. For instance, it is a social norm in our society to cover one's mouth when one is yawning. This norm does not constitute a new type of behavior, it merely regulates a behavior that we share, for instance, with chimpanzees. While it is unlikely that chimpanzees will want to cover their mouths too when they are yawning, there is nothing in Searle's theory that would rule out this possibility. And so the question remains why chimpanzees could not also adopt this rule as a social norm.

There is also a second consideration that would seem to weaken Searle's line of argument. Suppose one could show that institutional facts, as Searle describes them, can form a basis for social norms. Why could there not be something very similar in a complex society of animals? Many animals are capable of extended cooperative behavior. Why should this not allow them to assign a "value" to objects by following socially determined norm? The prohibition of infanticide that Andrews mentions seems to be just a case like that. The rule against infanticide ascribes a special value to the infants and those who do not recognize this value are punished. Perhaps this is not a "status function" in the

[2] Thanks to our referees for pressing us on this point.

technical sense in which Searle uses the term. It is nevertheless a function that has no physical, but a social foundation. Why could such functional ascriptions not serve as a basis for social norms?

We are therefore willing to admit that Searle's theory of status functions does not provide a complete argument for claiming that social norms are a distinctively human phenomenon because they require an assignment of *status* functions in contrast to an assignment of functions based on their physical properties. Yet, we believe that Searle's theory of institutional facts provides a good starting point for developing such an argument. Just think of how easy it is for Searle to explain why only humans have invented money as a social good. It is Searle's prime example of a functional ascription that requires continued cooperation. When humans invented money, they did so by cooperatively creating a status function of the form "X counts as Y in context C". In this way, a certain physical object X (for example a piece of paper) acquires a new status (for example the status of money) to which a function is attached by way of collective intentionality. Nothing like this, Searle claims, can be found in the animal kingdom. For animals, no object can acquire a status that goes beyond the brute physical functions that they can assign to physical objects. (p. 40)

We will come back to Searle's view at the end of this paper. Let us now consider a different line of argument against the view that some non-human animals exhibit a primitive form of social normative behavior. Tomasello rejects this view on the ground that it confuses principles of instrumental rationality with principles of genuine normativity:

> "Great apes may experience "instrumental pressure", for example, when they have a goal to eat food and they know that food is available at location X; this implies that they "must" go to location X. But this is just the way control systems with individual intentionality work: a mismatch between goal and perceived reality motivates action. In contrast, early humans began to self-monitor from the perspective of others and, indeed, self-regulated their behavioral decisions with others' evaluations in mind." (Tomasello 2014, 74-75).

Let us bracket for a moment the question what Tomasello means by "self-monitoring from the perspective of others". The claim is that chimpanzees, when they show an apparently moral behavior like in Andrews' example, only experience "instrumental pressure". They follow principles of *individual* practical rationality like "If I want to get food, I 'must' go to location X' or 'If I don't want to get in trouble with

the mother of this infant, I 'should' be tolerant but not aggressive against it'. What chimps are not able to understand, according to Tomasello's claim, are social principles like 'I 'should' behave in this but not in the other way, because this is *what others expect from me*'.

The question now is: How can we know that animals do not act in order to satisfy the expectations of others? This is the crucial step in Tomasello's argument. We would therefore like to quote him at length on this point:

"Finally, with respect to self-monitoring, the key is that being able to operate in this way communicatively requires individuals to self-monitor in a new way. As opposed to apes' cognitive self-monitoring, this new way was social. Specifically, as an individual was communicating with another, he was simultaneously imagining himself in the role of the recipient attempting to comprehend him (Mead, 1934). And so was born a new kind of self-monitoring in which communicators simulated the perspective of the recipient as a kind of check on whether the communicative act was well formulated and so was likely to be understood. This is not totally unlike the concern for self-image characteristic of early humans in which individuals simulate how they are being judged by others for their cooperativeness – it is just that in this case what is being evaluated is comprehensibility. Importantly, both of these kinds of self-monitoring are "normative" in a second-personal way: the agent is evaluating his or her own behavior from the perspective of how other social agents will evaluate it. […] This social self-monitoring for intelligibility in cooperative communication lays the foundation for modern human norms of social rationality." (Tomasello ibid., 58)

Can we conclude from this reasoning that animals do not understand social norms? It seems that Tomasello's argument warrants this conclusion if one accepts that a sharp distinction can be made between *individual* and *social* principles of rationality. But there are reasons speaking against such a clear dividing line, and this opens a way for defending Andrews' position. Take again the case of potato-washing mentioned in the quote from Searle earlier. If a chimp follows the principle "If I want to eat this dirty potato, I should wash it first", does he thereby follow a principle of individual or of social rationality? Formally, it is a principle of what he should do that has no implications about what others should do. But it is certainly important that this rule has been socially learned. And by "socially learned", we do not mean that one animal hands on knowledge to another one. It is a form of public learning. Whoever watches may pick up this rule and begin to follow it. So, it could be that a higher ranked animal in the group observes who is picking up the rule and who does not. This observation

could have consequences for the social status of those who adopt the practice. They may be considered to be smarter and better partners.

