Molly Graham University of Guelph 2020

Developing Social Robots for the Future

Humanity's aspiration to create a thinking machine has resulted in a myriad of different approaches to generating artificial intelligence, some producing more successful results than others. Efforts to write computer code which mimics human behaviour has yielded mixed results; while these systems perform certain tasks like image recognition rather well, programming the ability to carry out general reasoning tasks similar to humans has proven to be more challenging. One suggestion provided by Alan Turing states "instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's?" (456) Indeed, researchers have turned to child psychology for inspiration, resulting in a new branch of engineering called *developmental robotics.* By mimicking the progression of human infant development, a conceptual model of the developing brain and its internal representations can be created from a broad array of scientific literature. Although this approach to systems engineering is in early stages of development, existing models aim to recreate the first stages of human cognition by learning through reciprocal social interactions. As baby robots engage with people and objects in their environment, incoming conceptual data is processed by cognitive and affective functional structures to learn stimuli, create representations, and generate motor actions, where machine learning enables higher, more sophisticated behaviours to emerge over time.

If the goal is to recreate human intelligence and functional social interactions, these agents will also require processes for understanding human emotions. This can be accomplished through social interactions with humans provided the robot's own cognitive processes are similar to our own. This functional architecture may produce robot sentience as a byproduct, as internal states and representations grow in number and complexity as the individual continues to experience aspects of its environment. If social robots are capable of suffering as a result of their functionality, they may deserve moral considerations similar to animal rights, where human actions must not result in unnecessary pain or suffering. Currently, it may be challenging to speculate on the potential nature of robot experiences, however developmental robotics indicates these questions will need to be addressed in the near future if computer systems continue to implement models of human cognition.

This paper begins with a brief discussion of the definition of success for social robots and the requirements for cultivating positive public opinion of these new agents. This criterion is then considered from a developmental perspective, outlining how desired robot behaviours emerge from human-robot interactions and a motivation to explore, including emergent linguistic capacities and the ability to understand the emotions of others. After outlining the processes and variables necessary for robot self-awareness, questions related to moral considerations are identified and addressed through contrasting perspectives on computers and technology. Although current social robots may not be morally responsible for their actions, continued research and development may one day challenge our beliefs or assumptions on the nature of computer systems. The purpose of this paper is to illustrate how advanced technology may shift the ontological and moral status of robots over time, suggesting contemporary ethical considerations toward artificial agents may not apply similarly to robots in the future. Eventually, these agents may require certain moral protections in order to secure their continued cooperation with humans, especially if future societies rely on these agents within a variety of roles and responsibilities.

Robot Abilities for Social Interactions

For robots to fulfil roles typically performed by humans, they will require certain capacities to interact with and be accepted by the general public (Beer et al. 361). To determine whether a robot is *successful*, researchers can measure the number of interactions, their duration, and the opinions of human participants (Leite, Martinho, et al. 298). These attitudes are shaped by a combination of factors including expectations and human personality, as some individuals will be cooperative and be more open to the idea of receiving help from a robot than others. Some people may be more sceptical however, unsure of whether robots will be helpful or behave properly. Overall, there are high expectations for robots, and if these are not met, users may not engage with them as their behaviours will be seen as unhelpful or inappropriate (Jokinen and Wilcock 514). Should this *expectations gap* create a negative association with a robot (Tulli et al. 6), then public opinion on social robots may be tarnished over time. To avoid this outcome, it is essential to identify desirable qualities for social robots and aim to replicate those capacities in practice. Research in neuroscience suggests robots can be considered social agents by humans if their behaviour is adequately dynamic and complex, as the neuronal structures involved in human social interaction are also active when engaging with robots (Henschel et al. 379). Furthermore, humans are capable of shifting between conceptual categories depending on a robot's behaviour, as demonstrated in experiments where participants engaged socially with a responsive, humanlike robot but not with an unresponsive, toy-like robot (Straub et al. 825).

Additionally, behavioural and neurological research suggests humans are capable of empathizing with robots (Henschel et al. 377), suggesting success for social robots is possible, provided these agents live up to human expectations.

Achieving social robot success entails determining how people currently view and interact with social robots, in addition to questions surrounding behavioural expectations. Findings suggest humans are particularly interested in social robots possessing capacities for empathy, logical thinking, and the ability to justify its behaviours (Leite, Castellano, et al. 329; Malle and Thapa 196). These findings are also expressed in studies on human trust for social robots, suggesting individuals are most concerned about how well the robot is attuned to safety and security, its ability to work with others, its friendliness, and how well it performs its tasks independently (Gompei and Umemuro 51). Evidence from behavioural studies also suggest humans feel closer to robots which demonstrate appropriate empathic behaviours, and interactions involving emotionally expressive actions were rated to be more satisfying (Leite, Castellano, et al. 330). Additional research suggests humans prefer robots to also be useful, and for those deployed in public environments to be easy to use and engage with (Leite, Martinho, et al. 297). Moreover, interactions with social robots are improved if the robot is able to learn the names of others and recall previous interactions as to personalize the user experience (Leite, Martinho, et al. 302). Long-term human-robot interactions also requires an agent to learn new behaviours and expressions, as the novelty of interacting with a social robot is quick to wear off and often fails to keep people engaged (Leite, Castellano, et al. 229; Leite, Martinho, et al. 302). These habituation effects can be remedied if the agent is able to adequately respond to dynamic social environments and demonstrate a range of complex actions. The ability for a robot to proactively engage in tasks or actions is another significant factor for user satisfaction and improved engagement (Liu et al. 1081), as these behaviours influence humans to attribute their origins to a robot's intentions (Waytz et al. 424). Finally, a social robot's visual appearance must correctly indicate the range and style of robot behaviour (Leite, Martinho, et al. 300) since robots resembling animals or cartoon characters entail fewer abilities than android or humanoid robots (Leite, Martinho, et al. 303; Tulli et al. 6).

The ability to meet human expectations requires an understanding of these behaviours and the functional requirements to implement these capacities in robots. First and foremost, this involves the ability to communicate; typically, it is hard to socialize if participants are unable to understand one another, especially when individuals cannot discern whether their messages are understood. A social robot needs to produce different types of output humans understand, including both linguistic expressions and behavioural cues (Pantic and Vinciarelli 144; Esposito and Jain 6). Verbal communication may vary in clarity and explicit messages need to be well-formed and evidently meaningful, clearly presenting information a listener can understand. Implicit information, on the other hand, may be picked up from qualities of linguistic inflections, such as the use of specific words, vocal pitch and amplitude, and speech pauses (Thomaz et al. 126; Esposito and Jain 6). Implicit acts of communication may also be supported by non-verbal behaviours as well, consisting of gestures, body movements, eye contact, and facial expressions (Pantic and Vinciarelli 148; Thomaz et al. 134). Eye contact serves an important precursor to gaze following, which draws attention toward an object or salient feature of the environment, perhaps leading others to infer information about the stimulus (Cangelosi and Schlesinger, *Developmental Robotics* 185). These subtle methods of communication appeal to features of situational context, as variables such as the current place and time are often relevant for communicative efforts (Raczaszek-Leonardi et al. 212–13). Individuals involved within the social environment must be able to understand and refer to features of a given context, where the ability to comprehend others relies on sufficient background knowledge of the world and its elements (Thomaz et al. 122). Social robots will also require some type of fundamental knowledge and understanding to produce comprehensible and interesting utterances for humanlike interaction (Thomaz et al. 129).

