Development of Robot-enhanced Therapy for Children with Autism Spectrum Disorders

Project No. 611391

DREAM
Development of Robot-enhanced Therapy for Children with Autism Spectrum Disorders

Grant Agreement Type: Collaborative Project
Grant Agreement Number: 611391

D6.5 Self-monitoring sub-system

Due date: **31/3/2019**
Submission Date: **1/5/2019**

Start date of project: **01/04/2014** Duration: **60 months**

Organisation name of lead contractor for this deliverable: **University of Plymouth**

Responsible Person: **Tony Belpaeme** Revision: **1.0**

Project co-funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	**PU**
PP	Restricted to other programme participants (including the Commission Service)	
RE	Restricted to a group specified by the consortium (including the Commission Service)	
CO	Confidential, only for members of the consortium (including the Commission Service)	

Contents

Executive Summary

Deliverable D6.5 discusses the technical requirements for building a self-monitoring sub-system for a social robot: a system which monitors the robot's decisions and actions at a meta-level, and prevents the robot from taking actions which could cause distress or harm to the users. As the state of the art is not sufficiently mature to implement meta-level cognition for social robots, this deliverable takes the form of a white paper, discussing what technical components would be needed for an ethical social robot. This deliverable will also be submitted as paper to *IEEE Society on Social Implications of Technology*.

Principal Contributors

The main authors of this deliverable are as follows (in alphabetical order).

Tony Belpaeme, University of Plymouth
Hoang-Long Cao, Vrije Universiteit Brussel
Pablo Gómez, Vrije Universiteit Brussel
Daniel Hernández García, University of Plymouth
Kathleen Richardson, De Montfort University
Emmanuel Senft, University of Plymouth
Bram Vanderborght, Vrije Universiteit Brussel

Revision History

Version 1.0 (P.G. 01-05-2019)
Final version.

1 Towards building social robots that make ethical decisions

Robots are still largely used to assist us with physical tasks. From vacuum cleaning robots to robot arms in car factories, most robot operate in the physical environment and are not at all intended to inhabit the social world. This is about to change. As robots move away from executing constrained physical tasks, they will encounter our anthropocentric world in which perceiving, interpreting and acting in the social environment will be as important, if not more important, than operating in the physical environment. Self-driving cars not only move their charge from one location to another: when they share the road with human-driven cars, cyclist and pedestrians they will need to take the social behaviour of human road traffic into account as much as they need to localise and navigate in the physical world. When robots are explicitly designed to engage us on a social level, such as robots used in therapy, hospitality, education and retail, their functionality will be almost exclusively dictated by their ability to correctly respond the social signals and social context. Given the recent interest in ethical AI, and by extension ethical robotics, this means that robots will not only need to consider the ethics of their physical actions, but also of their social actions. Much ink has been spilled on what robots should or should not do. Asimovs three laws of robotics are almost 80 years later still used as the focal point for discussions around ethics, regulation and law for AI and robotics. And while Asimovs laws are very much flawed, illustrated by the fact that each of Asimovs stories which feature the three laws unfolds in not a very pleasant manner, the essence of the three laws somehow returns in most contemporary discussion on ethics in AI and robotics. The assumption being that AI systems and robots themselves will be able to make ethical decisions. This supposes that the human designer will not be able to build in appropriate ethical responses to every possible situation the system might encounter, and that the robot -using some form of AI- will need to make ethical decisions in uncertain or unfamiliar situations.

