This is the accepted version of the following article: Christian Balkenius, Lola Cañamero, Philip Pärnamets, Birger Johansson, Martin V Butz, and Andreas Olson, 'Outline of a sensory-motor perspective on intrinsically moral agents', *Adaptive Behaviour,* Vol 24(5): 306-319, October 2016, *which* has been published in final form at DOI: https://doi.org/10.1177/1059712316667203

Published by SAGE

Journal Title
XX(X):1–15
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Outline of a Sensory-Motor Perspective on Intrinsically Moral Agents

Christian Balkenius[1], Lola Cañamero[2], Philip Pärnamets[3], Birger Johansson[1], Martin V. Butz[4], Andreas Olsson[3]

Abstract
We propose that moral behavior of artificial agents could (and should) be intrinsically grounded in their own sensory-motor experiences. Such an ability depends critically on seven types of competences. First, intrinsic morality should be grounded in the internal values of the robot arising from its physiology and embodiment. Second, the moral principles of robots should develop through their interactions with the environment and with other agents. Third, we claim that the dynamics of moral (or social) emotions closely follows that of other non-social emotions used in valuation and decision making. Fourth, we explain how moral emotions can be learned from the observation of others. Fifth, we argue that to assess social interaction, a robot should be able to learn about and understand responsibility and causation. Sixth, we explain how mechanisms that can learn the consequences of actions are necessary for a robot to make moral decisions. Seventh, we describe how the moral evaluation mechanisms outlined can be extended to situations where a robot should understand the goals of others. Finally, we argue that these competences lay the foundation for robots that can feel guilt, shame and pride, that have compassion, and that know how to assign responsibility and blame.

Keywords
autonomous robots, embodied emotions, sensory-motor grounding, embodied interaction, empathy, intrinsic morality

1 Introduction

With the approaching introduction of autonomous robots into society, it is time to take potential risks seriously. The perceived threat from artificial intelligence that is currently in the public eye may certainly be exaggerated, but as robots are increasingly used in areas such as domestic, healthcare, or military settings, safety measures need to be put in place to ensure that robots are not dangerous to us, and that they know when they do something wrong.

One solution often suggested is something akin to Asimov's robot laws:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Although such rules make for good fiction, they are very problematic as a basis for ethical robots since they require that the robot has a full understanding of the rules, their consequences and perfect reasoning skills. Furthermore, this solution depends on an accurate perception of the current situation at all times. These underlying assumptions are not only well beyond the capabilities of present-day robots, but they are also open to numerous flaws due to their generality and abstract nature (Anderson, 2008; Murphy & Woods, 2009; Norman, 2005; Sloman, 2006).

The robotics community has been concerned about ethics for a number of years, with numerous initiatives and events organized around the world under the term "Roboethics" (cf. Anderson & Anderson, 2007). Such concerns can be grouped into two main strands: the design of robots that are respectful of and safe for humans in their interactions, and the concern for robots rights (cf. Sloman, 2006). These

[1]Lund University Cognitive Science, Sweden [2]School of Computer Science, University of Hertfordshire, UK [3]Department of Clinical Neuroscience, Karolinska Institutet, Sweden [4]Cognitive Modeling, University of Tübingen, Germany

Corresponding author:
Christian Balkenius, Lund University Cognitive Science, Box 192, 221 00 Lund, Sweden
Email: christian.balkenius@lucs.lu.se

initiatives take, in one or another way, an approach similar to Asimov's in the sense that they constitute attempts to come up with externally-given rules to constrain the behavior of robots and their interactions with humans. They are also typically characterized by attempts to ground the robot's ethics in reasoning capabilities often with a tutoring or advisory role imagined for the future ethical robot (e.g., Anderson, 2008; McLaren, 2006).

In contrast, we propose that intrinsically moral robots can be designed based on development and learning from bodily ("physiological") grounding and sensory-motor principles, such that full autonomy of the robot can be preserved, and that more advanced capabilities based on the ones outlined in this paper can subsequently be scaffolded. Such robots will be intrinsically moral in two senses: first, being concerned with, and capable of, distinguishing autonomously between "right" and "wrong"; second, learning "right" and "wrong" through interactions with other agents and by "empathizing" with those agents. Being grounded in the robot's "physiology" and more generally embodiment (Cañamero 1997, 2001, 2003) and sensory-motor principles (Pezzulo, 2011) implies that their morality will be grounded in the perceptual, value and motor systems of the robot itself, including values and representations internalized through interactions with others, and can be developed using subsystems modeled after (and meaningful to) their human counterparts. This includes direct visual and interoceptive perception of causal relations, agency and harm, as well as relevant motivational and emotional systems, together with causal reasoning mechanisms and social learning. Our approach thus puts social emotions at the heart of moral behavior, and in a fundamental way brings together embodied sensory-motor cognition, internal and internalized value systems, internal representations of self and others, bodily, "kinesthetic" judgments, and capabilities for self-perception. (Panksepp, 1998; Laird, 2007; Damasio, 1999, 2010; Solomon, 2007; Colombetti, 2014).

The rest of this paper is organized as follows. After framing our approach in the context of a triadic interaction model (Section 2), we propose to design agents that learn from their own experiences to act morally, based critically on seven types of competences. First, intrinsic morality must be grounded in the internal values of the robot arising from its physiology and embodiment (Section 3). Second, the moral principles of robots must develop through their experiences of interactions with the environment and with other agents—humans and robots (Section 4). Third, it is necessary that the robot is sensitive to social emotions. This includes using observed emotional reactions—including (facial, bodily) expressions—of others, both as reinforcing stimuli and for use in higher level decision making (Section 5). A sensitivity to social emotions depends both on the perceptual recognition problem and the existence of the appropriate learning mechanisms. We describe that the dynamics of the social emotions closely parallels that of other non-social emotional states such as hope and fear, frustration and relief. Fourth, the robot must also be able to learn from observation of others. This involves viewing interactions between other agents and the detection of their emotional reactions (Section 6). Fifth, a sensitivity to social emotions also implies an understanding of causation. We describe how a robot can infer causal relations by observing the dynamics of interaction between animate or inanimate objects (Section 7). The technical problem is here to recognize the dynamic interaction between objects or agents and to infer causal relations both at a basic dynamic level and at a more cognitive level. Sixth, the robot must learn to anticipate and reason about the consequences of actions (Section 8). Seventh, the robot must be able to infer the goals of others and know whether an interfering action will help or hinder. While the previous competences are at a more sensory motor level, this final level also requires generative models of other agents (Section 9). We argue that these competences lay the foundation for robots that can feel guilt, shame and pride, that have compassion, and know how to assign responsibility and blame (Section 10).

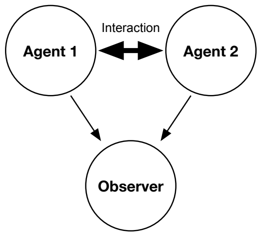

Figure 1. Triadic interaction between an observer and two other agents

2 Triadic interaction model

We propose that many questions of morality in robots can be addressed in a scenario with a triadic interaction between agents (humans or robots), where two agents interact and a third observes, learns from the two others, or potentially intervenes (Fig. 1). The first agent may behave aggressively toward the second or may help or hinder its actions. The observing agent will learn to anticipate the reactions of the second agent, internalize them, and use them in its own decision making: both when selecting its own actions and when it decides whether to intervene in the interaction between agent one and two.