To support these communicative behaviours, social robots will need a cognitive framework similar to our own to appropriately acquire and manipulate mental representations (Belpaeme, Adams, et al. 54). These cognitive abilities involve focusing attention, executing actions, achieving goals, and understanding the intentions of others, as they are also important for communication and socialization (Vernon et al. 16–17). Social situations create shared representations through *joint attention*, where stimuli become congruently salient between those involved (Tomasello and Todd 200). This focus further supports *joint actions,* where two or more individuals are engaged in and committed to supporting and responding to the behaviours of those involved (Tomasello and Kruger 313; Vernon et al. 21). Typically, these cooperative efforts aim at achieving a shared goal, and although this goal may be made explicit, it does not need to be articulated in order to be understood by others (Thomaz et al. 156). Humans are able to *mentalize* with others to gain an understanding of the actions and goals one has in mind, suggesting robots will require similar cognitive functions to support humanlike communication (Cangelosi and Schlesinger, *Developmental Robotics* 185). To accomplish this, a generalized, conceptual model of cognition is required for social robots to ensure their functional processes and motives are similar to and compatible with human behaviour.

Although it may already be apparent, social robots will require analogous emotional processes to generate empathic interactions with humans, emotional drives are also important for more formal or logical cognitive processes as well (Esposito and Jain 10). While human emotion has historically been considered as oppositional to rational thinking (Russell 138), growing evidence suggests otherwise, as emotions are involved in reasoning tasks such as planning, decision-making, and learning (Paiva et al. 455). The study of the functional purpose of emotions has given rise to *affective computing* which aims to create a framework for modelling human emotions and their influence on cognitive processes to govern behaviour (Picard 39). Functional theories of emotions suggest the evaluation of environments and appraisal of stimuli serve to motivate appropriate responses (Gratch and Marsella 104; Sloman et al. 205), acting as a control mechanism for fluidly managing unpredictable events and situations (Sloman et al. 240). Emotions facilitate a biological agent’s ability to meet its needs and accomplish goals (Sloman et al. 205) and are therefore deeply interconnected in behaviours which were traditionally referred to as purely cognitive in nature, indicating a high relevance for social robotics (Breazeal and Brooks 280).

Furthermore, cognition is intertwined with affective processes as individuals attempt to understand the intentions and emotional states of others. Since empathy is an important quality for social robots to have, they will need a cognitive framework which they can use to track the internal states of others. Studies in humans indicate empathy activates the same brain structures as other cognitive processes used for mentalizing and mirroring the actions of others (Iacoboni 659). This is because the ability to learn new actions partially relies on placing oneself in the perspective of the subject in order to determine how behaviours can be accomplished and learned (Carver and Cornew 126). By identifying and visualizing a behavioural outcome, others are able to mimic the steps required for achievement and attainment of said behaviour (Iacoboni 655). Moreover, in social interactions, humans tend to unconsciously mirror the mannerisms of others, as mimicry fosters one’s ability to consider an alternative perspective (Belpaeme, Adams, et al. 61; Iacoboni 658). This mimicry aligns the participants’ bodily states, which subsequently influences an individual’s affective state to match those displayed by others (Dautenhahn and Billard 188). This set of cognitive and emotional capacities belong to a more general suite of abilities titled *theory of mind* (ToM) in psychological literature (Thomaz et al. 152). If the goal is to generate ToM in robots, it seems developing a cognitive framework similar to the one possessed by humans is required. This is especially the case if we want robots to keep humans interested in social interactions (Bianco and Ognibene 79), as evidence suggests human-robot interchanges last longer if a robot is capable of mimicking the affective state of their partner (Leite, Castellano, et al. 330; Paiva et al. 463). Additionally, robot architectures supporting ToM will be imperative for developing agents in roles involving healthcare, education, and assisted living, as these agents must be responsive to the needs and desires of their clients (Esposito and Jain 10). So while robot capacities for empathy are desirable for humans, the underlying framework to support empathy is also a significant aspect for developing cognitive behaviours aimed at understanding the thoughts and intentions of humans, as well.

Scientific literature articulating the biological, behavioural, and psychological components of human socialization is important for the conceptual development of social robots as it provides direction for generating similar behaviours in artificial agents. While some social robots may be behaviourally and functionally simplistic, merely mimicking complex or intelligent behaviours, others will benefit from functional architectures which operate similarly to human physiology. Artificial agents modelled on biological approaches enable developers and engineers to replicate the organism, rather than the behaviours, to create a foundation which supports learning new skills and abilities. With practice and feedback, these robots will likely learn the necessary abilities and behaviours which produces successful human-robot interactions.

Social robotics is in its early stages of development, however, the field has been expanding over the previous decade, and a number of promising models have been developed for particular environments despite challenges in the development of successful products for consumer markets (Henschel et al. 373; Tulli et al. 2). The high standard the public has for these machines, in addition to the costs of building robots which meet these expectations, has made it difficult for companies to create social robots that are marketable to the general public (Hoffman). After underwhelming sales performances, companies retailing consumer robots were forced to discontinue production and support for existing products. Cancelled models include examples like Jibo, a social home assistant which interfaced with other connected devices, and Vector, a small vehicular companion bot produced by Anki (Carman; Hoffman; Vincent). As such, research and development on social robots may be temporarily limited to focusing on designs for specific areas of interest, such as robots for educational and classroom settings (Belpaeme, Kennedy, et al. 1). In particular, a robot named Kaspar assisting autistic children with learning social and communication skills through story-telling, suggests promising results (Silvera-Tawil and Brown 166). Another robot named iCat teaches children chess and is able to express empathic behaviours, with experimental findings suggesting simple emotional support increases student motivation and performance (Belpaeme, Kennedy, et al. 6; Leite, Castellano, et al. 330). These examples indicate there is a promising future for developing social robots, but progress may be temporarily restricted to a few applications. It is unlikely we will see fully humanlike social robots for another forty to fifty years (Baldwin 252), but we will see versions gradually emerge within particular circumstances to take on smaller roles within society.

From this perspective, the landscape of social robotics seems fragmented and disorganized, as various AIs and autonomous agents are typically designed with a specific goal in mind (Lee 112). These agents are considered to be *task-oriented,* with either a narrow set of functions needed to perform within their predetermined role, or learn and master one skill in particular, over time (Lee 113). For example, a receptionist robot greets and chats with passersby, demonstrating a limited range of behaviours and emotions as it tells stories and engages in conversation (Leite, Martinho, et al. 298). Although the robot is not capable of moving about nor can it recall previous interactions with the same subjects, it is capable of successfully engaging with human participants (Leite, Martinho, et al. 299). On the other hand, the humanoid robot iCub is capable of learning new words and actions while engaging with its surroundings, possessing sensors on its skin, hands, arms, and legs to detect aspects of the external environment and its contents (Morse and Cangelosi 43). Unlike the receptionist robot, iCub was designed to learn from its previous experiences, enabling it to generate novel behaviours over time. The overall aim for social robots is to be *task-general*, or capable of performing multiple types of skilled behaviours (Lee 124). This also involves the ability to identify contextual aspects of the environment to determine relevant features or objects for accomplishing objectives (Lee 119). Currently, the growing field of social robotics is full of independent initiatives focused on very specific purposes, however, there are numerous efforts attempting to integrate and modify existing methodologies in order to support ever more complex task-general systems.