This paper explores what would be required to build robots that are able to make ethical decisions. Earlier work provides guidelines on how robots should behave in order to be considered ethical [Malle, 2016, Shim and Arkin, 2017, Scheutz et al., 2015, Boden et al., 2017]. In this paper, we do not intend to provide yet another set of guidelines or a legal framework defining how robots should behave or interact, instead we aim to list what capabilities and technical functions robots would need to make ethical decision. We believe this to be one of the first works to address the technical requirements for building autonomous ethical robots, especially with a view on what is possible with the current state of the art. As a guidelines we will take three typical situations from the field of Human-Robot Interaction (HRI) in which a robot would have to make a decision which can be considered an ethical judgement. These situations are only a small subset of all situations in which a robot might be forced to make an ethical decisions, nevertheless, we believe they are representative examples for our technical analysis. Section 1.1 reviews the background in the field of robot ethics. Section 1.2 presents the three situations selected for this paper and analyses them to pinpoint the specific technical functionalities required to behave ethically in these situations. Section 1.2.4 classifies and summarises these technical requirements in categories and presents the current state of the art in robotics for each category, highlighting the differences between this state of the art and the technical desiderata for making robots ethical. Finally, this gap and its impact on the fields of HRI and robot ethics are discussed in Section 1.3.

1.1 Background

Robot ethics is concerned with ethical issues arising when robots are operating in society or when their impact is felt in society. These issues started to be considered as soon as the word “robot” was

coined and are strongly related to ethics concerning machine automation and technology and the social implications of bringing technology on our society. The term 'robot ethics" covers a wide ranges of issues including the ethical systems built into robots, the ethics of people who design and use robots, and the ethics of how people treat robots [Asaro, 2006]. Within the scope of this paper, we focus on the first meaning, i.e. building robots with the ability to make ethical decisions.

Ethical dilemmas used in human psychology and ethics research have been transferred to robots too. For example, one of the archetypical problems in classical ethics is the Trolley cart problem [Foot, 1967]. In the standard version, a trolley is heading toward five people on a track, while a track branching off from the main track has only one person on it. The participant has two options: do nothing and let the trolley run on, causing five casualties, or actively intervene and change the trolley's direction and thereby have only one casualty. This problem has been extended to autonomous artificial agents and cars, and used as a main focus point to discuss ethics applied to robots. Typically, in this version, an autonomous vehicle hurtling towards a solid obstacle cannot break in time and has to decide either to kill its driver by crashing into the obstacle or swerve to avoid the obstacle, but kill one or more pedestrians instead. However, as shown in [Goodall, 2016] this problem is ill-suited for robots and autonomous vehicles in multiples ways. First it assumes that the artificial agent has perfect perception of the world, knows precisely the impact of its actions and that it has a binary choice or at best a limited number of options. However, in the real world, none of these characteristics are present (sensors are often noisy, models of the world imperfect and actions continuous). Second, it diverts the attention towards an artificial problem, drawing the debate away from more genuine and more pressing problems. Finally, the robot does not have to be ethical, if indeed might be unable to assess the ethical implications of its actions and will simply follow what the designers have hard-coded or will take action on the output of some machine learning algorithm without considering the ethical implications of its decisions.

Many guidelines have been proposed to design robots with ethical behaviours. The very first ethical rules were drafted in 1942 by Isaac Asimov which are well-known as The Three Laws of Robotics [Asimov, 1950]. However, while these laws have been used by the media and legislators as a starting point or frame for robot ethics, Asimov's stories often showed how these –at first glance– intuitive laws to constrain an artificial agents behaviour often failed to do so [Murphy and Woods, 2009]. In 2011, the "Principles of robotics" was released, a multidisciplinary effort by experts from engineering and social sciences, they propose a set of five ethical principles and seven high-level messages to mitigate the potential negative impact of integrating robots in society [Boden et al., 2017]. The IEEE Ethically Aligned Design First Edition (EAD1e, https://ethicsinaction.ieee.org) provides useful and concrete recommendations for autonomous and intelligent systems that prioritize human well-being while taking into account cultural contexts [IEEE Standards Association, 2019]. These principles are released by an open community of experts from many perspectives and countries and the IEEE expects to involve more experts to move from principles to practice. The British Standard Institute released the BS8611 standard providing guidance to the ethical design and application of robots and robotic systems (BSI 2016). This standard does not only cover potential harm but also touches on other issues such as deception, addiction, and discrimination. The European Union is moving towards developing a set of ethical guidelines for the design, production and use of robots as future European civil law rules in robotics. A study of the European Parliaments Legal Affairs Committee in 2016 analysed ethical principles applied to robotics [Nevejans, 2016]. Also in 2018, the European Group on Ethics in Science and New Technologies proposed a set of basic principles and democratic prerequisites, based on the European fundamental values [European Group on Ethics in Science and New Technologies, 2018]. Other guidelines focus on specific application domains. For example, Lin, Bekey and Abney proposed programming approaches as well as relevant ethical theories and considerations for autonomous