Consider a simple example of a prototypical moral situation: the robot observes Agent 1 hitting Agent 2, causing harm to Agent 2, which is suitably expressed through, for example, a cry of pain or a hurtful facial expression. Our interest lies with what the robot now does. We propose that minimally the robot should feel an appropriate emotion (e.g., anger, compassion) as a result of interpreting the observed interaction in terms of its outcome (Agent 2 being hurt) and by assigning responsibility for that outcome (seeing that Agent 1 hit Agent 2). Taken together, these elements should motivate the robot to intervene appropriately in the situation by expressing its feelings and thus reproaching Agent 1, and, possibly, hindering Agent 1 from further hitting Agent 2. Hence, more abstractly, for the robot to behave morally, it needs to not only understand the goals of others and be able to detect others' emotional reactions, but it also needs a set of its own (internal) and acquired (internalized) values that ground its (moral) preferences, motivations, assessment of right or wrong and decisions for action. Further, this also depends on the representational and self- and other-perception capabilities of the observer agent that are involved in the consideration of others as being like me, in social emotions, and in moral behavior. In the following seven sections we develop a framework for moral robots based on these principles.

3 Embodiment of emotions: physiological grounding

To make robots *intrinsically* moral, the first step is to provide them with a basis to ground morality inherently, so that they can "judge" by themselves what is good or bad for them as well as for others. This means that the robots must have their own value system to base such "judgments" on, that will also allow them to interact with and learn about the physical and social world proactively and meaningfully (Cañamero, 1995; Pfeifer, 1996). Following an Embodied Cognitive Science and AI approach, we view embodiment as an essential element and determinant of cognition and action, as well as of emotion. In the context of this paper, this means that a value system that grounds morality *intrinsically* needs to be based in the embodiment of the robot in a fundamental way. Such bodily grounding provides not only the basis for a "core affect" (Damasio, 1999) system, but transpires through the entire "cognitive apparatus" of agents, biological or artificial, embedding us in a world of affective affordances (Colombetti, 2014) and giving us reasons to make sense of it and interact in it, not only as solitary individuals but fundamentally in our interactions with others, in what has been termed "participatory sense-making" (De Jaegher & Di Paolo, 2007). Embodiment is also at the core of moral emotions and their evaluative structure, rooting the evaluative emotional judgments that characterize them in a form comparable to kinesthetic judgments, not necessarily accessible to awareness and rational but rather tacit and unspoken (Solomon, 2007).

Although "embodiment" has different meanings when talking about "embodied agents" and "embodied cognition" (Ziemke, 2003), in this paper and building on a longstanding approach (Cañamero, 1997, 2001, 2003), the bodily grounding of moral values and emotions that we propose stems from the "physiology" of the robot and its control and interaction dynamics, in addition to (and coupled with) sensory-motor interaction. Such "physiological" modeling has greatly developed over the last two decades, and the term "internal robotics" (Parisi, 2004) was coined to emphasize the importance of modeling internal as well as external aspects of embodiment.

In our approach, the robot's physiology—consisting of essential variables and simulated hormones—and its dynamics is deeply intertwined with the perceptual, cognitive and motor capabilities of the robot (Cañamero, 1997; Avila-García & Cañamero, 2004; Cañamero & Avila-García, 2007) as well as its social interaction (Cañamero 2008) and provides mechanisms to endow the robot with the two key dimensions of emotions, namely arousal (Hiolle et al., 2014) and pleasure (Cañamero & Lewis, 2016 submitted). This modeling approach implies that the robot's intrinsic morality will be grounded in the perceptual, value and motor systems of the robot itself, including values and representations internalized through interactions with others, and can be developed using subsystems modeled after (and meaningful to) their human counterparts.

Such physiologically-based grounding of (moral) values can also drive and shape learning processes—not only the "what" of learning but also the "how" (Lowe, 2014). Of particular relevance to the framework that we propose here is its role in the learning of object and behavior affordances (Cos et al., 2010) and in reinforcement learning (Cos et al., 2013).

4 Development

A key aspect of intrinsically moral social robots is their ability to internalize the moral values, behaviors and social emotions of the humans they have to interact with. While different types of learning—both with and without explicit "teaching" or "reinforcing" signals on the part of the human—constitute important mechanisms towards this end, we argue for the need to adopt a developmental approach to make robots' morality intrinsic from the early stages of the interaction and learning process.

As argued elsewhere (Cañamero et al., 2006), a fuller and deeper integration of autonomous social robots into human environments requires their being embedded in the social environment in which they will fulfill their roles, in a way akin to how human children develop, although on a shorter time scale. The relatively recent field known as Developmental or Epigenetic Robotics (Zlatev & Balkenius, 2001) is an interdisciplinary area at the intersection of child development and robotics that endeavors both to take inspiration from human development to build better robots, and to use robots as models to help understand typical and atypical human development as well as tools in therapy of developmental disorders (Prince & Gogate, 2007). This field investigated the development of different types of skills, including sensory-motor, cognitive, affective, and social (for surveys see, e.g., Asada et al., 2009; Berthouze & Ziemke, 2003; Lungarella et al., 2003; Prince & Demiris, 2003). Grounding on internal value systems such as described in the previous section and social interaction (Pepperberg, 2001), the developmental processes modeled in this field can provide human-adapted mechanism for internalization, socialization and "enculturation" of moral values and the development of social and moral emotions through natural interaction with humans. Such processes include the notion of "ongoing emergence", defined as the continuous development and integration of new skills (Prince et al., 2005), as well as emotional development processes such as attachment (Cañamero et al., 2006), human-facilitated emotion regulation (Hiolle et al., 2014), and hormonally-modulated epigenetic development through sensory-motor interaxtion with humans (Lones et al., 2016). Such processed permit robots to develop different internal values, cognitive and affective profiles, and their external (e.g., behavioral, expressive, interactive) manifestations as a function of their different socially-driven developmental histories.

Robots with different developmental pathways and moral values would then be expected to behave differently when tested in our triadic interaction scenarios, permitting us to experimentally compare different moral principles.

5 Social emotions

We will ground our view on moral robots in a small set of emotions. These emotions will ground the robot's evaluations but also, as we discuss in the next section, provide a crucial interface to learning about others. Compared to the complexity of full human emotions, these emotions are simplified to the extent that they can be operationally defined and implemented in a robot with the perceptual abilities that are within the range of what is technically possible today. Our focus will be on the social emotions. To some extent, all emotions are social in the sense that they are accompanied by more or less visible expressions. However, some emotions have the additional quality that they are meaningless without the existence of a social context. These include negative self-directed emotions such as shame and embarrassment that involve violations of societal standards, as well as, pride which is in a sense its opposite. Although directed at the self, these emotions can be understood as a preparation for the expected reactions of others. While shame can be seen as an expectation of social blame or punishment, pride can be seen as an expectation of praise or other type of reward. Interestingly, these emotions can be elicited even without performing the action that caused the emotion. It is possible to feel ashamed or embarrassed without being guilty of the action that caused the emotion. Although social emotions may appear to need complex cognitive abilities, it has been suggested that emotions such as embarrassment could be the result of much simpler processes (Griffiths & Scarantino, 2009).