Tomasello might still argue that this is not sufficient for reaching the level of 'self-monitoring from the point of others'. Yet, it shows that there are quite sophisticated forms of social intelligence that do not require such a form of socially induced self-monitoring. And so the question remains, whether these forms of social behavior might provide the basis for a primitive form of normative understanding.

3. The case of children

So far, we have argued that the conception of a primitive sense of normativity that Andrews advocates and that includes an implicit understanding of social norms may be defensible. An advocate of this conception can counter the objections of Searle and Tomasello. We now want to argue, however, that there are still other problems with this conception. In order to explain what these remaining problems are, we turn now to recent studies on the early understanding of normativity in human children. These studies have been partly inspired by Searle's and Tomasello's work, but they take us a significant step further, as we shall try to show now. [3]

A growing body of developmental studies suggests that even toddlers possess a normative awareness that manifests itself in different settings, notably in conventional and in pretend games, with various competences, for example, dealing with property rights, with artefact functions, or with various kinds of entitlements. The main evidence in these studies is that children show signs of protest, presumably in reaction to the violation of a rule.

In a seminal paper, Rakoczy et al. (2008) proposed a new way of testing children's early understanding of norms. A basic idea underlying their approach is that norm violations may be perceived as a reason for protest, irrespective of what kind of norm is at stake. Hence, protesting could be a reliable indicator of normative understanding also outside a moral context.

[3] The following sections are based on previously published work in cooperation with Beate Priewasser and Eva Rafetseder. (see Brandl, Esken, Priewasser & Rafetseder 2015).

In order to measure this basis awareness of normativity, Rakoczy et al. observe children's reactions to a puppet's violation of a constitutive rule in a conventional game. The children tested are between 2 and 3 years old. They are first familiarized with two novel actions, for example:

- In the model phase an adult shows 2- and 3-year-old children new game actions (X counts as Y in C). The adult performs actions A1 and A2. A1 is marked as "daxing", A2 as an accidental mistake
- In the action phase it is the child's turn to play the game of daxing, and to learn how to dax
- In the test phase, a third person (a puppet) enters and announces: "I'm gonna dax now!"
- In the target condition, the puppet performs an action which is mistaken, given the structure of the game
- Children's responses to such mistaken actions, in particular protest and correction, are taken as indicators of their awareness of the rule structure of the game

According to Rakoczy et al., the 3-year-olds saw the puppet's actions as not conforming to the social norm of daxing, and enforced the norm. Hence, the experiments are taken to show that 3-year-olds understand social norms. "These studies demonstrate in a particularly strong way that even very young children have some grasp of the normative structure of conventional activities." (Rakoczy et al. 2008) This conclusion is based on the background assumption that social norms have a foundation in the assignment of status functions. Therefore it is crucial that children are informed about a new game in which familiar objects have to be used in accordance with arbitrary rules. In this way, Rakoczy et al. think these objects acquire a status function and that it is therefore a normative requirement that one should use these objects only in the way prescribed by the rules of 'daxing'.

Could one explain children's protest also as a response to social pressure? Do children experience in these experiments such a pressure or only the pressure of individual rationality? Let us quote Tomasello once more in order to illustrate the question we are asking here:

"Young human children are concerned with the social evaluation of others from preschool years on as they attempt to actively manage the impression they are making on them. […]From the point of view of normativity, this meant that in making their behavioral decisions, humans not only experienced individual instrumental pressure but also experienced second-personal social pressure from their partners in social

engagements. This constitutes one origin of what later become social norms of morality." (ibid., 75)

In the daxing-case, the child itself perfectly knows what it is supposed to do when it is her turn to "dax". It is therefore not the child that is under any kind of "pressure". Rather, it is the puppet that violates the daxing-rule that is put under pressure by the child. The role of the child is not to *experience*, but to let others experience a discomfort that is meant to change their behavior. We therefore need to put our question this way: Do children in these studies use their power as social agents to exert a "second personal pressure" on the puppet when it breaks the rules, or do they use their pressure merely to express their concern about a violation of some principle of individual rationality? Or do we have here another borderline case?