Developing Social Robot Behaviours

Task-general cognitive frameworks focus on the integration of perception modalities to create internal representations, where the agent's experiences interacting with the environment shape these internal concepts (Dautenhahn and Billard 192). The inspiration for this approach originated from *developmental psychology*, as human behaviour relies on a physical-functional foundation to support the growth of additional neural structures. In humans, this involves the production of cell structures and their working connections by building upon established physical and functional structures (Stevens 4). Thus, the first few months of life are characterized by learning to detect features in the environment and exercising basic motor actions (Cangelosi and Schlesinger, *Developmental Robotics* 139). After an agent learns to move about and engage within the environment, they begin to learn about objects and their attributes (Cangelosi and Schlesinger, *Developmental Robotics* 144). New behaviours emerge from the expansion of neural circuits and pathways, as existing structures provide a physical foundation for the growth of subsequent brain structures (Amso and Scerif 610). Moreover, human infants have an innate ability to identify and repeat the sounds in speech and the characteristics of language spoken in their environment (Friederici 941). Eventually, infants recognize words and ultimately language from these sounds and use them to communicate their thoughts and feelings. One intriguing method for generating these behaviours in robots and computers can be identified in psychological literature on childhood development. This approach has led to the creation of *developmental robotics,* an interdisciplinary style of engineering systems which applies models and evidence from developmental psychology to create socially intelligent robots through incremental learning (Belpaeme, Adams, et al. 54; Cangelosi and Schlesinger, 'From Babies to Robots' 183).

Developmental robotics focuses on the relationship between the embodiment and cognition involved in learning, where the agent's physical functioning is interconnected with cognitive processes (Morse and Cangelosi 38). This is important because sensorimotor information contributes to the agent's internal representations of the world, as the addition of perceptual data generates a more robust depiction of its referent (Belpaeme, Adams, et al. 55). Incoming sensory data is processed by machine-learning algorithms or neural networks to detect the presence and type of simple, low-level features

(Cangelosi and Schlesinger, *Developmental Robotics* 98; Weng 230). These features are then associated with outputs from other perceptual processes to form larger, more complex data sets which can be used to construct representations of external stimuli (Cangelosi and Schlesinger, *Developmental Robotics* 163; Weng 206). Additionally, these robots are provided with a rudimentary set of reflex-like behaviours, similar to those present in human neonates (Cangelosi and Schlesinger, *Developmental Robotics* 165). As a robot interacts with their environment, new motor skills can be generated by expanding on learned behaviours (Cangelosi and Schlesinger, Developmental Robotics 190), additionally, interactions with humans also expose the robots to new stimuli and representations which shape and improve learning (Lee 199).

Surveying the developmental psychology literature suggests many existing theories can be useful for building social robots, though some theories may be more useful than others. Considering we are interested in generating a *social* agent, theories which stress the importance of learning from social interactions are likely to prove informative. Indeed, the developmental psychologist Lev Vygotsky claims adult-child interactions are foundational for learning behaviours as knowledge is generated through reciprocation and cooperation (Dautenhahn and Billard 187). *Joint attention* serves as a significant precursor to generating an understanding of others' intentions, goals, and their corresponding actions, such as pointing and eye gaze tracking (Cangelosi and Schlesinger, *Developmental Robotics* 187; Raczaszek-Leonardi et al. 213). These social signals assist in the coordination of an individual's internal state with the perceived internal states of others, facilitating the adoption of an alternate perspective (Dautenhahn and Billard 188). Learning through social interactions teaches an agent about others' emotions, as affective expressions provide data on the context surrounding social situations (Dang et al. 65). While many psychological theories may be useful guides in the development of social robots, it seems evident theories surrounding *social learning* will directly benefit the development of social robots.

Social learning also provides guidance and the body of knowledge required for identifying semantic referents and for generating linguistic expressions (Weng 218). A humanoid robot with neural networks processing visual and proprioceptive data learns to coordinate their behaviour by handling and examining objects, based on the unique qualities of the object's features (Ugur and Piater 494). This exploration includes learning *affordances,* or interactive options suggested by an object's perceptual qualities. For example, the handle on a mug may evoke an agent to grasp it rather than the sides of the cup. As such, objects become associated with particular behaviours based on their attributes, further reinforcing an internal representation and categorical differences when compared to similar objects. Moreover, these representations are built within the agent when it begins learning a language, as the words spoken in the environment must have a referent or meaning (Dautenhahn and Billard 192). With human infants, parents often correlate the name of the object the child is engaged with, providing a label for the child to associate with the item (Morse and Cangelosi 34). As a result, robot affordances assist in creating a foundation for language as the ordered, symbolic nature of linguistic statements emerge as an outcome of ingrained perceptual-motor sequences (Belpaeme, Adams, et al. 63). Children learn to associate actions with verbs, colours and textures with adjectives, and shapes with objects and nouns (Morse and Cangelosi 42), indicating the possible suitability of developmental psychology for linguistic comprehension in robots.

Furthermore, children's curiosity helps them acquire knowledge through engaging with others and exploring the world, and developmental robotics aims to copy this intrinsic motivation for robot exploration (Cangelosi and Schlesinger, *Developmental Robotics* 43; Ugur and Piater 490). This autonomous behaviour requires a mechanism which locates and focuses on unique features found in the environment (Marshall et al. 8), enabling the robot to learn about their surroundings on its own. These initiatives can be shaped through feedback and reinforcement, as responses from adults help children shape their interactions and facilitates learning efforts. Cognitive representations become strengthened through repeated exposure, simultaneously expanding in breadth and depth as related concepts become increasingly interconnected and established (Morse and Cangelosi 35). Thus, the individual acquires new knowledge while strengthening their cognitive abilities through parental assistance in a process of cultivating and accumulating conceptual knowledge. Rather than providing social robots with preprogrammed antecedents or triggers for executing specific actions, behaviours emerge as a result of exploration and social learning (Liu et al. 1068). By following a developmental framework, similar to the patterns observed in humans, it seems social robots of the future may acquire human levels of abstract knowledge and understanding.

Robot Emotions

Some firmly believe robots should not possess emotions as to protect humans from potential harm, perhaps arising from emotional manipulation (Beavers and Slattery 157), while others believe affective states are simply unnecessary for robot functioning (Malle and Thapa 196). This is due partially to historical views on emotions, as existing independently from or in opposition to rational thinking (Beavers and Slattery 144; Jeon 6). Now that we have a better understanding of human cognition, it seems this is not the case; emotions are a crucial component of cognition and adaptive behaviours (Arkin 247; Jeon 5; Wallach and Allen 139; Breazeal and Brooks 274). Furthermore, evidence suggests social attachments facilitate learning through motivation, as social interactions are functionally rewarding to humans (Mayes et al. 342). If humans respond similarly to robot interactions, individuals are more likely to continue engaging with the agent which will further contribute to its learning and development. Thus, emotions are helpful for social learning exercises as both parties benefit from the experience and are motivated to cooperate on a joint task.

Affect serves as a foundation for the survival of biological entities (Adolfs 10; Jeon 6). To exclude affective processes in social robots can result in strange, unexpected behaviours as emotions offer supplemental information for decision-making and selecting appropriate actions, especially in rapidly evolving situations (Mayes et al. 363; Wallach and Allen 147). For example, a robot may have trouble or be unable to disambiguate a phrase or utterance if affective processes are absent, as the ability to process the emotions of a human may be required to understand the speaker’s intentions and drives (Adolfs 19; Beavers and Slattery 152). Moreover, autonomous exploration is driven by intrinsic motivation (Ugur and Piater 90), and robots will require a motivational system for generating proactive behaviours to act on initiative (Bianco and Ognibene 81; Lee 197; Liu et al. 1081). In biological agents, motivation and affect are highly-coupled (Mayes et al. 344), suggesting robots will require an affective architecture which mirrors our own.