military robotics [Lin et al., 2008]. Goodall (2014) introduced a concept of moral behaviour for an automated vehicle in which the vehicle should continuously assess and anticipate risks to itself or others. This can be done by hard-coded rules, machine learning techniques, or hybrid approaches [Goodall, 2014]. Finally, Stahl and Coeckelbergh discussed a number of ways to embed ethics in healthcare robotics projects by means of collaboration and dialogue within the entire team consisting of roboticists, therapists and ethicists [Stahl and Coeckelbergh, 2016].

1.2 Ethical situations for social robots

As a red line throughout this paper we are using three typical human-robot interactions each presenting a situation where a robot would have to make an ethical decision. These situations are not aiming to cover the whole field of HRI, but highlight typical cases which can inform the community on the technical components required to build ethical robots.

1.2.1 Reaction to a request

The first situation concerns short term interactions. A user request something from a robot, and the robot has to react to this demand, executing the requested action, clarifying the request or refusing the request. For example, the robot should probably refuse every request harming a living being. Both from an ethical and legal perspective, requests to cause harm should not be honoured. However, while some cases are clear cut, the appropriate and ethical response can be subtle, even for simple short term requests. For example, when faced with an ambiguous demand such as “Give me a drink”, a robot could ask the human to clarify the request (e.g. “What type of drink would you like”). However, when requested to “Drive over a pedestrian” a robot should not first ask a clarification question (e.g. “Which pedestrian, the one on the left? Or on the right?”) before executing the action or reporting that this action is unethical [Jackson and Williams, 2019]. On the contrary, an ethical robot should pick up that the request is unethical and realise the implications of the request and refuse the respond outright.

To behave ethically in such situations, a robot has to interpret the intentions of its users, understand the impact of its –potentially not yet executed– actions and possess knowledge of the ethical and legal framework which should guide its behaviour. Furthermore, its reasoning process and dialogue management have to be able to handle requests differently according to their ethical implications.

1.2.2 Long-term interaction

The second situation relates to longer term interactions, where the robot’s behaviour is expected to have a lasting impact well beyond the present and near future. One such example is exposure therapies for anxiety disorders, such as phobia. In such therapies, patients are progressively exposed to an increasing level of the source of the anxiety to fight irrational anxiety and fear [Powers and Emmelkamp, 2008]. This implies that a patient will experience a high level of discomfort or even distress for a short period of time to achieve long term positive impacts. However, a simplistic ethical robot involved in this type of therapy should normally avoid distressing people and would probably have to refuse to administer the treatment. However, a robot with the ability to take a long-term perspective and matching ethical cognition should on the contrary proceed with the therapy. To behave properly in this situation, the robot has to perceive the mental, affective and emotional state of the patient, possess a model of how the patient will respond to stimuli, have a Theory of Mind (to be able to attribute mental state and believes of the people it is interacting with) [Premack and Woodruff, 1978], know the expected impacts

of its action on people and the long term goals of the interaction (such as a therapy), and have the means to refer to an expert in case of ambiguous or unsolvable ethical situation.

1.2.3 Privacy, trust and responsibility

The last situation analysed in this paper concerns manipulation, privacy, trust and disclosure of information. This topic is especially important as robots are increasingly expected to operate in situations where it has access to personal and perhaps intimate details. Similarly to virtual assistants on smart phones or embedded in a technical artefact, a social robot operating in the private sphere raises issues about privacy [Henkel et al., 2019, Fosch-Villaronga and Albo-Canals, 2018]. Which information should it be allowed to share with the company producing it? Who exactly has access to personal information? Or should the government have access to recording in crime cases? Questions which are the subject of an ongoing debate all pertain to social robots as well. We expect that when a legal framework and societal consensus evolves, these will also be relevant to social robots.