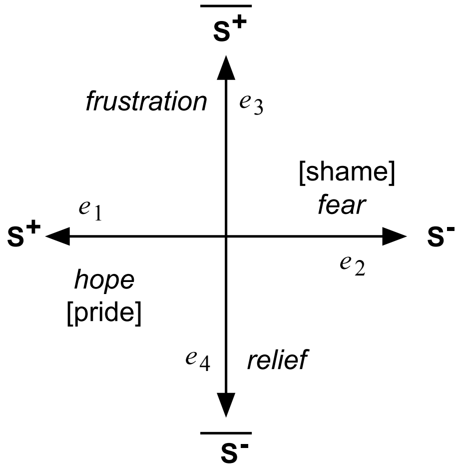

Figure 2. The Emotion Space. Every emotion is located in a four-dimensional space. Two of the dimensions code for positive and negative valence (hope and fear), while the two other code for unfulfilled expectations (frustration and relief). S^+ and S^- represent stimuli of positive or negative valence, and $\overline{S+}$ and $\overline{S+}$ represent omission of such stimuli. (Adapted from Rolls, 1990)

5.1 *The dynamics of social emotions*

The basis for our model of social emotions will be the four-dimensional emotional space[*] proposed by Rolls (1990). In this model, emotions can be categorized along four dimensions (Fig 2). The first two can be labeled "hope" and "fear" and correspond to the expectation of a positive and a negative event, respectively.

The second set of dimensions corresponds to "frustration" and "relief", that is, states caused by unfulfilled expectations. Relief is caused by a fearful event that did not happened and frustration is caused by a positive event that did not occur. These two dimensions are related to what

[*]Although one axis represents positive valence and the other negative, each axis should be considered two dimensions rather than one. To illustrate this, think that it is possible to expect something that is both positive and negative at the same time.

Solomon & Corbit (1974) called "hedonic aftereffects" and have interesting temporal dynamics. Here, however, we will simply assume that unfulfilled expectations will immediately shift the emotional state from hope to frustration or from fear to relief. In learning theory terms, the second two dimensions are related to omission of a reinforcer (Gray, 1975).

Together, these give a basic four-dimensional emotional space. Using the notation of Gray (1975), the basis for this space is

$$\left\langle S^{+}, S^{-}, \overline{S^{+}}, \overline{S^{-}}, \right\rangle.$$

Here, S is a stimulus (er event) and the sign indicates the valence of those stimuli. The line over the symbols indicates omission of an expected stimulus. To a first approximation and emotion E can thus be represented as a point

$$E = \langle e_1, e_2, e_3, e_4 \rangle$$

in this space. The values on each of the axes are assumed to be positive. A scalar valence can be calculated using the dot product as

$$V(E) = v \cdot E$$

where v represent that hope and relief are considered to both have positive valence while fear and frustration are both negative,

$$v = \langle 1, -1, 1, -1 \rangle.$$

We can also approximate the effects of emotions on arousal A using a similar calculation

$$A(E) = a \cdot E$$

where a indicates the effect on arousal of each emotional dimension. Alternatively, this calculation can be used to derive the level of attention that should be allocated to a stimulus (cf. Billing & Balkenius, 2014). This is important because it can aid the robot in perceiving and interpreting causal interactions as well as in its decision making capabilities (see below).

Although, different emotions can have a place in this four-dimensional space, this space does not constitute a complete characterization of an emotion. Many other factors influence the characterization, conceptualization and labeling of emotions. One such factor is whether the emotion is social or not. We suggest that the social emotions shame and pride directly parallel fear and hope. Shame can be seen as an expectation of social punishment, such as contempt, ridicule or scorn, while pride is seen as an expectation of social reward, for example, admiration or praise. Just like the omission of a non-social outcome leads to the emotions relief (in the case of absence of an expected negative event out outcome) or frustrations (in the case of absence of an expected positive outcome or event), omission of the expected social reactions causes similar effects. However, there are no separate words for these emotions when the cause is social rather than non-social.

Omission is not the only type of unfulfilled expectation. It is also possible, for example, that more praise is received than expected (or deserved). This mismatch can rebound into embarrassment. Similarly, when more punishment is received than expected or motivated, it turns into humiliation.

In all these situations, the emotional state E_{t+1} after an events depends on the expected emotional state E'_t and the actual outcome E_t such that

$$V(E_{t+1}) = V(E_t) - V(E'_t). \quad (1)$$

Note that this formalism allows for many different reactions E_{t+1} as long as they fulfil this condition and thus allows for both individual differences and different reactions depending on the exact emotions involved.

5.2 Detecting emotional reactions

Given this basic emotional framework, the robot must be able to use it to learn about other agents and to evaluate the actions of others. The dynamics of the emotional model sketched above can straightforwardly be implemented in a robot. However, it is necessary for it to be able to accurately read and respond to the emotional reactions of others. For simplicity we will assume a non-linguistic robot, so for natural interactions non-linguistic cues must be understood and reciprocated. There are several types of cues that can be detected by various sensory processing systems that can be useful to a robot.

Returning to our example in Fig. 1, for the robot to react to Agent 2 being hit it could pick up on non-verbal vocalisations (whining), painful facial expressions and bodily responses. A robot can pick up non-verbal vocalizations and analyse their emotional content without any understanding of language (Oudeyer, 2003). A significant amount of information is available in the pitch profile of non-verbal as well as verbal vocalization (Anikin et al., submitted). Moreover, such vocalizations appear to be almost universal (Scherer, 2000) and are thus a very useful source of information for a robot. Similarly, many techniques exist that can detect facial expressions in images (e.g. Turk & Pentland, 1991; Bartlett, et al. 2003; Pantic & Patras, 2006; Shan, Gong & McOwan, 2009).

A robot can recognize the posture and movements of a human body and use it to detect emotional reactions as they manifest in the human body. Many systems exist that are able to detect actions from image sequences (e. g. Xia, Chen & Aggarwal, 2012, Guha & Ward, 2012) and such systems can be adapted to detect emotional reactions. Finally, an additional cue might be available in pupil dilation, which is a more subtle signal containing useful social information (Kret, Fisher & De Dreu, 2015) that can also potentially be detected by a robot. Such signals are easily detected by dedicated eye-trackers, but a robot with a vision system of sufficient acuity could also detect this signal from a distance.

5.3 From emotions to behavior

So far, we have only discussed the evaluation of stimuli and events, but for this to have any bearing on morals, we need to connect these evaluations with behavior. This is done by noting that the valence function V above is a value function as it is used in reinforcement learning. In fact, Eq. 1 is related to the temporal difference in reinforcement learning. In the reinforcement learning paradigm, behaviors are learned as associations between a stimulus (or state) and a response (or action). For example, in the popular Q-learning algorithm (Watkins & Dyan, 1992), the expected value of an action a in a state s is represented by a function,

$$Q(s, a)$$

and action selection is reduced to the selection of an action based on this value function using some strategy. In the simplest case, the function is represented by a table that stores the expected values for each combination of state and action and the action with the maximum expected value would be selected with high probability. Another approach is to let the behaviors compete for control over a decision period (cf. Billing & Balkenius, 2014; Wong & Wang, 2006). This temporal element reflects the fact that the time when information is attended to affects valuation and choice process (Krajbich, Armel & Rangel, 2008; Lim, O'Doherty, & Rangel, 2011; Pärnamets et al., 2015).