While we do not want to deny that the reported reactions may be indicative of a general awareness that something has gone 'wrong', or that something 'wrong' has been done, it would be premature to call these responses a 'normative protest' merely for this reason. As soon as there is any evidence that a young child, an animal or an artificial system learns to distinguish between 'right' and 'wrong', we may attribute to it a basic awareness of normativity. One could define normativity simply as a standard that allows us to distinguish between 'right' and 'wrong', correct/incorrect, pleasant/unpleasant, expected/unexpected, or whatever. However, such categorizations can be understood like any other classification as purely factual. Simply calling something 'right' or 'wrong' does not yet make it right or wrong in a normative sense. Therefore, children might merely follow a certain pattern or regularity, when they first learn to use such terms, without grasping the deontology that we associate with them.

Consider a baby that reliably shows signs of protest in various circumstances, e.g. when her bottle is too hot, when the lid of the bottle is blocked, when the bottle does not taste sweet enough, etc. In this way even a baby can distinguish between 'right' and 'wrong' and would therefore manifest some normative understanding if we use the term 'normative' without any restrictions. But babies do not show the right kind of protest when we present them with rule-violations in conventional games like the 'daxing'-game.

What we conclude from this example is that there are two very different

kinds of protest:

(a) protesting against some condition that is perceived as wrong, as unpleasent, or as unexpected, etc.

(b) protesting against rule-violations in conventional games.

The baby mentioned above is capable of expressing protest only in sense (a), while the protest of children in the 'daxing'-game is of the quite different kind (b). While we agree with Rakoczy et al. on this point, it still needs to be shown whether this suffices for demonstrating a form of *normative* protest.

4. Three steps towards normative understanding

The protest behavior of young children raises a similar problem of indeterminacy that we already encountered in the discussion of Andrews' notion of primitive normativity. This leads us back to the fundamental question what it means to understand social norms. In the case of Andrews' examples of primitive normativity, we have seen that the behavior of chimpanzees can be interpreted as following either individual or social principles of rationality. Likewise, the protesting studies with 2- and 3-year old children leave it open whether children at this age really protest against the violation of social norms or some principle of individual rationality. No appeal to an understanding of social norms is necessary, for example, when a child learns from his caregiver how to open a water-tap: "You have to do it in this way, not that way". If the protesting behavior can be explained in the same way, then their protest does not indicate anything about their understanding of social norms.

We now start a third and last attempt to clarify what a basic understanding of social norms could mean. Let us therefore go back once more to Tomasello's idea to ground such an understanding in what he calls the ability to "self-monitor from the perspective of others". Without any doubt, the 2- and 3-year old children in the "daxing"-games are acquainted with the experience of social pressure. They may already experience a basic form of shame when they engage in social self-monitoring. Why then is it that it is still unclear, as we tried to show,

whether they protest against the violation of a social norm? Something needs to be added here to the explanation that Tomasello offers. We now want to suggest that for answering this question it is critical to consider the effect that the authority of the experimenter has on children's response.

The authority of the experimenter derives from the fact that he tells the child (and the puppet) how to play the game of daxing. So far, it is his authority that decides which moves in the game are correct (appropriate) or incorrect (inappropriate). While this requires a social interaction between the one who introduces the game and those who are supposed to follow its rules, it is not yet a form of interaction that requires an understanding of social norms. There is social pressure due to the instructions handed out by someone who has the authority to do so, and it therefore seems that Tomasello's condition for normative understanding is fulfilled. But that is not necessarily the case since social pressure does not exclusively arise when social norms are operative. Authorities are an independent source of social pressure, even when they do not act according to social norms.

In order to make sure that a child is sensitive to norms, we have to contrast the demands of an authority with what a social norm requests.

The Ronnie situation

Consider a child, Ronnie, that cries when she is put to bed and the lights are turned off. Suppose that Ronnie is old enough to notice a watch at the wall and that she observes that it is not yet 8 o'clock. When her Mum appears, presumably to turn off the lights, Ronnie does not immediately protest. She turns to the watch – perhaps pointing at it – to pass on her observation that the time to go to sleep has not yet come. But when her Mum ignores this and turns off the light anyway, Ronnie starts to protest heavily.