However, due to their stronger embodiment and agency, social robots raise additional questions. For example, a robot companion for a child will be in a position where its behaviour can influence the child's actions and beliefs [Vollmer et al., 2018]. Or the child might entrust the robot with confidential information, assuming the secret to be safe with the robot. The ethical responsibility of such a robot is not even clear today, should the robot bias the child's behaviour to serve a particular outcome, perhaps set by an external programme, a therapist, a teacher or a parent? Can a robot assess information disclosed by the child and make an appropriate and ethical decision on how to take action on that information? It seems likely that children will entrust such robots with secrets, and for some of them, an ethical decision might be to share them with the parents, or even report information to the authorities (for example in the case of child abuse).

While this seems futuristic, we already find ourselves in similar situations with "robotic toys" such as the "smart" barbie. The ethically appropriate response of the robot is often not clear, and will require the AI to have abilities which are beyond those available today. The robot needs to perceive the emotional state of the child, understand the implications of the information disclosed by the child, understand the potential outcomes of disclosing information. This robot should also not only decide what would be the best course of action, but also have the means to perform the required actions. At present, the best technology available is a crude keyword spotting – when a concern is raised data is often indiscriminately shared with owners or managers of the service.

1.2.4 Technical requirements

The previous section presented three situations where social robots would face ethical decisions. Each of these situation requires specific technical competencies for the robot that we grouped into the three classic categories: perception, reasoning and action. For each category, we highlight key features that robots require to behave ethically, and compare these requirements to the current state of the art in the field.

Perception To behave ethically, robots first need to perceive their environment effectively, which includes the need to have social perception. This includes the understanding of the emotional state and intention of the people it is interacting with. It also needs to be aware of the context of the interaction and perceive any changes to this context. Finally, the robot needs to analyse the state of the world and the information it is provided, through for example verbal utterances provided by the users.

While Deep Learning recently showed important progress in the fields on image recognition [LeCun et al., 2015], speech recognition or emotion recognition [Kahou et al., 2016], the level of accuracy, precision and semantic understand required to interpret a social scene with enough detail to behave ethically in still very much beyond the state of the art. Current methods suffer from simple artefacts, such as gender bias in data-driven natural language processing models [Caliskan et al., 2017] or speech recognition performing poorly for young children [Kennedy et al., 2017]. Similarly, efforts in image recognition and description focus mainly on labelling objects or providing short descriptions for visual scenes, but context inference and understanding is still very much out of reach for AI. One poignant example relevant to social robotics is the fact that emotion recognition is still based on data provided by people acting out emotion rather than from natural data. As such, emotion recognition cannot capture subtleties of affect and emotions which govern most of our day to day activities, and does not transfer to the real world. Only recently datasets have become available with video captured "in the wild" which have sufficiently rich labelling [Kollias et al., 2019]. Finally, most of the research currently done in HRI requires a controlled environment to test hypotheses, however these specific and constrained contexts limit the transfer of the results to applications in the wild [Baxter et al., 2016].

Reasoning Once a robot has access to processed inputs, it has to decide its next actions, or plans to achieve a specific goal. To be able to plan efficiently in complex social environments, robots needs precise models of the world they use to reason, both on a physical, social and societal or institutional level. Robots also needs to possess a Theory of Mind, allowing them to understand the state of their human partners, their knowledge, goals and intentions. Furthermore, robots also have to know the impact of their actions, to evaluate if some of their actions could produce physical or mental distress to their people. And finally, when interacting in domains requiring expertise, such as therapies, they have to be provided with specific domain-knowledge, such as the steps of a therapy, the specificities of their users or the goal of an interaction.