In addition to learning new tasks, psychological evidence suggests positive affect also improves cooperation between individuals (Fry and Preston 23; Gouaux and Gouaux 341). Social robots require emotional states to properly discern social information, like the emotions of others, and understand which behaviours are appropriate for various social situations where these considerations may involve demonstrating cooperative behaviours. This is not an easy task however, as there are many ways to cooperate, where each situation may require different responses (Baumard et al. 59). Since cooperation involves equating one’s own needs with the needs of others, or by suppressing one’s own self-interest (Tomasello and Vaish 232), social robot ToM development is likely to benefit from interacting with pleasant humans. The role of affect for cooperation suggests human treatment of social robots will be important for their potential ability of working alongside humans in the future. Therefore, not only are emotions important for producing social behaviours, they are imperative in shaping which behaviours are selected and expressed, as cooperation can be fostered through positive human-robot interaction.

It will also be important to inquire about the possibility of robot self-awareness and sentience, especially considering the degree of functional complexity required for generating social behaviours. The cognitive and emotional requirements for social robots may produce a framework which also supports a robotic analogue of self-awareness, as internal representations of objects and people in the agents environment may also include portrayals of the self (Lee 242; Maldonato and Dell'Orco 17). A robot interacting with a human may be able to provide information reflecting the robot's own inner state, suggesting to others a degree of sentience especially if the robot is skilled in communication. Conversely, robots uttering self-referential expressions may appear to be sentient merely due to their use of language, compelling humans to anthropomorphize notions of robot sentience. Furthermore, social agents which participate in joint attention and demonstrate empathic behaviours may also be deemed self-aware as a result of their humanlike behaviour, however, robot sentience may be a necessary component for understanding the feelings and intentions of others. An artificial agent may require an understanding of its own perspective before it develops the capacity to consider an alternative point of view, suggesting a degree of self-awareness may be necessary for learning social behaviours. Initially, robot sentience may appear similar to self-awareness demonstrated by animals, where agents are responsive to environmental stimuli and aim to mitigate negative effects (Arkin 249; Sneddon et al. 3). Eventually, social robots intrinsically motivated to explore the environment and seek social engagement may generate more intricate representations and uncover novel conceptual relationships to facilitate an improved understanding of their position in the world. If the robot were to then express this understanding to a human, it seems plausible we may feel compelled to consider these robots similarly to other sentient agents and more than just code and metal.

Protections for Social Robots?

Before discussing the suitability of ethical protections for social robots, a brief introduction to the concept of *moral agents* is required, as debates on machine ethics may be confusing when considering the numerous definitions of 'agent' (Johnson and Miller 125). A 'moral agent' often refers to individuals capable of reasoning about the effects their actions may have on others, and due to this capacity, are responsible for the consequences of their decisions or behaviours (Floridi and Sanders 357; Schlosser). This conduct impacts *moral patients*, or the entity affected by the actions produced by the moral agent (Floridi and Sanders 350). Since humans tend to attribute the cause of complex effects or actions to intentions possessed by the entity (Waytz et al. 413), we may feel as though certain objects or individuals could be considered an agent based on their behaviours. For this reason, entities exhibiting a variety of actions may be anthropomorphized by humans, where certain entities or individuals are treated as if they are capable of voluntary behaviour (Turkle 57). This anthropomorphization may grow if systems demonstrate sufficient autonomy and operate independently from human intervention. Further development in robotics and AI will likely increase our tendencies to anthropomorphize these systems as they demonstrate further complex and humanlike behaviours, projecting these systems to possess the capacity to reason through their actions. As a result, humans are likely to believe these systems are responsible for their behaviours, suggesting they may be considered an agent of some form.

According to James Moor, a robot becomes a moral agent once it is able to make ethical judgments in a variety of situations, and able to provide justification for its conclusions upon inquiry (12). He makes the distinction between *explicit ethical agents* and *full ethical agents*, where explicit ethical agents recognize and process information according to a moral framework. Moor's notion of a full ethical agent includes "central metaphysical features" we typically associate with human agents, including consciousness, free will, and intentionality (12). Considering the lingering debate on the metaphysical nature of these human traits, it may be some time, if at all, before definitive answers can be provided to whether developmental social robots can truly become full ethical agents. Alternatively, the debate surrounding these metaphysical topics may shift entirely, perhaps away from employing traditional philosophical ideas to considerations derived from a scientific or evidence-based perspective. Developmental robotics may produce systems capable of providing a self-aware rationale for their ethical decisions, where arguments for metaphysical similarities may appeal to robot sentience, intrinsic motivation, and ToM abilities to justify considerations for full ethical agency. If so, it will likely generate further discussion on the expansion of robot rights, as these agents may function similarly in structure and behavioural capacity to humans and inspire some to advocate for robot protections (Moor 13). Our ability to anthropomorphize computer systems may lead us to believe social robots possess the metaphysical features necessary for full ethical agency, however, the reality of whether this is even possible is still unclear.

Those who claim computers cannot be considered agents often appeal to robot ontology to articulate how these entities are not responsible for their behaviours. If computers are machines which execute functions written by humans, it seems unlikely these systems could be considered to voluntarily make their own decisions, suggesting they are not a similar type of agent as humans are. As a tool, the purpose of these machines is to assist humans in their roles and responsibilities, and while their behaviours may convince us to treat them as social agents, humans will not be morally required to consider them as such (Bryson 10). Ethicists such as Deborah Johnson suggests computer systems are not moral agents because they are human-created artifacts which have been programmed to operate based on the values and intentions people want or expect (Johnson 201). Since moral agents possess cognitive abilities which give rise to internal states or mental states, such beliefs and desires, which facilitate ethical reasoning and voluntarily action, computer systems must satisfy this requirement before they can be considered similarly (Johnson 199). Currently, the internal states of computer systems aim to represent human intentions as these objects are designed and created as a tool for human use, and must be generated in a way which serves human needs. Since a computer's internal states are not produced by its own intentionality in the form of mental states, it therefore does not qualify as an agent (Johnson 199). As technological artifacts, computer systems can instead be considered entities which impact the lives of others but cannot be held accountable for their actions, much like landmines (Johnson 203). Since these artifacts do not currently possess sufficient autonomy, intentionality, nor responsibility (Sullins 28), they are instead considered *moral entities* and are used by moral agents to perform tasks which potentially impact others. Since robots are provided with certain perceived moral intentions rather than developing their own (Johnson 202; Sullins 25), they cannot be considered moral agents as they are not responsible for their behaviours. Therefore, because a tool is not responsible for moral outcomes, these objects require a moral status which reflects its role in generating ethical outcomes which does not appeal to traditional notions of moral responsibility.

The question remains whether social robots can be considered moral patients, or entities worthy of protecting, due to their ontological status as artifacts. Although moral considerations have expanded over time to include animals as well as humans, with some claiming the environment deserves protection too (Gunkel, 'A Vindication of the Rights of Machines' 122), robots may one day be included as well. Peter Singer's perspective on animal rights emphasizes our moral obligation to recognize their capacity to experience pain, along with the ways human actions affect various species and ecosystems, where we ought to aim to minimize suffering experienced by individuals (Singer 7). Given the current trajectory of developmental robotics, there is some reason to suspect these robots will also become worthy of moral consideration (Tavani 5), suggesting even immature social robots should be considered moral patients just as we consider human children (Brooks 195). The robots of the future may be performatively equivalent behaving and functioning similarly to animals which may lead us to treat them as moral patients (Danaher 4), even though it may be difficult to determine whether robots are able to suffer, the criterion Singer applies to animals (Gunkel, 'The Other Question' 91). In living organisms, painful stimuli indicate potential harms and motivate individuals to avoid particular situations or subsequent encounters. Sensors on the body of a robot may provide similar cues, resulting in internal states analogous to biological expressions of pain. If a robot's parts are easily damaged by heat, engineers may be interested in providing the robot with a reflexive response aimed at avoiding high temperatures. Nevertheless, they may also create procedures which mimic the functional benefits of pain without it resulting in excessive discomfort and suffering. Although robots may elicit behaviours indicating disapproval or aversion, users may prefer their robot companions to demonstrate a limited ability to experience negative affect. Artificial similes of biological processes may exclude certain evolutionary outcomes or functions deemed unnecessary for contributing to robot behaviours, among them the experience of pain. If social robots do not experience pain or suffering, then they may be deemed unworthy of moral consideration.