This is a case, we believe, that helps us to see how normative understanding can manifest itself in protest behavior. The fact that Ronnie looks up to the watch when her Mum comes in, gives us good reason to say that she understands the normative force of the rule "When it is 8 o'clock, it is time to go to bed". It does not matter that Ronnie might also protest when her Mum turns off the lights in accordance with the rule. It is her observing the watch that indicates her knowledge about

when the lights *should* be turned off.

At first, the Ronnie-example may seem to be a simple case in which merely regulative rules are operative. We want to argue, however, that even in this case there is room for applying Searle's famous distinction between regulative and constitutive rules (see Searle 1969). As defined by Searle, regulative rules "regulate antecedently or independently existing forms of behavior" (1969, p.33). In our example, the rule to brush one's teeth before going to bed would be an example of a regulative rule, because the brushing can be performed independently of the rule. But what about the rule "go to bed at 8"? Surely, one can go to bed at other times, and thus perform the action antecedently and independently of the given rule. And one may do other things at 8, instead of going to bed, just as one can go to bed without brushing one's teeth. If one takes the rule merely in this regulative sense, however, it would not lead to a clear manifestation of normative understanding. No such understanding is needed to behave according to the instructions of adults requesting that such rules are obeyed.

In our example, Ronnie seems to exhibit a normative understanding, however. He does this by insisting that he does not have to go to bed *now*, but only at 8, because that is what the rule says. The action he wants to perform is "going to bed at the right time", and *this* action does not exist independently of a rule that defines 8 o'clock as the right time to go to bed. For this reason, we think that the rule here counts as a constitutive rule, like the daxing rule or other conventional game rules. It is a constitutive rule because it generates a new social fact. It is not a naturally given fact that 8 o'clock is the right time to go to bed. (Without a rule that establishes this, 8 o'clock is neither the right, nor the wrong time to go to bed). If children know that these rules are not freely (idiosyncratically) invented by an adult, they recognize an authority-independent ("objective") fact that generates a norm: 8 o'clock counts as the time to go to bed in the context of the evening routines in Ronnie's family.

At this point we agree with Searle's claims regarding the irreducibility of deontic phenomena (Searle 1995, p. 70). This does not mean to deny, however, that an understanding of normative facts builds on children's experience with social pressure from authorities. From very early on in their lives, children are confronted with authorities that tell them what to do or not to do: drink this, do not touch that, go to sleep, etc. They come

to understand such commands and know which circumstances prompt authorities to express them. At some point, they will find out that there are constraints for these commands to *be justified*. Even authorities cannot issue commands simply at their own will. There are rules that *entitle* them to do so depending on the circumstances. If it is 8 o'clock, Mum is entitled to turn off the lights. Once children know that there are these constraints even for authorities, they learn to understand something about normative constraints.

Taking all this into account, we can now differentiate between three stages in social cognition leading up to an understanding of social norms. In presenting our model here, we are aware that much more needs to be said in order to flesh out this model in detail. Also, it would be interesting to compare our model with similar models that have been proposed in the literature on moral development, including the classical. theory of stages of moral development by Laurence Kohlberg (cf. Kohlberg 1981 and 1982). Although Kohlberg's model targets a cognitive development in much older children, beginning at the age of 9, we notice certain parallels between how Kohlberg describes the early stages of this development and our model. Thus, when Kohlberg claims that, at a pre-conventional stage, children initially operate with a sense of right and wrong that is based on what feels good or bad, this description fits with our observation that social creatures develop a sensitivity for norms even before they are able to grasp the normative and/or conventional character of social rules.[4]

A three stage model of normative understanding

Stage 1: Social creatures adapt to regularities that have a social foundation.

At this stage we find, for example, chimpanzees that obey a rule against infanticide. For Andrews this is already a form of implicit understanding of social norms. However, the social foundation of this behavior only consists in the fact that chimpanzees learn from others and create the

[4] Among psychologists, it is not completely unusual to apply Kohlberg's ideas in devising a stage model of moral development, beginning already in infancy. See for instance the entry on "Moral Development" in the online Psychology Encyclopedia: http://psychology.jrank.org/pages/431/Moral-Development.html">Moral Development - STAGES OF MORAL DEVELOPMENT.

rules as a group. Following Tomasello's argument, this kind of behavior requires only individual and no social pressure to adapt to an existing social order. The apes need not understand what others expect from them, they may be driven only by instrumental rationality.

Stage 2: Social creatures make things happen that *should be* the case, or prevent things from happening that *should not be* the case (according to what authorities demand).

