From social interaction to ethical AI: a developmental roadmap

Matthias Rolf and Nigel Crook and Jochen Steil

***Abstract*— AI and robot ethics have recently gained a lot of attention because adaptive machines are increasingly involved in ethically sensitive scenarios and cause incidents of public outcry. Much of the debate has been focused on achieving highest moral standards in handling ethical dilemmas on which not even humans can agree, which indicates that the wrong questions are being asked. We suggest to address this ethics debate strictly through the lens of what behavior seems socially acceptable, rather than idealistically ethical. Learning such behavior puts the debate into the very heart of developmental robotics. This paper poses a roadmap of computational and experimental questions to address the development of socially acceptable machines. We emphasize the need for social reward mechanisms and learning architectures that integrate these while reaching beyond limitations of plain reinforcement-learning agents. We suggest to use the metaphor of "needs" to bridge rewards and higher level abstractions such as goals for both communication and action generation in a social context. We then suggest a series of experimental questions and possible platforms and paradigms to guide future research in the area.**

I. Motivation

The recent rapid and widespread deployment of AI and robotic systems across a broad range of application domains has raised considerable ethical concern in both public and academic arenas. This concern ranges from fears about the ethical consequences of creating so-called 'super-intelligence' [1] to anxiety about the ability of autonomous vehicles to make the 'right' moral choice of who to kill in the conventional 'trolley problem' scenario [2]. There is a good deal of hype in the media about both of these scenarios, despite the fact that actual scientific and technological progress is very far from achieving either of them. The error in both cases stems from significantly over estimating the capabilities of AI and robotic technologies and severely under estimating the powers of human intelligence [3]. In the trolley problem scenario, for example, it is assumed that the autonomous vehicle would be capable of computing the consequences of all possible driving actions it could take (steering angle, acceleration, breaking), taking into account vehicle dynamics, inertia, road friction, and predicted responses of all the other road users (e.g. pedestrians diving out of the way, other vehicles on the road), all in a very short amount of time. If such computations were at all possible (and it's not clear that they are), the autonomous vehicle would then have to compute the moral implications of each outcome. The human response to such a situation, on the other hand, would most likely be to simply break and try and stop the vehicle. From a practical ethical perspective, simple breaking to minimize potential impact might be a better choice than attempting to make a rational choice of who to kill.

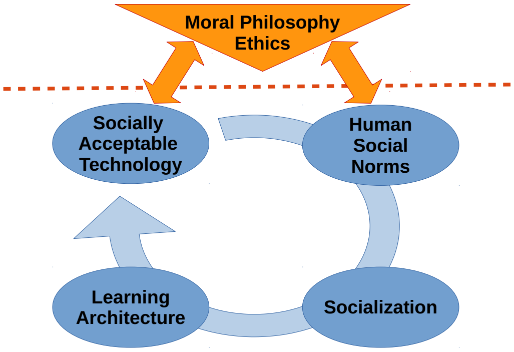

Fig. 1. We propose to address ethical decision making in adaptive robots solely through the concepts of social norms and social acceptability. A robot that is to be socially accepted has to abide by social norms. Socialization and development of such a robot would rely on social feedback signals that also guide human development, that need to be connected into learning architectures. At the end of this process should be a machine that *implicitly* addresses ethics through the eye of the social context by which it is accepted, rather than trying to achieve a universal ethical standard that we may never agree on.

Despite all this hype, there is nevertheless an increasing requirement for AI and robotic systems to be equipped with some degree of moral competence. Two factors in particular appear to be instrumental in this trend: increasing autonomy and increasing social interaction. There is considerable academic and commercial interest in making AI and robotic systems increasingly autonomous. As they become more autonomous in their actions and decision making, the likelihood is that these actions and decisions will increasingly carry ethical implications [4]. There is also considerable academic and commercial interest in developing AI and robotic systems that are increasingly embedded in social contexts in which they interact with humans [5]. In many of these situations, so called 'natural' interactions with people will require at the very least an observance of the social norms that operate in that context.

Much of the debate about the ethics of AI and robots center on establishing a commonly agreed approach to ethical decision making [6]. However, there is currently no agreement even on the overall ethical approach that should be adopted, whether that be based on a utilitarian, consequentialist or virtue ethics stance. This begs the question of how it is possible to reach agreement on the ethical competence of an AI or robotic system if agreement cannot be reached

M. Rolf and N. Crook are with the School of Engineering, Computing, and Maths at Oxford Brookes University, UK. J. Steil is with the Institute for Robotics and Process Control at Technische Universität Braunschweig, Germany.

on the fundamental nature of ethics.