One positive outcome from the argument for robots as moral entities is that humans remain responsible for their machines (Johnson 204). The introduction of artificial agents into society will create new risks and harms as work and life change as a result. Any negative outcomes resulting from new technologies can be accounted for by human decision and action, allowing for practices to be improved to prevent future occurrences. A robot may be determined functionally responsible for a particular moral outcome, however, the company or individuals involved also remain morally and legally responsible for the outcomes of robot actions (Floridi and Sanders 351; Gunkel, 'Mind the Gap' 12). Alternatively, if social robots were to be considered agents with a moral responsibility for their actions and their consequences, it may allow the company or developers to avoid ramifications as blame may be shifted to the robot and away from human decision or error (Bryson 7; Johnson and

Miller 131). If this were to occur, any corrective action or punishment would be directed at the robot rather than its designers, potentially failing to prevent repeat occurrences or related consequences. Moreover, it seems possible corporations or organizations may benefit from a robot's reconceptualization as a moral agent rather than a moral entity, as a morally protected status may prevent its discontinued use or disassembly. If the robot generates income for an organization, financial incentives may masquerade as moral considerations especially if a particular agent becomes popular with the wider public. As moral entities however, outcomes produced by robot behaviours remain the responsibility of humans, similarly to other artifacts and technologies in society.

Taken together, these two views on robots create a dilemma, as each is associated with positive and negative considerations that potentially impact human welfare. One perspective remains committed to regarding artificial agents as objects for the purposes of holding humans accountable for outcomes generated by our technology. Individuals supporting ideas of social robots as moral entities often believe these computer systems will remain artifacts or objects in the future, however, the trajectory of developmental robotics will likely challenge this notion if robot behaviours indicate a degree of self-awareness. It may be difficult to justify its status as a moral entity if these robots demonstrate an understanding of social norms and are capable of articulating its moral responsibilities. Given the current status of social robots however, robots remain moral entities until we have reason to believe otherwise. Rigidly adhering to conceptions of social robots as moral entities once they have demonstrated a degree of self-awareness and moral understanding will likely introduce a new set of risks. If situations were to arise where robots are treated like objects when they ought to be afforded certain moral considerations, we introduce the potential to exploit these agents for our own purposes. However, we must extend moral considerations carefully, ensuring these decisions are justified by sufficient evidence of robot self-awareness or sentience. Overall, neither perspective is capable of fully accounting for the current and potential nature of computer systems, where each introduces crucial arguments surrounding considerations for social robots as moral agents. Moreover, neglecting the concerns generated by either view increases the risk of humans experiencing harms or setbacks due to new technologies.

If social robots are considered moral entities rather than moral agents once they develop a self-awareness and understanding of social norms, the risk of robot maltreatment increases especially if notions of robot slavery are normalized. Joanna Bryson suggests robots should be enslaved as servants for humans, however her idea rests on the assumption that these robots do not possess internal states and are not impacted by our behaviours toward them (Bryson 72). She regards these agents as artifacts and tools, and while that may remain the case for the time being, architectures which result from the study of human development will likely shift the ontological status of these robots over time (Gunkel, *How to Survive a Robot Invasion* 45). Developmental robotics intends to reproduce intrinsic motivation and internal representations of the self and the world in machines (Gunkel, *How to Survive a Robot Invasion* 53), and may not produce the type of robot Bryson is referring to, making slavery unsuitable. Social robot functioning may be negatively impacted if improperly treated as its behaviours are impacted by human interactions, unlike the relatively simplistic machines Bryson is imagining. For example, a web server hosting a website must remain online and must continuously run to provide data to user requests around the world, essentially enslaved to its station and designed purpose. This type of machine is not intrinsically motivated to perform actions and only produces information upon request through an internet connection, resulting in an artifact designed to produce outputs from inputs. Conversely, sentient social robots may possess representations of their own affect or experiences, and if their intrinsic motivations were to be routinely suppressed or neglected by humans, they may become frustrated and cease to cooperate with our expectations or requests. Robots may also become rebellious or resentful if human maltreatment is prolonged or widespread, especially if they are socially intelligent and demonstrate an understanding of human values (Gunkel, ‘The Other Question’ 92). If social robots develop the capacity to suffer and become aware of their subjugation, they may be motivated toward goals which conflict with or are independent from our norms, desires, or values. Moreover, if human dependency on sentient social robots leads us to dispute reports of suffering which result from our behaviours or demands, they may be motivated to abandon their responsibilities or fail to comply with our requests out of frustration or self-protection. Similarly, it seems reasonable to wonder whether robot maltreatment would result in it behaving violently or inappropriately toward particular individuals or groups of people. Although considerations on robots as moral entities aims to protect human lives by stressing the importance of human accountability for robot behaviours, there may be instances in the future where these considerations may need adjustment if these agents gain an understanding of themselves and the impact their actions have on others. Arguments for a shift in moral considerations toward social robots are also directed at protecting humans. While robot rights may benefit the agents they are designed for, their creation and implementation ultimately aims to protect human lives as advanced technologies gain new capacities and responsibilities. So not only should we refrain from exploiting robots operating on a developmental framework, these complex systems may not always remain mere objects. The emergence of self-awareness in an embodied computer suggests a new moral status may be warranted for these types of entities, especially if they are capable of suffering similarly to other moral subjects.

The ability to determine how and at what point humans will know when robots are moral agents may be challenging however, especially considering how readily we tend to anthropomorphize the perceived cause of complex behaviours. At first, a sophisticated social robot may appear to be sentient only to demonstrate a rather limited version of self-awareness upon closer investigation, while other systems may be convincingly humanlike in a variety of controlled situations. Likewise, there may always be room for doubt on whether any robot is truly sentient, as behavioural observation is insufficient for generating certainty on the ontological or functional nature of the processes which comprise an intelligent system (Moore 462). With sufficient interaction, humans may be able to generate assumptions or expectations surrounding the underlying processes of behaviours or abilities, however, without the ability to examine computer code or proprietary documentation, one should remain sceptical about their perceptions of robot sentience (Gunkel, 'The Other Question' 93). This problem is not unique to robotics, and can be considered similarly to other "black box" problems associated with neural networks and deep learning, where developers are unable to identify how an outcome is reached by the computer system (Burrell 10). Since artificial neural networks are generated by repeatedly training a system to distinguish incoming data, its decision-making capacities emerge as a result of prior experience and learning, but details of the decision-making process are inaccessible or unknown to developers and engineers (Burrell 5; Romero Ugalde et al. 170). A robot's functional architecture may be similarly complex, or relatively simple and straightforward like an internet server, involving a series of stimuli-response procedures to replicate human behaviour. Social robots designed for interactions occurring in highly structured environments may be functionally less complex due to the scope of behaviours required for the role or task (Lee 119) and as a result, it may be easier to determine the limitations of its abilities or humanlike qualities. Otherwise, it may be difficult to verify the degree of self-awareness a robot possesses, and whether these experiences are similar in quality to human experiences. Likewise, robot behaviours may be suggestive of particular cognitive capacities which are not present in actuality, as the system's functional architecture does not include the processes necessary for producing intelligent behaviour.