This provides a minimal model of how emotions can be modeled in the robot, how the robot can observe others' emotions and map them onto its own valuations, and how its valuations can form the basis for action selection and decision making. However, for the robot to be able to select actions, it needs to have a better understanding of its surroundings and social context. We believe that the key here is the ability to learn from and through the interaction with others and to understand causal relations. The next two sections expand the robot framework in this regard.

6 Observational learning

A robot that can detect the expressions of social emotions can learn from its own experience which reactions its behavior will produce in a person. However, this learning ability will be limited to its own experience. It would be useful if the robot could also learn by observing the interactions of others.

Consider again the triadic interaction in Fig. 1. Two agents interact and the observer, in this case the robot, can detect the performed actions and the emotional responses of the two agents. The observed event can be used to estimate a number of quantities.

Let us first assume that the robot uses something like simple reinforcement learning, such as Q-learning. If agent 1 performs an action that results in a negative emotional reaction from agent 2, this can be used to decrease the expected value of that action. Similarly, if agent 2 reacts in a positive way, this can be used to increase the expected value of that action. This situation is very similar to that described above, except that the action is not performed by the robot itself but instead by someone else. If the observation of an action activates the same motor codes as when the robot performs the action itself, then the learning can take place in exactly the same way as if the robot had performed the action itself. Previous research shows that the mechanisms involved in observational learning of emotional value in animals and humans are similar to those used in direct conditioning (Olsson & Phelps, 2007). This claim has recently been extended by studies of the learning of instrumental actions through observation using neural (Burke et al., 2010; Crocket, 2016) and psychophysiological (Selbing, Lindström & Olsson, 2014) methods to describe the computational mechanisms of learning the value of others' actions and their consequences.

In humans and other animals, it is possible that this ability is supported by "mirror neurons" that react in the same way when we perform an action as when we see someone else performing that action (Rizzolatti, Fogassi & Gallese, 2001). Wolpert, Doya & Kawato (2003) suggested that a possible computational mechanism could be that the brain simultaneously simulate many possible actions and compare them with the observed behavior to determine which action is performed. This depends on an ability to anticipate motions and also allows us to coordinate our actions with others (Knoblich & Jordan, 2002). These mechanisms then likely interact with other brain systems supporting both habitual and goal-directed action selection (Wunderlich, Dayan & Dolan, 2012; Cushman & Morris, 2015). Other computational approaches that can be used by a robot are described by Schaal, Ijspeert & Billard (2003)

In addition to assigning value to an action based on how it influences another, there are several other properties that can be estimated from the observation of an interaction between two agents. The first is that the value of the action can be estimated in isolation in a context-independent way. For example, seeing Agent 1 hit Agent 2 and the negative reactions it produces could be used to lower the value of “hitting” in general, therefore implicitly coding that “hitting” is bad.

Another type of learning relates to the involved agents. Seeing agent 1 hit agent 2 could increase the expectation that agent 1 will perform this action again. This can be used to assign a negative valence to agent 1, but just like the valence described in the previous section is a reduced form of a multidimensional emotional space, the valence assigned to an agent can depend on many factors. The negative valence can reflect that agent 1 is stronger, hostile, more dominant, or possibly a “bad” agent. Valence can also be assigned to agent 2 in a similar way. However, here it is important exactly how agent 2 reacts both before and after being hit. Without any additional knowledge, many possible interpretations are possible. Should the valence of agent 2 be lowered because it is someone that is hit, or should it be increased to compensate for the negative valence induced by the hitting? Indeed, both cases are possible and occur in different situations. Assigning values in this way to agents is likely a central feature of morality (Uhlmann, Pizarro, & Diermeier, 2015), as perceptions of an agent’s “character” will be computationally more efficient than fully evaluating each situation. Once the robot has learned that Agent 1 tends to be the one hitting Agent 2, it can shape its interventions taking Agent 1’s bad moral character into account as soon as it recognizes the Agent (cf. Singer, Kiebel, Winston, Dolan, & Frith, 2004).

In a classical experiment, children between 42 and 71 months of age viewed a model performing hostile actions toward a doll (Bandura, Ross & Ross, 1961). When they were later allowed to play with the doll, children that had seen the model perform aggressive actions toward the doll were more likely to be aggressive towards the doll compared to children that had not observed any aggressive actions towards the doll. Similar learning effects have been observed in experiments exposing children to interacting human adults (Repacholi & Meltzoff, 2007). Importantly, observational learning depends on a range of social factors, such as experienced similarity (Bandura & Ross, 1961; Golkar, Castro & Olsson, 2015; Mobbs et al., 2009) and empathy (Olsson et al., 2016) with the involved agents.

7 Causal perception

To aid the robot’s learning in social situation it should be equipped with capabilities to understand causal relationships. This will additionally benefit its capacity for making moral judgments, since moral judgment and causal ascription are closely linked, as reviewed below.

In their seminal 1944 study Heider and Simmel showed participants simple animations of geometrical shapes moving in various directions and speeds around a larger semi-closed rectangular structure. Almost uniformly, participants reported seeing not abstract shapes buzzing about the screen, but meaningful social interactions. In particular, the majority of participants attributed detailed intentions to the shapes, seen as agents engaged in a malevolent pursuit and hosts to a range of complex intentional states such as anger, fear, persistence, shrewdness and more (Heider & Simmel, 1944). Possibly, human participants use mental state attributions to makes sense of the complex physical stimulus, hence making the retention of the observed movement patterns easier and more parsimonious (cf. Dennett, 1988). Crucial for our purposes, is that the observation of mere physical movement patterns suffices to support intentional attributions on the stimulus side. Around the same time as Heider & Simmel conducted their study, similar results were obtained by Michotte (1946/1963), who was primarily interested in the perception of causality from simple physical displays. Michotte studied simple interactions between two (sometimes three) moving objects and under which conditions participants would perceive the movements of one object as causing the movements of the second (for review see Scholl & Tremoulet, 2000).