A truly ethical robot would require all these functionalities to be present, efficient and fully integrated. However, the current state of the art in the fields of planning, social modelling or dialogue management is still far from reaching these milestones [Thomaz et al., 2016]. The real world is far from deterministic, and consequently, most of the robots interacting there only have limited planning capabilities [Alterovitz et al., 2016]. Similarly, social interactions have not yet been fully understood: communication and collaboration are hard to achieve and positive results can only be achieved in highly constrained environments today [Lemaignan et al., 2017].

Action Perceiving and reasoning are not sufficient for a robot to be ethical, it needs to be able to execute a planned course of action, and as shown in the previous section, the range of actions required to do so can span a large number of modalities. As robots are increasingly using natural language, they will need to have dialogue management systems not only able to produce complex and rich language, but language that is deemed appropriate, sensitive and ethical. They will also need to move in the physical world, navigating to relevant locations and performing required physical manipulations. They would also need other type of social modalities, such as displaying appropriate emotions and social behaviours (e.g. expressing empathy), observing correct proxemics, and producing proper responses to other cultural and social norms of conduct. Finally, they might also have to use technological modalities such as phones or internet to contact remote entities, such as the authorities or the parents of a child.

While progress is made on all these actuation fields, displaying ethically appropriate behaviour for social robots is still a challenge today [Thomaz et al., 2016]. Something as simple as facial expres-

sions, for example, are still very impoverished. Popular robots used in HRI such as the Nao or Pepper have no physical actuators on their face, thus limiting their expressiveness and potentially limiting the repertoire of expressions. While this does not mean that current robots will take ethically questionable actions, the lack of expressivity might mean that the robot's intentions are misread with potentially negative consequences. A richer spectrum of expressivity should be allow to mitigate this. There is cause for optimism, for example, the problem of generating human-like voices with emotion and backchanneling seems almost resolved [Van Den Oord et al., 2016]. Additionally, physical manipulation in the real world is still an active research field, as demonstrated by the number of physical manipulation challenges being tackled across disciplines [Quispe et al., 2018].

Summary In summary, to be able to behave ethically, robots would require as yet unavailable capabilities ranging from perception over reasoning to action. Each sub part of these three main categories represents field of research on their own (such as dialogue management, or the modelling and display of affect and emotion), and are still in active development today. While progress is being made, at the time of writing we are still a long way from reaching the level required for robots to behave intrinsically ethically.

1.3 Discussion

The state of the art in robotics has not yet reached a level allowing robots to behave ethically, as illustrated by the three scenarios above. Additionally, it should be noted that while covering typical ethical situations for social robots, these three scenarios, and consequently their subsequent technical requirements, only cover a small subset of situations social robots would actually face in the wild. To build truly ethical robots, roboticists still require a number of other fields (such as natural language understanding, dialogue management, emotion display, social modelling…) to reach significant breakthroughs. However, by providing the basic tools required to generate ethical behaviours, these achievements would only represent the first step toward reaching intrinsically ethical robots. All these components will have to be integrated together, and then an ethical controller will have to be designed or trained if desired.

Despite progress in robot ethics, the question of the achievability of artificial ethics is still unanswered. As of today, the scientific community has no support that this arguably human trait can be obtained by technological means. Nevertheless, building the tools required to build ethical robots would benefit the field of HRI, and while a human level –or super-human level– of ethics might not be achievable for robots, robots could still possess sufficient ethical behaviour for selected applications. This sufficient, or perhaps even superficial, ethical behaviour is behaviour that would be produced by an agent lacking intrinsic ethical capabilities (for example a robot without Theory of Mind), but that would appear ethical in the eyes of external observers. An analogy could be made with artificial consciousness or a partial Chinese Room problem [Searle, 1980]. If a robot has a correct set of rules to define its reaction to inputs, it could mislead observers and appear intelligent, conscious or self aware without being truly any of these. For example, chatbots have been optimised by engineers to appear as human-like as possible, but most of the time they consist of stimulus-response behaviour — even if some use more complex systems based using ontologies and memory [Abdul-Kader and Woods, 2015]. While being far from straightforward, this type of superficial ethical behaviour would be much simpler to achieve than intrinsic ethics. In this case, roboticist would have to identify ethical situations and design in advance behaviour deemed ethical. However, even in this case, the exact details of what the robot should do are far from being defined. As explained in Section 1.2, many guidelines exist about designing robots [Malle, 2016, Shim and Arkin, 2017,