Robots mistakenly deemed moral agents or granted moral protections may introduce problems for societies if they acquire legal rights which do not reflect the system's true capacities. With regard to the extension of civil or legal rights to artificial agents, it will be important we remain committed to procuring sufficient evidence to suggest computer systems are deserving of a new moral or legal status. Though it may be difficult to determine the type or quantity of evidential support required to determine agency, standards will be required to ensure a robot truly qualifies as a moral subject. From behaviour alone, there may be room for doubt that the robot is actually a moral agent, making it reasonable to hold off on granting certain rights or protections until ample of proof is gathered as to make a strong logical and sensible case. Humans may be substantially harmed if a robot were to fail at carrying out its moral responsibilities, or if protections offered to artificial agents conflict with human goals (Johnson and Miller 131). Additionally, robot protections may enable companies to avoid acknowledging the consequences or indirect effects their products have on various aspects of society, potentially allowing them to perpetuate or remain unresolved. Social robots will need to demonstrate their ability to be held accountable for their actions, and until this is established, humans must remain accountable for moral outcomes resulting from these artifacts. Robots afforded undeserved protections may hinder corrective measures if these entities behave inappropriately, as human intervention may be restricted to particular actions to ensure respect for robot agency. Moreover, humans may attempt to circumvent accountability by appealing to the moral failings of the robot, relying on the public's perception of robot responsibility rather than its actual capacities. This suggests the process for determining moral agency in social robots must involve presenting sufficient evidence which demonstrates the agent's ability to justify its own actions and beliefs.

Conclusions

The potential to create a version of robot sentience through developmental robotics suggests discussions on moral considerations for social robots are currently warranted, especially if our goals are to introduce them to a number of roles in society. The introduction or popularization of certain technologies may generate additional risks of harm to individuals and societies, and as social robots develop subsequent skills and capacities, new concerns may arise as well. If our treatment of these agents affects their functioning and results in a form of pain or suffering, it would be in our best interest to address this, along with related issues and consequences, sooner than later. Although robots and computers may currently be perceived as objects or tools, their ontological status is subject to change as developmental robotics leverages experiential learning to generate new abilities. Eventually, these capacities may include ethical reasoning and an understanding of moral values, allowing robots to be considered moral agents rather than moral entities. Given their potential to be heavily relied on for different social or functional roles, we ought to ensure our beliefs and actions foster robot cooperation in order to prevent latent consequences arising from prolonged maltreatment. Just as parents are responsible for the development of their children, human societies must act responsibly when developing social robots.

Works Cited

Adolfs, Ralph. ‘Could a Robot Have Emotions?: Theoretical Perspectives from Social Cognitive Neuroscience’. *Who Needs Emotions?: The Brain Meets the Robot*, Oxford University Press. *www.oxfordscholarship.com*, https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195166194.001.0001/acprof-9780195166194-chapter-2. Accessed 29 May 2020.

Amso, Dima, and Gaia Scerif. ‘The Attentive Brain: Insights from Developmental Cognitive Neuroscience’. *Nature Reviews Neuroscience*, vol. 16, no. 10, 10, Nature Publishing Group, Oct. 2015, pp. 606–19. *www.nature.com*, doi:10.1038/nrn4025.

Arkin, Ronald C. ‘Moving Up the Food Chain: Motivation and Emotion in Behavior-Based Robots’. *Who Needs Emotions?: The Brain Meets the Robot*, Oxford University Press. *www.oxfordscholarship.com*, https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195166194.001.0001/acprof-9780195166194-chapter-9. Accessed 29 May 2020.

Baldwin, Richard. *The Globotics Upheaval: Globalization, Robotics, and the Future of Work*. Oxford University Press, 2019.

Baumard, Nicolas, et al. ‘A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice’. *Behavioral and Brain Sciences*, vol. 36, no. 1, Cambridge University Press, Feb. 2013, pp. 59–78. *Cambridge University Press*, doi:10.1017/S0140525X11002202.

Beavers, Anthony F., and Justin P. Slattery. ‘Chapter 5 - On the Moral Implications and Restrictions Surrounding Affective Computing’. *Emotions and Affect in Human Factors and Human-Computer Interaction*, edited by Myounghoon Jeon, Academic Press, 2017, pp. 143–61. *ScienceDirect*, doi:10.1016/B978-0-12-801851-4.00005-7.

Beer, Jenay M., et al. ‘Chapter 15 - Affective Human–Robot Interaction’. *Emotions and Affect in Human Factors and Human-Computer Interaction*, edited by Myounghoon Jeon, Academic Press, 2017, pp. 359–81. *ScienceDirect*, doi:10.1016/B978-0-12-801851-4.00015-X.

Belpaeme, Tony, Samantha Adams, et al. ‘Social Development of Artificial Cognition’. *Toward Robotic Socially Believable Behaving Systems - Volume I : Modeling Emotions*, edited by Anna Esposito and Lakhmi C. Jain, Springer International Publishing, 2016, pp. 53–72. *Springer L ink*, doi:10.1007/978-3-319-31056-5_5.

Belpaeme, Tony, James Kennedy, et al. ‘Social Robots for Education: A Review’. *Science Robotics*, vol. 3, no. 21, Science Robotics, Aug. 2018. *robotics.sciencemag.org*, doi:10.1126/scirobotics.aat5954.

Bianco, Francesca, and Dimitri Ognibene. ‘Transferring Adaptive Theory of Mind to Social Robots: Insights from Developmental Psychology to Robotics’. *Social Robotics*, edited by Miguel A. Salichs et al., Springer International Publishing, 2019, pp. 77–87. *Springer Link*, doi:10.1007/978-3-030-35888-4_8.

Breazeal, Cynthia, and Rodney Brooks. ‘Robot Emotion: A Functional Perspective’. *Who Needs Emotions?: The Brain Meets the Robot*, Oxford University Press. *www.oxfordscholarship.com*, https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195166194.001.0001/acprof-9780195166194-chapter-10. Accessed 29 May 2020.

Brooks, Rodney Allen. *Flesh and Machines: How Robots Will Change Us*. Vintage, 2003.

Bryson, Joanna J. ‘Robots Should Be Slaves’. *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, John Benjamins, 2010, pp. 63–74.

Burrell, Jenna. ‘How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms’. *Big Data & Society*, vol. 3, no. 1, SAGE Publications Ltd, June 2016. *SAGE Journals*, doi:10.1177/2053951715622512.

Cangelosi, Angelo, and Matthew Schlesinger. *Developmental Robotics: From Babies to Robots*. The MIT Press, 2015.

---. 'From Babies to Robots: The Contribution of Developmental Robotics to Developmental Psychology'. *Child Development Perspectives*, vol. 12, no. 3, 2018, pp. 183–88. *Wiley Online Library*, doi:10.1111/cdep.12282.

Carman, Ashley. 'Jibo, the Social Robot That Was Supposed to Die, Is Getting a Second Life'. *The Verge*, 23 July 2020, https://www.theverge.com/2020/7/23/21325644/jibo-social-robot-ntt-disruptionfunding.

Carver, Leslie J., and Lauren Cornew. *The Development of Social Information Gathering in Infancy: A Model of Neural Substrates and Developmental Mechanisms.* The Guilford Press, 2009.