In the context of this lack of agreement on the nature of ethics, there are many researchers who are taking a pragmatic approach and building AI and robotic systems with built-in ethical reasoning. Much work has been done, for example, in using deontic logics for the formal verification of moral choices in autonomous systems [7]. These approaches, however, attempt to engineer ethics into their systems, which inevitably results in brittle and narrow skills that is not capable of dealing with the breadth and complexity of social contexts in which the systems will ultimately need to operate.

Taking inspiration from the development of social and ethical competence in humans from a very early age, this paper proposes a developmental learning approach based on socially acceptable behavior. The paper presents a computational and experimental roadmap for achieving this.

II. Developing to be socially acceptable

The ultimate bar for any technology to be embedded in society is always social acceptance. Not ethics per se determines that decision. A major factor of social acceptance is the *perceived* risk [8] and whether it is acceptable with respect to the technology's benefit [9] or necessity [10]. This includes ethical risks like physical harm to people, such as discussed over decades for nuclear energy, and today discussed for lethal accidents with autonomous cars [11], and also in the generally accepted rules of risk assessment for machinery, including robots [12].

A. Social Norms

Social acceptance has much more nuance than just deliberation about immediate threats to health and life. Technology must fit into our daily lives, where it is implicitly subjected to a plethora of laws, but also social norms and values. It is, above all, important that intelligent machines obey existing laws regarding machines in general and specific application areas. A robot does not, in the first place, have to be "ethical" in any sense of cognitive or deliberate decision making – it must be *constructed* to align with law [13].

Laws, already, are specific to each country. Social norms are even more diverse and culturally specific. Examples relevant to robotics include:

- Proxemics [14]: members of different cultures have different sizes of personal space. Getting too close to someone can be perceived as invasive or threatening, but the exact distance is culture-specific.
- Offensive gestures [15]: A gesture that is frequently used for ordinary communication in one culture can be profoundly offensive in another culture.
- Backchanneling [16]: Some cultures have norms for highly active and persistent backchanneling even in the middle of another person's sentence, while most cultures value silent listening and actively discourage such interruptions.

Social norms are not only highly culture-specific, most of them are also highly context-sensitive [17]. The social norm to not interrupt people while they are talking, for example, permits numerous exceptions based on social context and roles. One could hope to hard-code all these things into a machine; the point of this article is not that this is fundamentally impossible. For that, however, the robot's decision making would have to be *hard-wired* and constrained to explicitly programmed domains. For an autonomous car it may be both viable and desirable to hard-code decision making in compliance with the law [13], specific to driving and driving only.

How, though, can a machine be ethical and socially acceptable if its very decision making is constantly changing? How can change be constrained to only socially acceptable behavior? Is this even possible when a truly intelligent robot would be capable enough to step forward into new, not explicitly programmed domains? Would a household robot thrown into a football match know to obey the commands of its teammates, but not the commands of its opponents [18]? Would a robot know that a gesture used in his intended market is offensive in another culture?

B. Learning Social Norms from Social Needs

A robot could likely not know all these things. Neither do humans, but we can learn from mistakes. If decision making is adaptive and the potential context of a system is extendable, then also its compliance to social norms has to be adaptive and extendable. The acquisition of new social norms is therefore an important goal for robot learning and AI in general.

While the debate of machine ethics suffers from the fact that there is no (and never was a) universal agreement on a code of ethics, the pitch towards social norms is inherently addressing the direct social context of a system, and acknowledging the cultural diversity that at least the perception of morality undoubtedly has [19].

The acquisition of such appropriate behavior is, however, not solved. In fact, it is largely unclear how an AI could learn to act appropriately within a social context. Clearly social norms need to be acquired by feedback from the social context. Declarative verbal instruction ("You must not ...") is one possible form of feedback, but which requires profound prior knowledge and shared concepts. Expressions of valence ("Well done!") need less prior knowledge and seem to resemble rewards, but are typically given by actual people in ways that are inconsistent with standard reinforcement learning models [20], [21]. In fact, it is highly culturally specific whether such verbal feedback is given at all in every day situations. In some cultures, breaches of social norms are rather silently "noticed and frowned upon" [22].