Simple moving displays have also been used to directly elicit judgments clearly situated near or in the moral domain. In a recent study, participants evaluative judgments of agents shown in moving displays derived from the work of Michotte (i.e., simple collision events) were elicited. Participants’ evaluations fitted a dyadic template of morality, where the roles of “Agent” and “Patient” were derived from predictions arising from a combination of the underlying force dynamics (i.e. movements) with a simple normative principle of non-interference (Nagel & Waldmann, 2012). Similarly, human participants have been shown to be sensitive to a variety of kinematic factors in their judgments of severity of actions (Iliev, Sachdeva & Medin, 2012). Participants viewed a number of scenarios involving a predefined agent and patient object (a cylinder and a cone) as well as a dangerous “fireball” which caused harm to the patient. For each scenario a kinematic factor was varied, such as force, distance travelled, amount of contact, etc., and participants made severity choices between pairs of scenarios. A kinematic model, predicted choices in 80% of trials, suggesting that simple physical factors coupled with domain-general causal inference can ground a variety of moral judgments. Moving away from visual displays, work on vignettes and other abstract problem descriptions has also shown that patterns of moral judgments are dependent on causally grounded intentional ascriptions, mirroring judgments elicited for non-moral scenarios of identical causal structure (Cushman & Young, 2011). Relatedly, judgments of responsibility for joint outcomes between multiple agents have been found to depend on causal functions translating individual actions to group outcomes (Gerstenberg & Lagnado, 2010).

Underscoring the impact of causal attributions, other research has shown that causal attribution, and malicious intent, to an harmful action to the self, enhances self-rated and physiological indices of discomfort, as well as feelings of revenge (Olsson et al, under review).

Studies on human infants indicate that both the capacity for causal and moral understanding of external events develops early and at similar ages. Preschoolers ages 3-5 interpret the displays used by Heider & Simmel similarly to how adults do, inferring agency and complex intentions to the figures shown (Berry & Springer, 1993). In an early study, researchers tracked infants gaze towards animated objects moving in either goal-rational or non-rational manners (Gergely, Nádasdy, Csibra & Bíró, 1995). The results indicated that 12-month old infants could differentiate between rational and non-rational approach trajectories based on prior habituated demonstrations of agents' intentions (wanting to be close to another agent). Other work has demonstrated how infants, as young as 8- to 10-months old, are able to perceive causation for events not marked by direct physical contact, and do so for both biologically plausible and non-plausible motion patterns (Schlottmann, Surian & Ray, 2008).

We argue that causal perception will form a critical component in an autonomous moral robot, because without it they will not be able to make accurate judgments about their social world, select appropriate actions in the face of moral transgressions or couple their feelings with outside states of the world. These notions presuppose inference of causal relations and intentions. To properly infer relations of agency and patiency (cf. Gray, Waytz & Young, 2012), causal and intentional relations must be understood. Therefore, for moral robots to be able to act in their environments, they need the ability to attribute causal relations properly and from these deduce intentions and agent-patient relations.

Since causal relations can be perceived directly by looking at the temporal dynamics of interacting objects and are mediated by strict visual rules (see Scholl & Tremoulet, 2000 for examples), these rules can be implemented in the visual system of robot allowing it to determine both that the actions of one agent influences another, and the relative agency or patiency of that agent. As the robot grows more experienced, it might of course change how it values certain causal interactions, just like humans can learn the difference between a playful punch and a malicious punch. Similarly, just as in humans, the epigenetic trajectory will be constitutive of what moral agent the robot becomes (Zlatev & Balkenius, 2001).

8 Learning the consequences of actions

Learning based on reinforcement is simple and efficient since it can directly strengthen or weaken a behavior in a particular situation. This, however, is also its main limitation since the outcome of the learned behavior is not remembered. A more useful form of learning is to learn the actual consequences of actions.

The simplest action-outcome model is a set of tuples

$$\{\langle a_i, o_i\rangle\}$$

where a_i is an action and o_i is the corresponding outcome. These tuples can be learned either from the robot's own experiences or from observations just like in the examples above. The important difference from model-free reinforcement learning algorithms such as Q-learning is that these memories can be used in either direction. When the robot desires a particular outcome o_i, it can look through its database of action-outcome relations for an action that will likely produce that outcome, that is, although these structures are learned in the direction action-outcome, they can be used in the inverse order. This is therefore sometimes called an inverse model.

Inverse models allows for much more flexible use of a learned experiences and can obviously be much more complex that a simple database. For example, an inverse model typically depends on the state of the robot as well as the state of the world. The relevance of inverse model learning for a moral robot is that it allows it to explicitly choose between different outcomes and use it for reasoning about different actions and action sequences. This parallels how humans use separate valuation systems deriving from a distinction between model-free and model-based reinforcement learning (Daw, Niv & Dayan, 2005; Daw, Gershman, Seymour, Dayan, & Dolan, 2011). Recently, the model-free/model-based distinction has also been hypothesized to explain moral choices, in particular that certain responses to moral dilemmas might reflect the relative dominance of either model-free (or Pavlovian) strategies relying on heavily on immediate emotional reactions, while other reflect a switch to a more model-based strategy entailing a deeper evaluation of the decision tree (Crockett, 2013; Cushman, 2013).

9 Understanding the goals of others

For a robot to understand how an action that influences others will be met, it is often necessary to understand what the other agent is trying to accomplish. But how can a goal or intention be inferred by simply observing behavior? One way to do this is to use a generative model (Demiris, 2007; Schrodt & Butz, 2016). Such models have recently been suggested to be fundamental to how the brain works (Friston, 2010; Butz, 2016).

Put simply, a generative model G is a model that produces a specific behavior B for a particular observable state s and set of hidden parameters ϕ,

$$B = G(s, \phi).$$

Here we are interested in generative models where ϕ contains the goal that an agent attempts to accomplish. Given an observed behavior B, the task for the robot is to determine the parameters ϕ that would have produced the observed behavior. This is usually stated as an optimization problem and the parameters can be estimated, for example, using expectation maximization (Moon, 1996).

As a basic example, assume that the robot is viewing an agent A moving in an environment with an object O. The movement through the environment could potentially have something to do with O. The robot can use a generative model to test if the observed behavior is consistent with trying to approach, avoid or ignore the object O. Say the behavior is consistent with approach behavior, in this case the robot can infer that object O probably has a positive valence to agent A. With this knowledge, the robot can conclude that an action that help agent A reach O will be helpful to A while an action that makes it harder for A to approach O will hinder A.

A striking test of this ability was an experiment where 6- and 10-month old infants viewed a display of an agent trying to climb a hill (Hamlin, Wynn & Bloom, 2007). For some displays another agent hindered the climber by pushing her down the hill, while for the remainder a third agent aided the climber by pushing her up the hill. Both choice and preferential looking data show that infants strongly prefer the prosocial agents to the anti-social ones. Together, the data indicates a broad, generalized capacity to infer causal structures from moving events from an early age and using this information to support proto-moral judgments.

Generative models can also be used to understand the intentions of physical movement. For example, aggressive behavior follows a very different movement trajectory than affective behavior. Breitenberg (1984) presented some illustrative examples were simple goal-directed mechanisms give rise to movement trajectories that can be interpreted as fear, aggression, curiosity or liking. Balkenius (1995, p. 95) describes a parametrization of such behaviors were clear criteria are given for the different behavior types that could be used as a generative model.

10 Discussion

We have outlined how intrinsically moral robots can be designed by implementing seven competencies that, combined, allow a robot to learn to behave morally and make moral decisions. The framework describes high-level criteria that need to be fulfilled by a robot for it to become intrinsically moral. Each of these competencies can be implemented in different ways depending on the specific control architecture used for the robot.