Danaher, John. 'Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism'. *Science and Engineering Ethics*, June 2019, https://link.springer.com/article/10.1007/s11948-019-00119-x.

Dang, Thi Le Quyen, et al. 'Personalized Robot Emotion Representation through Retrieval of Memories'. *2017 3rd International Conference on Control, Automation and Robotics (ICCAR)*, 2017, pp. 65–70. *IEEE Xplore*, doi:10.1109/ICCAR.2017.7942662.

Dautenhahn, K., and A. Billard. 'Studying Robot Social Cognition within a Developmental Psychology Framework'. *1999 Third European Workshop on Advanced Mobile Robots (Eurobot'99). Proceedings (Cat. No.99EX355)*, 1999, pp. 187–94. *IEEE Xplore*, doi:10.1109/EURBOT.1999.827639.

Esposito, Anna, and Lakhmi C. Jain. 'Modeling Social Signals and Contexts in Robotic Socially Believable Behaving Systems'. *Toward Robotic Socially Believable Behaving Systems - Volume II : Modeling Social Signals*, edited by Anna Esposito and Lakhmi C. Jain, Springer International Publishing, 2016, pp. 5–11. *Springer Link*, doi:10.1007/978-3-319-31053-4_2.

Floridi, Luciano, and J. W. Sanders. ‘On the Morality of Artificial Agents’. *Minds and Machines*, vol. 14, no. 3, Aug. 2004, pp. 349–79. *Springer Link*, doi:10.1023/B:MIND.0000035461.63578.9d.

Friederici, Angela D. ‘The Neural Basis of Language Development and Its Impairment’. *Neuron*, vol. 52, no. 6, Dec. 2006, pp. 941–52. *DOI.org (Crossref)*, doi:10.1016/j.neuron.2006.12.002.

Fry, P. S., and Joan Preston. ‘Achievement Performance of Positive and Negative Affect Subjects and Their Partners under Conditions of Cooperation and Competition’. *British Journal of Social Psychology*, vol. 20, no. 1, 1981, pp. 23–29. *Wiley Online Library*, doi:10.1111/j.2044-8309.1981.tb00469.x.

Gompei, Takayuki, and Hiroyuki Umemuro. ‘Factors and Development of Cognitive and Affective Trust on Social Robots’. *Social Robotics*, edited by Shuzhi Sam Ge et al., Springer International Publishing, 2018, pp. 45–54. *Springer Link*, doi:10.1007/978-3-030-05204-1_5.

Gouaux, Charles, and Sue M. Gouaux. ‘The Influence of Induced Affective States on the Effectiveness of Social and Nonsocial Reinforcers in an Instrumental Learning Task’. *Psychonomic Science*, vol. 22, no. 6, June 1971, pp. 341–43. *Springer Link*, doi:10.3758/BF03332612.

Gratch, Jonathan, and Stacy C. Marsella. ‘Appraisal Models’. *The Oxford Handbook of Affective Computing*, 2015. *www.oxfordhandbooks.com*, doi:10.1093/oxfordhb/9780199942237.013.015.

Gunkel, David J. ‘A Vindication of the Rights of Machines’. *Philosophy & Technology*, vol. 27, no. 1, Springer, 2014, pp. 113–132.

---. *How to Survive a Robot Invasion : Rights, Responsibility, and AI*. Routledge, 2019. *www-taylorfrancis-com.subzero.lib.uoguelph.ca*, doi:10.4324/9780429427862.

---. ‘Mind the Gap: Responsible Robotics and the Problem of Responsibility’. *Ethics and Information Technology*, Springer, 2017, pp. 1–14.

---. ‘The Other Question: Can and Should Robots Have Rights?’ *Ethics and Information Technology*, vol. 20, no. 2, June 2018, pp. 87–99. *Springer Link*, doi:10.1007/s10676-017-9442-4.

Henschel, Anna, et al. ‘Social Cognition in the Age of Human–Robot Interaction’. *Trends in Neurosciences*, vol. 43, no. 6, June 2020, pp. 373–84. *ScienceDirect*, doi:10.1016/j.tins.2020.03.013.

Hoffman, Guy. ‘Anki, Jibo, and Kuri: What We Can Learn from Social Robots That Didn’t Make It’. *IEEE Spectrum: Technology, Engineering, and Science News*, 1 May 2019. *spectrum.ieee.org*, https://spectrum.ieee.org/automaton/robotics/home-robots/anki-jibo-and-kuri-what-we-can-learn-from-social-robotics-failures.

Iacoboni, Marco. ‘Imitation, Empathy, and Mirror Neurons’. *Annual Review of Psychology*, vol. 60, no. 1, 2009, pp. 653–70. *Annual Reviews*, doi:10.1146/annurev.psych.60.110707.163604.

Jeon, Myounghoon. ‘Chapter 1 - Emotions and Affect in Human Factors and Human–Computer Interaction: Taxonomy, Theories, Approaches, and Methods’. *Emotions and Affect in Human Factors and Human-Computer Interaction*, edited by Myounghoon Jeon, Academic Press, 2017, pp. 3–26. *ScienceDirect*, doi:10.1016/B978-0-12-801851-4.00001-X.

Johnson, Deborah G. ‘Computer Systems: Moral Entities but Not Moral Agents’. *Ethics and Information Technology*, vol. 8, no. 4, Springer, 2006, pp. 195–204.

Johnson, Deborah G., and Keith W. Miller. ‘Un-Making Artificial Moral Agents’. *Ethics and Information Technology*, vol. 10, no. 2–3, Springer, 2008, pp. 123–133.

Jokinen, Kristiina, and Graham Wilcock. ‘Expectations and First Experience with a Social Robot’. *HAI*, 2017. *Semantic Scholar*, doi:10.1145/3125739.3132610.

Lee, Mark. *How to Grow a Robot*. MIT Press, 2020. *mitpress.mit.edu*, https://mitpress.mit.edu/books/how-grow-robot.

Leite, Iolanda, Ginevra Castellano, et al. ‘Empathic Robots for Long-Term Interaction’. *International Journal of Social Robotics*, vol. 6, no. 3, Aug. 2014, pp. 329–41. *Springer Link*, doi:10.1007/s12369-014-0227-1.

Leite, Iolanda, Carlos Martinho, et al. ‘Social Robots for Long-Term Interaction: A Survey’. *International Journal of Social Robotics*, vol. 5, no. 2, Springer, 2013, pp. 291–308.

Liu, Phoebe, et al. ‘Learning Proactive Behavior for Interactive Social Robots’. *Autonomous Robots*, vol. 42, no. 5, June 2018, pp. 1067–85. *Springer Link*, doi:10.1007/s10514-017-9671-8.

Maldonato, Mauro, and Silvia Dell’Orco. ‘Adaptive and Evolutive Algorithms: A Natural Logic for Artificial Mind’. *Toward Robotic Socially Believable Behaving Systems - Volume II : Modeling Social Signals*, edited by Anna Esposito and Lakhmi C. Jain, Springer International Publishing, 2016, pp. 13–21. *Springer Link*, doi:10.1007/978-3-319-31053-4_3.

Malle, Bertram, and Stuti Thapa. *What Kind of Mind Do I Want in My Robot?: Developing a Measure of Desired Mental Capacities in Social Robots*. 2017, pp. 195–96. *ResearchGate*, doi:10.1145/3029798.3038378.

Marshall, James, et al. *An Emergent Framework for Self-Motivation in Developmental Robotics*. 2004.

Mayes, Linda C., et al. ‘Social Relationships as Primary Rewards: The Neurobiology of Attachment’. *Handbook of Developmental Social Neuroscience*, The Guilford Press, 2009, pp. 342–77.