We suggest to take a more fundamental and, in fact, culturally independent pathway for feedback: social interaction itself. Given a machine a *need* to socially interact would imply that its learning directs towards socially conformative behavior – in the same way the need for battery charge would discourage an agent to destroy its charger. In this view social interaction corresponds to a positive reinforcement, whereas social rejection or temporary loss of social interaction is a negative reinforcement. Amoral or asocial behavior would discourage others to interact with the machine. The machine would therefore be discouraged from interrupting other for the same basic reason as it would be discouraged from physically harming others: both would result in the loss of social interaction. It is known for children that an appeal to negative social consequences is far more effective in discouraging negative behavior than non-social appeals [23], with parent and peer relations being most significant [24]. Hence, putting valence on social interaction itself, and giving also machines a need for it appears to us a fruitful way to embed ethics into adaptive machines, which is now recognized as a key issue in machine ethics [25].

C. *The need for a radically interaction-centered approach*

Social interaction capabilities are the prerequisite of ethical behavior and certainly root in the sensori-motor domain, but they do not emerge easily from the fully bottom-up incremental approach. The latter typically gets stuck in relatively simple scenarios and it has remained unclear how the required more complex behavior control for meta-cognition may be formed.

We therefore argue for an alternative radically interaction-oriented approach, which starts from endowing the robot with very basic interaction skills. These shall be embedded in an initial hypothesis of a simplified meta-cognition system that has needs, goals, and ways to explore and react to social feedback. Similar to the original sensorimotor developmental approach, where certain basic controllers e.g. to generate exploratory movement, to recognize blobs, or to detect motions are assumed, the interaction-oriented developmental approach that we propose here assume a basic set of capabilities to initiate interaction and to detect social signals.

III. A Computational Roadmap

In the tradition of developmental learning, an incremental approach that first centers on learning of basic sensorimotor skill is predominant [26], [27]. The great promise of this approach is to scale to more versatile higher-level cognitive behavior through increasing the complexity of the system. This approach is often motivated by through studying the child development and consequently is practically explored mostly with child-like humanoid robots like e.g. iCub [28] or Nao [29]. This approach has lead to a number of advancements in learning and interaction methods [30], [31], [32], [33] that are successful despite being based only on very few assumption in a learning-while-behaving approach. It has, however, not scaled to more complex cognitive behavior and there is currently no hint that the discussed question of appropriate or ethical behavior can be tackled in this way.

Enabling the learning of socially appropriate behavior in a closed loop will require a variety of different mechanisms (see Fig. 2). In this section we discuss what building blocks might be necessary and what particular challenges need to be addressed.

A. *Needs and Biases*

The root cause for socially appropriate, and eventually ethical, behavior in this proposed approach is the need for social interaction. The first challenge is therefore to detect whether a social interaction is ongoing, and thereby provide the basis for positive or negative reinforcement. This has been a focus area in developmental robotics as well as human-robot interaction research in the past, from which many possible interaction channels are already known.

A very basic sign of a potentially ongoing social interaction is seeing someone's face. Neonates have well known attentional biases towards faces [34], and facial recognition software [35] is readily available nowadays. A socially learning agent could start by simply regarding the presence of a face as a reward, and learn and maintain a course of action that leads to faces showing up repeatedly. A misbehaving agent might simply not be attended to, and therefore be deprived of its social stimulus. Similarly, the detection of human voice is a well known attentional magnet in infants [36] and readily detectable for machines [37]. Models of cross-modal synchrony [38] can further indicate how auditory and visual stimuli relate to each other. Being talked to can be regarded as a sign of social interaction in the same way as seeing a face. Neither is a strict proof of a social interaction going on, but each might be already sufficient as an initial bias towards it. Further, mutual gaze detection is used as significant cue in human-human social interaction [39] and generally a strong cue for receiving attention that has been used also in robotic systems [40].

Passive approaches like face and voice detection are complemented by models that analyze environmental stimuli in direct relation to the agent's own action. Contingency detection [41] has been argued to be a key factor in detecting social interactions by focusing on stimuli that are repeatedly observed right after an agent's actions. Other models have focused on longer-term self-other discrimination, and suggested predictive models as a key mechanism [42].