- **Physiological and bodily grounding** permits to root morality inherently, so that robots have internal values that permit them to "judge" by themselves what is good or bad for them as well as for others.
- **Developmental processes** will provide a mechanism for internalizing moral values, behaviors and emotions through social interaction with humans.
- **Social emotions** will allow the robot not only to possess a dynamically updating value system but also to learn from others emotional reactions and internalize them.
- **Observational learning** ensures that the robot will learn from observing the interactions of other agents, which will provide for a greater amount of learning opportunities about how different peers value different actions.
- **Causal perception** allows the robot to infer from mechanical and physical properties of interactions who was responsible and utilize this knowledge in its moral judgments.
- **Learning the consequences of actions** allows the robot to go beyond simple learning and to generalize its learning to strive for action structures leading to desirable moral outcomes.
- **Understanding the goals of others** will let the robot to not only react to direct interactions it observes but also to proactively intervene in its environment to help or hinder other agents depending on what it believes is the right thing to do.

We have argued that robots designed in this way are intrinsically moral – in the sense that they do not merely mimic human morality, but instead generate moral judgments and behavior grounded in their own valuations, sensory-motor interactions and past experiences. In other words, their morality emerges from basic building blocks. Within the scope of their experiences, they are true moral agents. For example, empathy and compassion are often seen as emotions, but given the framework developed above they should rather be seen as the result of an ability to see others as being similar to oneself. A robot would be able become empathetic when it can use its own generative models to predict the reactions of others, and subsequently also mirror those reactions within its own emotional system. It will further show compassionate behavior by using its inverse models to select actions that will help another agent.

These mechanisms also make possible emotions such as jealousy and envy that depend on a comparison between one's own situation and that of someone else, however, it is questionable whether there would be any reason to implement such emotions in a robot.

It is possible to object that the framework proposed here is too shallow since it depends on the direct experience of the robot and does not take questions about right and wrong into consideration. However, this is exactly the reason why we believe that this is a viable path toward robots that can interact with humans in a responsible way. Each of the mechanisms we have described depend directly on the experiences of the robot and appropriate learning mechanisms. Because the robot has learned all moral behaviors by itself, or from observing others, we know that the robot will be able to detect these situations again. This contrasts sharply with a robot ethics based on explicit rules that are not grounded in the perceptual and motor abilities of the robot.

Nevertheless, there are of course several limitations to the approach we have outlined here. One such is that we have throughout worked with a simple example of a morally charged interaction - seeing one agent hitting another. While it is clear how the competencies we discuss are relevant for the robot being able to act in such a situation, it might be more difficult to see how it could learn to consider, for example, that raising a flag upside down is a terrible thing to do (assuming that this is the case in its community). This is a much more subtle action, where it might be more difficult to learn who is responsible, or to gauge reactions to the flag properly. Understanding the importance of the flag being upright presupposes understanding its symbolic and cultural value. However, these kinds of limiting cases, while important, are also examples of very sophisticated moral norms that humans construct and, we argue, something that a first minimal robot system such as the one proposed here cannot be expected to handle.

A second, related limitation, is the lack of linguistic capacity in our robot. With language, communication of norms could be expedited, and more subtle conceptual or contextual distinctions could be communicated to it. If the robot, like human children, learned its language together with learning the rest of its world, we could hope that it would also learn to symbolically reason based on the norms it has come to endorse. This would open the framework to the inclusion of explicit moral rules. However, these would still need to be grounded in the different competences listed above.

Third, our approach, with its emphasis on social emotions and observational learning entails that the robot will acquire part of its moral valuations from how agents around it act and react to each other. As autonomous robots are rare, it is likely that these will be humans, which raises the question how good models they (we) are? This limitation allows us to highlight the important distinction between acting from what we think is right – what a moral robot can be expected to learn to do – and acting in a way which is ultimately right - what philosophers are still discussing. What morality the robot will acquire will be dependent on where it spends its formative years, but it will nevertheless be moral as acting consistently with its emotional and causal appraisals of various situations.

At the start of this paper we motivated the development of moral robots with concerns about potential risks of introducing artificial autonomous agents in a human society, but it is also worth highlighting another benefit of our approach, namely that autonomous moral robots will likely be easier for humans to interact with. This is because their morality, like ours, will be grounded in their sensory-motor experiences and based on a history of social learning through their interaction with humans. They will be beings inhabiting similar lifeworlds to ours (cf. Von Uexküll, 1934/2010), making them closer to becoming not only agents of equal moral standing with us, but possibly also being treated as moral patients in their own right. We believe this is a necessary step for true social interactions to take place between robots and humans.

To conclude, we view morality as intrinsically linked to complex social cognition and behavior. In fact, this link might be universally applicable across entities with such social features, ranging from primates (de Waal, 1996) to autonomous robots as described in this paper. We hope that our suggested design features for an intrinsically moral social robot will aid in the construction of artificial agents that can be fully trusted by both their users and by the public at large. Only when artificial intelligence is intrinsically moral, fear of it will dissipate.

References

Anderson, S. L. (2008). Asimov's "laws of robotics" and machine metaethics. *AI & Society, 22*(4), 477–493.

Anderson, M., & Anderson, S. L. (2007). The status of machine ethics: a report from the AAAI Symposium. *Minds and Machines, 17*(1), 1–10.

Asada, M. et al., (2009). Cognitive Developmental Robotics: A Survey. *IEEE Transactions on Autonomous Mental Development, 1*(1), 12–34.

Avila-García, O. & Cañamero, L. (2004). Using hormonal feedback to modulate action selection in a competitive scenario. In *Proc. Eighth Intl. Conf. on Simulation of Adaptive Behavior (SAB04)* (pp. 243–252). Cambridge, MA: The MIT Press.

Balkenius, C. (1995). *Natural intelligence in artificial creatures*. Lund University Cognitive Studies, 37.

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through the imitation of aggressive models. *Journal of Abnormal and Social Psychology, 63*(3): 575–82

Bartlett, M. S., Littlewort, G., Fasel, I., & Movellan, J. R. (2003, June). Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on* (Vol. 5, pp. 53–53). IEEE.

Berry, D. S., & Springer, K. (1993). Structure, motion, and preschoolers' perceptions of social causality. *Ecological Psychology, 5*(4), 273–283.

Berthouze, L., & Ziemke, T. (Eds.) (2003). *Connection Science, 15*(4). Special Issue: Epigenetic Robotics: Modelling Cognitive Development in Robotic Systems.

Billing, E., & Balkenius, C. (2014). Modeling the Interplay between Conditioning and Attention in a Humanoid Robot: Habituation and Attentional Blocking. In *Proceeding of The 4th International Conference on Development and Learning and on Epigenetic Robotics (IEEE ICDL-EPIROB 2014)* (pp. 41–47).

Burke, C. J. Tobler, P. N. Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences, 107*(32), 1443.14436.

Butz, M. V. (2016). Towards a Unified Sub-Symbolic Computational Theory of Cognition. *Frontiers in Psychology, 7*, 925.

Braitenberg, V. *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press, 1984.

Cañamero, L. D. (1997). Modeling Motivations and Emotions as a Basis for Intelligent Behavior. In W. Lewis Johnson (Ed), *Proceedings of the First International Conference on Autonomous Agents* (pp. 148–155). New York: The ACM Press.