Scheutz et al., 2015, Boden et al., 2017], but they can be conflicting and complex to apply to each situation. Furthermore, even in these simple cases, identifying all the parameter required for making a simple decision (such as the trolley problem) are not available today.

An alternative way to help robots display ethical behaviours is to include a human in the robots action selection loop. For example, the DREAM project [Esteban et al., 2017] developed new Robot Enhanced Therapies for Children with Autism Spectrum Disorders, which had the stringent requirement to ensure that the robots actions were at all times ethically appropriate. However, due to the lack of technical features available to do so, the designers settled on keeping a therapist in the action selection loop. Using the concept of Supervised Autonomy [Esteban et al., 2017], the robot was mostly autonomous, but domain experts such as therapists at all times had the ability to override actions about to be executed if they deemed the actions to be incorrect or unethical. By keeping a human in control of the robots behaviour, they could palliate the robot's eventual suboptimal behaviour due to the missing technology, while having the robot act mostly autonomously. Alternatively, a robot identifying a situation requiring an ethical decision could delegate the decision o what to do next to a human, thus addressing the lack of intrinsic ethics. However, while scaling more easily than Supervised Autonomy, requiring robots to identify ethical situation still necessitates important sensing and reasoning capabilities and even this way of partially bypassing the technical issues will raise additional concerns, such as privacy or responsibility. This "human in the loop" controller could also be augmented with learning [Senft et al., 2015] which would allow the robot to progressively build its autonomous ethical controller. Similarly to humans learning to be ethical from other humans with more experience [Driscoll and Driscoll, 2005], having robots learn ethical behaviour from people might be in the end the best, if not the only way, for robots to reach levels of ethics on par with humans.

However, as demonstrated by the amount in ethics research for humans, as of today, it is not always clear what would be an ethical behaviour for humans. Furthermore, as demonstrated in [Malle et al., 2015] there is no guarantee that humans would expect the same behaviour from humans and robots. Robots might be expected to have a more pragmatic behaviour, while humans should be more empathetic. A last issue, is the lack of legal framework for robots to interact autonomously, and the mismatch between legislators knowledge of the field and the technical reality of building robots and social robots in specific. While not answering ethical concerns and scientific questions, providing such legal framework would at least allow roboticists to know what is expected from these autonomous robots interacting in the wild and who bears the responsibility in these complex situations.

References

[Abdul-Kader and Woods, 2015] Abdul-Kader, S. A. and Woods, J. (2015). Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, 6(7).

[Alterovitz et al., 2016] Alterovitz, R., Koenig, S., and Likhachev, M. (2016). Robot planning in the real world: research challenges and opportunities. *Ai Magazine*, 37(2):76–84.

[Asaro, 2006] Asaro, P. M. (2006). What should we want from a robot ethic. *International Review of Information Ethics*, 6(12):9–16.

[Asimov, 1950] Asimov, I. (1950). Runaround. i, robot. *New York: Bantam Dell*.

[Baxter et al., 2016] Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., and Belpaeme, T. (2016). From characterising three years of hri to methodology and reporting recommendations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 391–398. IEEE.

[Boden et al., 2017] Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., et al. (2017). Principles of robotics: regulating robots in the real world. *Connection Science*, 29(2):124–129.

[Caliskan et al., 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[Driscoll and Driscoll, 2005] Driscoll, M. P. and Driscoll, M. P. (2005). *Psychology of learning for instruction*. Pearson Allyn and Bacon Boston, MA.