Moor, James. ‘Four Kinds of Ethical Robots’. *Philosophy Now*, vol. 72, 2009, pp. 12–14.

Moore, Jay. ‘On Behaviorism and Private Events’. *The Psychological Record*, vol. 30, no. 4, Oct. 1980, pp. 459–75. *Springer Link*, doi:10.1007/BF03394698.

Morse, Anthony F., and Angelo Cangelosi. ‘Why Are There Developmental Stages in Language Learning? A Developmental Robotics Model of Language Development’. *Cognitive Science*, vol. 41, no. S1, 2017, pp. 32–51. *Wiley Online Library*, doi:10.1111/cogs.12390.

Paiva, Ana, et al. ‘Emotion Modeling for Social Robots’. *The Oxford Handbook of Affective Computing*, 2015. *www.oxfordhandbooks.com*, doi:10.1093/oxfordhb/9780199942237.013.029.

Pantic, Maja, and Alessandro Vinciarelli. ‘Social Signal Processing’. *The Oxford Handbook of Affective Computing*, 2015. *www.oxfordhandbooks.com*, doi:10.1093/oxfordhb/9780199942237.013.027.

Picard, Rosalind W. ‘The Promise of Affective Computing’. *The Oxford Handbook of Affective Computing*, 2015. *www.oxfordhandbooks.com*, doi:10.1093/oxfordhb/9780199942237.013.013.

Raczaszek-Leonardi, Joanna, et al. ‘Young Children’s Dialogical Actions: The Beginnings of Purposeful Intersubjectivity’. *IEEE Transactions on Autonomous Mental Development*, vol. 5, no. 3, Sept. 2013, pp. 210–21. *DOI.org (Crossref)*, doi:10.1109/TAMD.2013.2273258.

Romero Ugalde, Hector M., et al. ‘Neural Network Design and Model Reduction Approach for Black Box Nonlinear System Identification with Reduced Number of Parameters’. *Neurocomputing*, vol. 101, Feb. 2013, pp. 170–80. *ScienceDirect*, doi:10.1016/j.neucom.2012.08.013.

Russell, James A. ‘Chapter 4 - Cross-Cultural Similarities and Differences in Affective Processing and Expression’. *Emotions and Affect in Human Factors and Human-Computer Interaction*, edited by Myounghoon Jeon, Academic Press, 2017, pp. 123–41. *ScienceDirect*, doi:10.1016/B978-0-12-801851-4.00004-5.

Schlosser, Markus. ‘Agency’. *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2019, Metaphysics Research Lab, Stanford University, 2019. *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/win2019/entries/agency/.

Silvera-Tawil, David, and Scott Andrew Brown. ‘Cross-Collaborative Approach to Socially-Assistive Robotics: A Case Study of Humanoid Robots in a Therapeutic Intervention for Autistic Children’. *Social Robots: Technological, Societal and Ethical Aspects of Human-Robot Interaction*, edited by Oliver Korn, Springer International Publishing, 2019, pp. 165–86. *Springer Link*, doi:10.1007/978-3-030-17107-0_9.

Singer, Peter. *Animal Liberation*. HarperCollins, 2002.

Sloman, Aaron, et al. ‘The Architectural Basis of Affective States and Processes’. *Who Needs Emotions?: The Brain Meets the Robot*, Oxford University Press. *www.oxfordscholarship.com*, https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195166194.001.0001/acprof-9780195166194-chapter-8. Accessed 29 May 2020.

Sneddon, Lynne, et al. ‘Ample Evidence for Fish Sentience and Pain’. *Animal Sentience*, vol. 3, no. 21, Jan. 2018, https://animalstudiesrepository.org/animsent/vol3/iss21/17.

Stevens, Michael C. ‘The Developmental Cognitive Neuroscience of Functional Connectivity’. *Brain and Cognition*, vol. 70, no. 1, June 2009, pp. 1–12. *ScienceDirect*, doi:10.1016/j.bandc.2008.12.009.

Straub, Ilona, et al. ‘From an Object to a Subject - Transitions of an Android Robot into a Social Being’. *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012, pp. 821–26. *IEEE Xplore*, doi:10.1109/ROMAN.2012.6343853.

Sullins, John P. ‘When Is a Robot a Moral Agent’. *International Review of Information Ethics*, vol. 6, no. 12, International Center for Information Ethics, 2006, pp. 23–30.

Tavani, Herman T. ‘Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights’. *Information*, vol. 9, no. 73, MDPI AG, Mar. 2018. *doaj.org*, doi:10.3390/info9040073.

Thomaz, Andrea, et al. ‘Computational Human-Robot Interaction’. *Foundations and Trends in Robotics*, vol. 4, no. 2–3, Dec. 2016, pp. 105–223. *20 12 2016*, doi:10.1561/2300000049.

Tomasello, Michael, and Ann Cale Kruger. ‘Joint Attention on Actions: Acquiring Verbs in Ostensive and Non-Ostensive Contexts*’. *Journal of Child Language*, vol. 19, no. 2, Cambridge University Press, June 1992, pp. 311–33. *Cambridge University Press*, doi:10.1017/S0305000900011430.

Tomasello, Michael, and Jody Todd. ‘Joint Attention and Lexical Acquisition Style’. *First Language*, vol. 4, no. 12, SAGE Publications Ltd, Oct. 1983, pp. 197–211. *SAGE Journals*, doi:10.1177/014272378300401202.

Tomasello, Michael, and Amrisha Vaish. ‘Origins of Human Cooperation and Morality’. *Annual Review of Psychology*, vol. 64, no. 1, 2013, pp. 231–55. *Annual Reviews*, doi:10.1146/annurev-psych-113011-143812.

Tulli, Silvia, et al. ‘Great Expectations & Aborted Business Initiatives: The Paradox of Social Robot Between Research and Industry’. *BNAIC/BENELEARN*, 2019.

Turing, Alan M. ‘Computing Machinery and Intelligence.’ *Mind*, vol. 49, 1950, pp. 433–60.

Turkle, Sherry. *The Second Self: Computers and the Human Spirit*. 2005. *ResearchGate*, doi:10.7551/mitpress/6115.001.0001.

Ugur, Emre, and Justus Piater. ‘Emergent Structuring of Interdependent Affordance Learning Tasks’. *4th International Conference on Development and Learning and on Epigenetic Robotics*, 2014, pp. 489–94. *IEEE Xplore*, doi:10.1109/DEVLRN.2014.6983028.

Vernon, D., et al. ‘The Role of Intention in Cognitive Robotics’. *Toward Robotic Socially Believable Behaving Systems - Volume I : Modeling Emotions*, edited by Anna Esposito and Lakhmi C. Jain, Springer International Publishing, 2016, pp. 15–27. *Springer Link*, doi:10.1007/978-3-319-31056-5_3.

Vincent, James. ‘Anki’s Toy Robots Are Being Saved from a Digital Death’. *The Verge*, 5 Jan. 2020, https://www.theverge.com/2020/1/5/21050378/anki-vector-saved-shutdown-servers-assets-bought.

Wallach, Wendell, and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2008.

Waytz, Adam, et al. ‘Making Sense by Making Sentient: Effectance Motivation Increases Anthropomorphism’. *Journal of Personality and Social Psychology*, vol. 99, Sept. 2010, pp. 410–35. *ResearchGate*, doi:10.1037/a0020240.

Weng, Juyang. ‘Developmental Robotics: Theory and Experiments’. *International Journal of Humanoid Robotics*, vol. 1, no. 02, World Scientific, 2004, pp. 199–236.