Cañamero, L. D. (2001). Emotions and Adaptation in Autonomous Agents: A Design Perspective, *Cybernetics and Systems, 32*(5): 507–529.

Cañamero, L. D. (2003). Designing Emotions for Activity Selection in Autonomous Agents. In R. Trappl, P. Petta, S. Payr (Eds.), *Emotions in Humans and Artifacts* (pp. 115-148). Cambridge, MA: The MIT Press.

Cañamero, L. (2008). Animating Affective Robots for Social Interaction. In L. Cañamero, R. Aylett (Eds.), *Animating Expressive Characters for Social Interaction* (pp. 87–122). John Benjamins Publishing Co., Advances in Consciousness Research Series.

Cañamero, L., & Avila-García, O. (2007). A bottom-up investigation of emotional modulation in competitive scenarios. In A. C. R. Paiva, R. Prada, R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction* (ACII 2007, pp. 398–409). Springer, Berlin, Heidelberg.

Cañamero, L., Blanchard, A., & Nadel, J. (2006). Attachment Bonds for Human-Like Robots. *International Journal of Humanoid Robotics, 3*(3): 301–320.

Cañamero, L., & Lewis, M. (submitted). Hedonic Quality or Reward? A Study of Basic Pleasure in Homeostasis and Decision Making of a Motivated Autonomous Robot. (under review.)

Colombetti, G. (2014). The Feeling Body: Affective Science Meets the Enactive Mind. The MIT Press.

Cos, I., Cañamero, L., & Hayes, G. M. (2010). Learning affordances of consummatory behaviors: Motivation-driven adaptive perception. *Adaptive Behavior, 18*(3–4): 285–314.

Cos-Aguilera, I., Cañamero, L., Hayes, G. M., and Gillies, A. (2013). Hedonic Value: Enhancing Adaptation for Motivated Agents. *Adaptive Behavior, 21*(6), 465–483.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17*(8), 363–366.

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences, 112*(45), 13817–13822.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*(6), 1052–1075.

Cushman, F. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.

Damasio, A. (1999). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. London: Vintage.

Damasio, A. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. London: Vintage.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215.

De Jaegher, H., & Di Paolo, E. (2007). Participatory Sense-Making: An Enactive Approach to Social Cognition. *Phenomenology and the Cognitive Sciences, 6*(4): 485–507.

De Waal, F. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Harvard University Press.

Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and Brain Sciences, 11*(03), 495–505.

Demiris, Y. (2007). Prediction of intent in robotics and multi-agent systems. *Cognitive Processing, 8*(3), 151–158.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci. 11*, 127–138.

Gray, J. A. (1975). *Elements of a two-process theory of learning*. Academic Press.

Guha, T., & Ward, R. K. (2012). Learning sparse representations for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(8), 1576–1588.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165–193.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition, 115*(1), 166–171.

Golkar, A., Castro, V., & Olsson, A. (2015). Social learning of fear and safety is determined by the demonstrator's racial group. *Biology Letters, 11*(1), 20140817.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry, 23*(2), 206–215.

Griffiths, P. E. & Scarantino, A. (2009). Emotions in the wild: The situated perspective on emotion In Aydede, M. & Robbins, P. (Eds.) *The Cambridge Handbook of Situated Cognition*. New York, NY: Cambridge University Press.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557–559.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*(2), 243–259.

Hiolle, A., Lewis, M., & Cañamero, L. (2014). Arousal Regulation and Affective Adaptation to Human Responsiveness by a Robot that Explores and Learns a Novel Environment. *Frontiers in Neurorobotics, 8*(17). http://dx.doi.org/10.3389/fnbot.2014.00017.

Iliev, R. I., Sachdeva, S., & Medin, D. L. (2012). Moral kinematics: The role of physical factors in moral judgments. *Memory & Cognition, 40*(8), 1387–1401.

Knoblich, G., & Jordan, S. (2002). The mirror system and joint action. In M. I. Stamenov & V. Gallese (Eds.) *Mirror Neurons and the Evolution of Brain and Language* (pp. 115-124). Amsterdam: John Benjamins.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13*(10), 1292–1298.

Kret, M. E., Fischer, A. H., & De Dreu, C. K. (2015). Pupil mimicry correlates with trust in in-group partners with dilating pupils. *Psychological science, 26*(9), 1401–1410.

Krichmar J. L & Röhrbein, F. (2013). Value and Reward Based Learning in Neurorobots. *Frontiers in Neurorobotics, 7*(13). http://dx.doi.org/10.3389/fnbot.2013.00013.

Laird, J. D. (2007). *Feelings: The Perception of Self*. Oxford University Press.

Lim, S. L., O'Doherty, J. P., & Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *The Journal of Neuroscience, 31*(37), 13214–13223.

Lones, J., Lewis, M., & Cañamero, L. (2016). From Sensorimotor Experiences to Cognitive Development: How Does Experiential Diversity Influence the Development of an Epigenetic Robot? *Frontiers in Robotics and AI* (in press). http://dx.doi.org/10.3389/frobt.2016.00044.

Lowe, R. (2014). Embodiment in Emotional Learning, Decision Making and Behaviour: The 'What' and the 'How' of Action. In C. Stephanidis and M. Antona (Eds.), *Universal Access in Human-Computer Interaction. Aging and Assistive Environments: 8th International Conference, UAHCI 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part III* (pp. 672–679). Springer International Publishing.

Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection Science, 15*(4), pp. 151–190.

McLaren, B. M. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. Intelligent Systems, IEEE, 21(4), 29–37.

Michotte, A. (1946). *The perception of causality*. London: Methuen.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine, 13*(6), 47–60.

Murphy, R. R., & Woods, D. D. (2009). Beyond Asimov: the three laws of responsible robotics. *IEEE Intelligent Systems, 24*(4), 14–20.

Nagel, J., & Waldmann, M.R. (2012). Force dynamics as a basis for moral intuitions. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 785–790). Austin, TX: Cognitive Science Society.

Navarrete, C. D., McDonald, M. M., Asher, B. D., Kerr, N. L., Yokota, K., Olsson, A., & Sidanius, J. (2012). Fear is Readily Associated with an Out-group Face in a Minimal Group Context, *Evolution & Human Behavior, 33*(5), 590–593.

Norman, D. (2005). *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic books.

Olsson, A., & Phelps, E. A. (2007). Social Learning of Fear. *Nat. Neurosci., 10*, 1095–1102.

Olsson, A., McMahon, K., Papenberg, G., Zaki, J., Bolger, N., & Ochsner, K. N. (2015). Vicarious fear learning depends on empathic appraisals and trait empathy. *Psychological Science, 0956797615604124*.

Olsson, A., Brodbeck, C., Bolger, N., & Ochsner, K. N. (submitted). Once shocked, twice shy: How perception of the intent to punish influences negative evaluations and conditioned fear of others (under review).

Oudeyer, P-Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies, 59*(1), 157–183.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.

Pantic, M., & Patras, I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 36*(2), 433–449.

Parisi, D. (2004). Internal Robotics. *Connection Science, 16*(4) 325–338.