[Esteban et al., 2017] Esteban, P. G., Baxter, P., Belpaeme, T., Billing, E., Cai, H., Cao, H.-L., Coeckelbergh, M., Costescu, C., David, D., De Beir, A., et al. (2017). How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder. *Paladyn, Journal of Behavioral Robotics*, 8(1):18–38.

[European Group on Ethics in Science and New Technologies, 2018] European Group on Ethics in Science and New Technologies (2018). Statement on artificial intelligence, robotics and autonomous systems. Accessed: 2018-09-18.

[Foot, 1967] Foot, P. (1967). *The problem of abortion and the doctrine of double effect*.

[Fosch-Villaronga and Albo-Canals, 2018] Fosch-Villaronga, E. and Albo-Canals, J. (2018). Robotic therapies: Notes on governance. In *Workshop on Social Robots in Therapy: Focusing on Autonomy and Ethical Challenges*.

[Goodall, 2014] Goodall, N. J. (2014). Machine ethics and automated vehicles. In *Road vehicle automation*, pages 93–102. Springer.

[Goodall, 2016] Goodall, N. J. (2016). Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, 30(8):810–821.

[Henkel et al., 2019] Henkel, Z., Baugus, K., Bethel, C. L., and May, D. C. (2019). User expectations of privacy in robot assisted therapy. *Paladyn, Journal of Behavioral Robotics*, 10(1):140–159.

[IEEE Standards Association, 2019] IEEE Standards Association (2019). Ethically aligned design.

[Jackson and Williams, 2019] Jackson, R. B. and Williams, T. (2019). Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 401–410. IEEE.

[Kahou et al., 2016] Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., et al. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111.

[Kennedy et al., 2017] Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., and Belpaeme, T. (2017). Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 82–90. ACM.

[Kollias et al., 2019] Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., and Zafeiriou, S. (2019). Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

[Lemaignan et al., 2017] Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A., and Alami, R. (2017). Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence*, 247:45–69.

[Lin et al., 2008] Lin, P., Bekey, G., and Abney, K. (2008). Autonomous military robotics: Risk, ethics, and design. Technical report, California Polytechnic State Univ San Luis Obispo.

[Malle, 2016] Malle, B. F. (2016). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, 18(4):243–256.

[Malle et al., 2015] Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 117–124. ACM.

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

[Nevejans, 2016] Nevejans, N. (2016). European civil law rules in robotics. *European Union*.

[Powers and Emmelkamp, 2008] Powers, M. B. and Emmelkamp, P. M. (2008). Virtual reality exposure therapy for anxiety disorders: A meta-analysis. *Journal of anxiety disorders*, 22(3):561–569.

[Premack and Woodruff, 1978] Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

[Quispe et al., 2018] Quispe, A. H., Amor, H. B., and Christensen, H. I. (2018). A taxonomy of benchmark tasks for robot manipulation. In *Robotics Research*, pages 405–421. Springer.

[Scheutz et al., 2015] Scheutz, M., Malle, B., and Briggs, G. (2015). Towards morally sensitive action selection for autonomous social robots. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 492–497. IEEE.

[Searle, 1980] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

[Senft et al., 2015] Senft, E., Baxter, P., Kennedy, J., and Belpaeme, T. (2015). Sparc: Supervised progressively autonomous robot competencies. In *International Conference on Social Robotics*, pages 603–612. Springer.

[Shim and Arkin, 2017] Shim, J. and Arkin, R. C. (2017). An intervening ethical governor for a robot mediator in patient-caregiver relationships. In *A World with Robots*, pages 77–91. Springer.

[Stahl and Coeckelbergh, 2016] Stahl, B. C. and Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*, 86:152–161.

[Thomaz et al., 2016] Thomaz, A., Hoffman, G., Cakmak, M., et al. (2016). Computational human-robot interaction. *Foundations and Trends in Robotics*, 4(2-3):105–223.

[Van Den Oord et al., 2016] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *SSW*, 125.

[Vollmer et al., 2018] Vollmer, A.-L., Read, R., Trippas, D., and Belpaeme, T. (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics*.