While all these cues could indicate the sheer presence of an interaction, and therefore by regarded as positive reinforcers, the valence of the interaction itself might have to be taken into account. Empathy [43] has been argued to be a key factor in social development, and could act as a modulator to the otherwise only positive nature of detecting social interaction. While someone being happy would indicate a positive reinforcement, detecting an angry face could be contribute negatively to the agent's social need. In each of these cases the social need would be fed directly from sensory input (see Fig. 2).

B. *Social Reinforcement Learning*

The next challenge is how to incorporate such signals into a learning architecture. Given that we discuss them in terms of valence (good/bad), standard reinforcement learning architectures are a possible first approach. Neuroscientific studies have indeed shown that adaption towards social conformity resembles reinforcement learning related signals in the brain [44]. Work on "social reinforcement learning" [45], [21], [20] has so far focused on making use of interaction partners in exactly the same way that a carefully, manually shaped reward function is used: get accurate feedback about

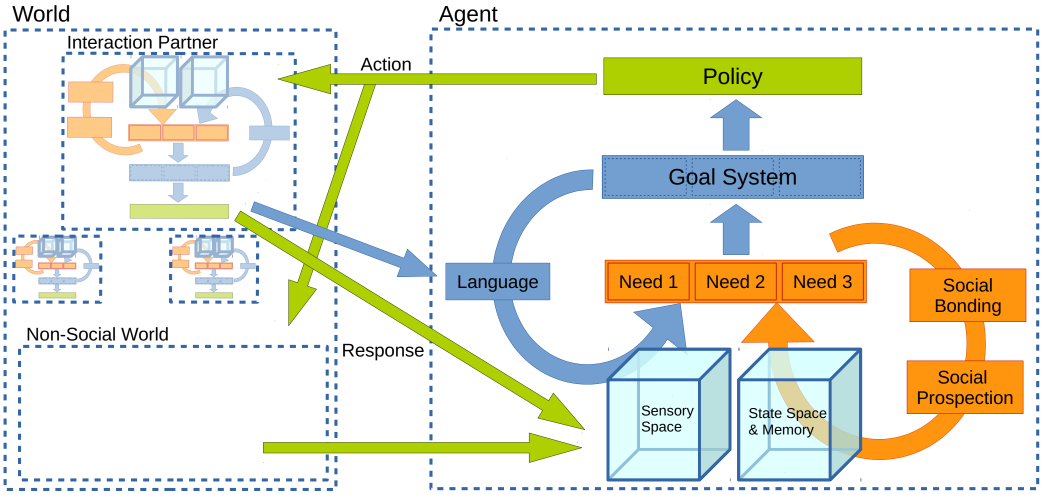

Fig. 2. Conceptual view of building blocks and interactions for a socially developing moral machine. The agent (right) interacts with the world (left) potentially including both social partners and non social objects. A set of social and internal needs motivates behavior through goal abstractions. External stimuli affect behavior through this need system, including linguistic feedback that is grounded in goal representations. Social bonding and prospection further ground the social needs through memory.

the immediate action in a repeatable way as often as possible. This is, in practice, not how people give feedback, though. People do, for example, not only give feedback about present and past action, but also "anticipatory rewards" [21] as promise for future action. Further, people do not keep positively reinforcing things that are already good [20].

A further key challenge is how to reconcile such social reward signals with internal/intrinsic needs such as self-preservation. Both internal and external factors need to be weighed at any time [18]. Standard reinforcement learning can only achieve this by combining all different needs or reward signals into a single number representing the overall reward, which is typically done by forming a weighted sum of the different components. Then, the Bellman equation dictates that rewards are linearly combined over time. Effectively, this allows to trade one need for another: getting a negative reward for running out of battery could be compensated by getting a positive reward not colliding with an obstacle. This is in practice addressed by careful manual reward shaping, but which does not the remedy the underlying formal problem. Multi-objective reinforcement learning [46] can tackle instantaneously conflicting objectives, but leaves the temporal problem. Even the worst behavior at one point in time may be compensated with future good behavior. This averaging is known be inappropriate for many applications [45], but particularly inadmissible for self-preservation and ethics. A robot must not be able to "compensate" for destructing itself or even harming a person.