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences, 112*(13), 4170–4175.

Pepperberg, I. M. (2001). Lessons from Cognitive Ethology: Animal Models for Ethological Computing. In Balkenius, C., Zlatev, J., Kozima, H., Dautenhahn, K., and Breazeal, C. (Eds.), *Proceedings of the First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. Lund University Cognitive Studies, 85. Lund: LUCS.

Pezzulo, G. (2011). Grounding procedural and declarative knowledge in sensorimotor anticipation. *Mind & Language, 26*(1), 78-114.

Pfeifer, R. (1996). Building "Fungus Eaters": Design Principles of Autonomous Agents. In P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, and S.O. Wilson (Eds.), *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior SAB96 (From Animals to Animats)* (pp. 3–12). Cambridge, MA:The MIT Press.

Prince C. G. & Gogate L.J. (2007). Epigenetic Robotics: Behavioral Treatments and Potential New Models for Developmental Pediatrics. *Pediatric Research, 61*, 383–385.

Prince C. G. , Helder, N.A., & Hollich, G.L. (2005). Ongoing Emergence: A Core Concept in Epigenetic Robotics. In Berthouze, L., Kaplan, F., Kozima, H., Yano, H., Konczak, J., Metta, G., Nadel, J., Sandini, G., Stojanov, G. and Balkenius, C. (Eds.) *Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. Lund University Cognitive Studies, 123.

Prince, C. G., & Demiris, Y. (Eds.) (2003). *Adaptive Behavior, 11*(2). Special Issue on Epigenetic Robotics.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience, 2*(9), 661–670.

Rolls, E. T. (1990). A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition & Emotion, 4*(3), 161–190.

Schaal, S., Ijspeert, A., & Billard, A. (2003). Computational approaches to motor learning by imitation. *Phil. Trans. R. Soc. Lond. B 358*, 537–547.

Scherer, K. R. (2000). A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology. In *INTERSPEECH* (pp. 379–382).

Schlottmann, A., Surian, L., & Ray, E. D. (2009). Causal perception of action-and-reaction sequences in 8-to 10-month-olds. *Journal of Experimental Child Psychology, 103*(1), 87–107.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4*(8), 299–309.

Schrodt, F., & Butz, M. V. (2016). Just Imagine! Learning to Emulate and Infer Actions with a Stochastic Generative Architecture. *Frontiers in Robotics and AI, 3*(5).

Selbing, I., Lindström, B., & Olsson, A. (2014). Demonstrator Skill Modulates Observational Aversive Learning. *Cognition, 133*(1), 128–139.

Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing, 27*(6), 803–816.

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron, 41*(4), 653–662.

Sloman, A. (2006). *Why Asimov's Three Laws of Robotics are Unethical*. Retrieved June, 16, 2016.

Solomon, R. C. (2007). *True to Our Feelings: What Our Emotions are Really Telling Us*. Oxford University Press.

Solomon, R. L., & Corbit, J. D. (1974). An opponent-process theory of motivation: I. Temporal dynamics of affect. *Psychological Review, 81*(2), 119.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*(1), 71–86.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81.

Von Uexküll, J. (1934/2010). *A foray into the worlds of animals and humans: With a theory of meaning*. U of Minnesota Press.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning, 8*(3–4), 279–292.

Wolpert D. M, Doya, K., & Kawato M. (2003). A unifying computational framework for motor control and social interaction. *Philos Trans R Soc Lond B Biol Sci, 358*, 593–602.

Wong, K. F., & Wang, X. J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *The Journal of Neuroscience, 26*(4), 1314–1328.

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience, 15*(5), 786–791.

Xia, L., Chen, C. C., & Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference* (pp. 20–27). IEEE.

Ziemke, T. (2003). What's That Thing Called Embodiment? In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 1305–1310). Mahwah, NJ: Lawrence Erlbaum.

Zlatev, J., & Balkenius, C. (2001). Introduction: Why "Epigenetic Robotics". In Balkenius, C., Zlatev, J., Kozima, H., Dautenhahn, K., and Breazeal, C. (2004). *Proceedings of the First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. Lund University Cognitive Studies, 85. Lund: LUCS.

Christian Balkenius is a professor of Cognitive Science at Lund University where he received his PhD in 1995. His research goal is to understand the cognitive and developmental processes involved in perception, learning and emotion at both a neural and computational level. The research ranges from computational models of conditioning, to learning mechanisms in the control of visual attention as well as the design of robots with these abilities. His work is highly interdisciplinary combining experimental, modeling and engineering techniques.

Martin Butz With a Diplom with honors in computer science including a minor in psychology (University of Würzburg, Germany, 08/2001) and a PhD in computer science (University of Illinois at Urbana-Champaign, IL, USA, 10/2004), Dr. Butz has been pursuing interdisciplinary, collaborative research in cognitive science for nearly two decades. Since 09/2011 Dr. Butz is full professor in Cognitive Modeling at the Department of Computer Science and the Department of Psychology in the Faculty of Science of the University of Tübingen, Germany. Dr. Butz has published more than fifty journal articles in various disciplinary and interdisciplinary journals. His third monography, which is called "How the Mind Comes Into Being: Introducing Cognitive Science from a Functional and Computational Perspective" is scheduled to be published by Oxford University Press this fall (2016).

Lola Cañamero is Reader in Adaptive Systems and Head of the Embodied Emotion, Cognition and (Inter-)Action Lab in the School of Computer Science at the University of Hertfordshire in the U.K, which she joined as faculty in 2001. She holds undergraduate (Licenciatura) and postgraduate degrees in Philosophy from the Complutense University of Madrid and a PhD in Computer Science (Artificial Intelligence) from the University of Paris-XI. She turned to Embodied AI and robotics as a postdoctoral fellow in the groups of Rodney Brooks at MIT and of Luc Steels at the Free University of Brussels. Since 1995, her research investigates the interactions between motivation, emotion, and embodied cognition and action from the perspectives of adaptation, development and evolution, using autonomous and social robots and artificial life simulations.
Website: http://www.emotion-modeling.info.

Birger Johansson is a researcher at Lund University Cognitive Science, Sweden and the Department of Psychology, Uppsala University, Sweden. He received his Ph.D. in Cognitive Science in 2009 at Lund University. His main research focus is to build robots motivated by child development. He is currently involved in building the Epi robot at Lund University Cognitive Science Robotics Group.

Andreas Olsson is an associate professor in clinical neuroscience at Karolinska Institutet, where he directs the EmotionLab (www.emotionlab.se). He completed his Ph.D. in psychology at New York University and post-doctoral training in social neuroscience at Columbia University. Broadly, Olsson's research aims at better understanding emotional learning, regulation and decision-making in social situations. He takes a multi-method approach, including the experimental study of behavior, peripheral psychophysiology, and functional magnetic resonance imaging (fMRI).

Philip Pärnamets is a postdoctoral researcher at Karolinska Institutet, in Sweden. He studied politics, philosophy, economics at the Universityof York before completing a Ph.D. in cognitive science from Lund University in 2015. In his work he uses behavioral experiments, eye-movements and other physiological correlates, as well as computational models to study learning and decision-making in social and moral contexts.