Children reach this stage when they protest in the "daxing"-experiments. In contrast to stage 1, they are now aware that agents do not only pursue individual goals according to principles of instrumental rationality. They now become sensitive to social pressure and therefore begin to monitor their own behavior and the behavior of others in order to avoid such pressure. Rakoczy and Tomasello take this to imply a new form of *social practical rationality*. But for the reasons pointed out, their sensitivity to social pressure should not be confused with an understanding of social norms prescribing what one *should do.* (We say more about the distinction between what *should be* and what *should be done* below).

Stage 3: Social creatures recognize social rules and know that they apply in all situations of a certain kind (independently of what an authority dictates).

When children reach this third stage, they learn to understand rules *as* rules in an abstract sense. This means that they are able to grasp the rule as something that exists independently of a concrete situation and can therefore apply to a certain *kind* of situation. They are then able to act in such situations because *the rule* requires it, not because it would be unwise to act against an authority. That is, we submit, when children begin to understand the nature of social norms.

5. The case of robots

Let us now turn to the case of robots. How far can the above considerations help to sort out our conflicting intuitions about artificial agents?

Starting with stage 1, we want first to reach an agreement that robots are social agents at least in the sense that they can adapt to regularities with a social foundation. To illustrate this first point, consider the following

example of an adaptive social behavior of robots: There is a vast body of work on algorithms and control methods for groups of decentralized cooperating robots, called a "swarm" or "collective". These algorithms are generally meant to control collectives of hundreds or even thousands of robots. Each robot has the basic capabilities required for a swarm robot. For example, hundreds of small robots are moving on a table and create together the letter *A*. They are equipped with an algorithm to control each other: if one of them stops at the wrong position on the table, the others are able to inform the misaligned robot about its mistake.[5]

The controversial question is whether robots might also exhibit a form of social practical rationality that includes a sensitivity to social pressure (provided by authorities). One might argue that this would grant too much to an artificial agent on the ground that robots are not able to follow rules. However, there is a further distinction to be made here, since we must distinguish behavior at stage 2 from rule-following at stage 3. Following Wilfried Sellars, one might introduce at this point a sophisticated distinction between "ought to dos" and "ought to bes":

"Pattern-behavior of such and such a kind ought to be exhibited by trainees, hence we, the trainers, ought to do this and that, as likely to bring it about that it is exhibited." (Sellars 1974, 423).

The trainee here might be a dog, an infant, or a robot. As its trainers or educators, we set the rules how it ought to behave. But as long as the dog, the infant or the robot does not understand the normative force of our rule, we should not describe it as a stage 3 case of following a rule, abstracted from the demands of concrete authorities. According to Sellars, we operate with such an abstract notion of rule-following when we reason about what one ought *to do*. While the trainee at stage 2 may not be able to participate in such reasoning, he may nevertheless understand what ought *to be*, and in this sense he, she or it may be able to follow a given rule. Without making this distinction, we would also have to deny chimpanzees that they can follow social rules, and we take that to be an empirically established fact.

Doubts still remain whether robots actually can be social agents in the sense of stage 2. In this case they would have to engage in a form of

[5] Cf. http://www.eecs.harvard.edu/ssr/projects/progSA/kilobot.html

social self-monitoring that Tomasello describes. That means they would have to consciously experience that others expect from them to behave in certain ways and they would have to be able to evaluate their own behavior from a second person perspective. Even if there are forms of robot behavior that come close to such self-monitoring, it is still very much an open question whether one might not reduce their social behavior to stage 1. So far we do not know whether robots can have any conscious awareness of what should or should not be the case and what others expect from them.

An interesting case study here is provided by the robot 'Leonardo', built by Cynthia Breazeal and Brian Scassellati from the MIT Media Lab (cf. https://en.wikipedia.org/wiki/Leonardo_(robot).[6] Some researchers claim that Leonardo has a rudimentary theory of mind in the sense that it can model the beliefs and intentions of an interlocutor, including the interlocutor's beliefs about itself. But is such a description really warranted? Leonardo's motors, sensors, and cameras allow it to mimic human expressions, which helps humans react to the robot in a familiar way. Some authors describe it as mimicking human facial expressions, thereby distinguish between itself and others, and even to take the perspective of others. If this should mean that Leonardo understands something about the mental perspective of others and therefore possesses „mind-reading"-abilities, we strongly doubt this claim for the following reason. There is no evidence here that could not also be explained by saying that Leonardo keeps track of the interrelations between the facial expressions and the actions of an interlocutor, and in this way detects mismatches between them. It reads *behavior* but not *minds*.

At this point, the case of child development might be brought up as a further argument. Children experience social pressure from early on at an affective level, long before they develop secondary emotions like shame or pride. At this stage, social self-monitoring manifests itself as an awareness of being the center of attention of others. A child may feel uncomfortable (or secure) at an affective level without understanding what it is that makes her feel that way. When we try to transfer this non-conceptual consciousness of social relationships to the case of robots, we

[6] Thanks to Johanna Seibt for drawing our attention to this case. For more about Leonardo and other social robots, see http://robotic.media.mit.edu

face the problem how we should justify ascribing any form of consciousness to artificial systems. But even if we set aside this problem, there is reason to remain skeptical at this point. We also do not know what kind of evidence would be sufficient to find a connection between social self-monitoring and affective consciousness in non-human animals. Yet, if we granted robots the ability to experience social pressure, we would have to grant this also to dogs and to apes that have been enculturated for several generations. Given the difficulties in answering this question in the case of animals, it is unlikely that research on artificial systems will be able to provide this kind of evidence in the near future.

Finally, we see no way how a swarm of robots would ever meet the conditions necessary for reaching stage 3. Not only are members of such a swarm probably not experiencing any social pressure (stage 2), they are hardly in a position to understand rules *as* rules, i.e. as abstract principles (in contrast to observable processes and abstractions from them) that possess an independent ("objective") normative force. We doubt that robots will ever be able to fulfill this condition.

Our conclusion of this paper is therefore a skeptical one. There is no evidence in sight warranting the claim that robots are social agents in some strong sense that requires an understanding of social norms. Despite this negative conclusion, our investigation also contains a positive suggestion how to make sense of the social behavior of robots. In our view, we may conceive of it in the same way in which we conceive of our social interaction with nonhuman animals or with infants when we train or educate them. In this case, we set the rules and bring it about that other agents conform to them. We enable them to adapt their behavior to our rules even before they can understand social norms.

References

[1] Andrews, K. (2009). Understanding norms without a theory of mind. *Inquiry*, Vol. 52, No. 5, 433-448.

[2] Andrews, K. (2013). Ape autonomy? Social norms in other species. In: *Philosophical perspectives on animals: Mind, ethics, morals*, edited by K. Petrus and M. Wild, Transcript, 2013, 173-196

[3] Brandl, J., Esken, F., Priewasser, B. & Rafetseder, E. (2015). Young children's protest: What it can (not) tell us about early normative understanding. In: *The*

roots of normativity, edited by G. Satne, Special Issue, *Phenomenology and the Cognitive Sciences*, Berlin, New York: Springer, 719-740.
[4] Daeg de Mott, Daena K.: Moral Development, Psychology Encyclopedia (retrieved, July 29, 2016): http://psychology.jrank.org/pages/431/Moral-Development.html">Moral Development - STAGES OF MORAL DEVELOPMENT.
[5] Ginsborg, H. (2011). Primitive normativity and skepticism about rules. *The Journal of Philosophy*, Vol. CVIII, No. 5, 227-254.
[6] Mead, G. H. (1934). *Mind, self, and society*. Chicago: Chicago University Press.
[7] Kohlberg, L. (1981). Essays on moral development, Vol. I: The philosophy of moral development. San Francisco, CA: Harper & Row.
[8] Kohlberg, L. (1982). Moral development. In: *The cognitive developmental psychology of James Mark Baldwin: Current theory and research in genetic epistemology*, edited by J.M. Broughton & D.J. Freeman-Moir, NJ: Ablex Publishing Corp.
[9] Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology, 44*(3), 875-881.
[10] Rakoczy, H., Brosche, N., Warneken, F., & Tomasello, M. (2009). Young children's understanding of the context-relativity of normative rules in conventional games. *British Journal of Developmental Psychology*, *27*(2), 445-459.
[11] Searle, J.R. (1969). *Speech acts. An essay in the philosophy of language*, Cambridge: Cambridge University Press.
[12] Searle, J.R. (1995). *The construction of social reality*. New York: The Free Press.
[13] Searle, J.R. (2010). *Making the social world. The structure of human civilization*. Oxford: Oxford University Press.
[14] Sellars, W. (1974). Meaning as functional classification. *Synthese* Vol. 27, 417–37.
[15] Tomasello, M. (2014). *A natural history of human thinking*. Cambridge, MA: Harvard University Press.
[16] Von Rohr, C., Burkart, J. & van Schaik, C. (2011). Evolutionary precursors of social norms in chimpanzees: A new approach. *Biology and Philosophy*, No. 26, 1-30.