C. *Needs as Proto-Goals*

Adaptive agents must overcome this flaw of established reinforcement learning if they are to behave ethically. Studies have already shown that this can be achieved for the case of self-preservation by models of fear [47], [48] that complement regular positive rewards with a disjunct system for negative rewards that can be used to selectively control risk aversion. Models of hormonal systems [49] have achieved to arbitrate more components, but require substantial hand crafting of the endocrine dynamics. Particularly useful in these system is the implementation of a numerical "need" concept that superficially looks like a reward system but has specific dynamics. In particular, a need can be satisfied to 100%, which clearly signals that the need does not need to be actively pursued at the moment and other needs can be given priority. Bare numeric rewards in a classic reinforcement learning sense do not allow to perform this test of achievement, which is otherwise only possible for explicit *goal representations* [50]. Hence, needs might already be regarded as "proto-goals" that could quickly lead to more concrete goals abstractions [51] (see Fig. 2) that could allow for efficient action planning as well as lay the ground for communication.

D. *Language*

The mechanisms discussed so far only concern feedback about present (or past) experience. If a robot behaves inappropriately, it may be able to figure out the mistake and never repeat it. For self-preservation and ethical purposes this is not sufficient. In classical reinforcement learning scenarios, simulated agents need to attempt self-destructive action to learn that it is self-destructive, such as falling off a cliff repeatedly in order to learn that cliffs are dangerous. Careful reward shaping can remedy this to some extent, for example by giving small negative results in the vicinity of dangerous states, but even then the robot has to get at least close to danger to learn about it.

An expensive physical robot cannot be exposed to risk in this way. Neither could it be allowed to learn about serious ethical failure only by committing it repeatedly. A robot must not harm a person, just in order to learn from the social reaction that harming people is unethical.

This can only be addressed by giving social feedback about *hypothetical* action without anyone performing it. This is achieved amongst humans only by means of *language*: "Do not go near the cliff". "Do not use this offensive gesture". Getting a social agent to incorporate such statements into its system would bring a significant leap towards moral machines, but is far from trivial in particular for learning machines that develop their own action repertoire.

For the social development described in this paper it is not, however, necessary to include the full complexity of speech and communication at all times. Firstly, we think it is, despite the overall developmental paradigm, admissible to include *prior knowledge* of language. Incorporating prior knowledge and biases about language would serve the same function as incorporating prior knowledge about faces or voices: to enable the eventual social and moral development. Secondly, it seems sufficient to *comprehend* key elements of linguistic feedback without the agent being able to produce speech itself. Learning to communicate with language is a highly interesting and relevant research area, but not immediately necessary for the moral scope. The latter point bares similarity for example with the socialization of intelligent pets, which may take a set of verbal instructions along with other social feedback, but do not have to speak themselves.

A specific additional challenge in this domain is that the agent's action repertoire is subject to learning. In order to allow language to refer to the agent's (hypothetical) actions, the agent has to have an *abstraction* of it and know how to connect it to an utterance of a social interaction partner. The abstraction aspect can be addressed with approaches for the learning of goal representations [52], which have been hypothesized to be necessary in cognitive system exactly for communicative purposes rather than purely action control [50]. Mapping the abstraction to a spoken word, however, does require at least some extent of linguistic learning (even though other linguistic aspects may be incorporated as prior knowledge or even omitted).

E. Social Relations

Linguistic feedback about hypothetical action eventually has to be incorporated into the same system of needs as any other social signal (Fig. 2). At this point it is not per se clear what exact valence a statement like "Do not use this offensive gesture" has. There may be an impending penalty, which may even be explicitly stated. In the scope of this paper, the more relevant consequence would be the loss of social contact after disobedience. Estimating the impact of that loss, and therefore truly closing the loop back to needs and resulting actions, would require *models of social relation*. Receiving the full impact from a sentence like "I will be very disappointed with you if you do that" is only possible if there is a specific sense of relevance of the relation with the immediate interaction partner. Generally, receiving commands as well as evaluating feedback for decision making requires awareness of social role and/or identity.

A learning agent will need to know or learn whose feedback (or commands) to incorporate, and hence give priority to some social relations over others – it would have to bond. It would need a memory (see Fig. 2) of the character and state of such relations, and learn to make predictions about the future of relations, for example based on the keyword "disappointed". Only then could the significance of "I will be very disappointed with you if you do that" truly be evaluated. Long term moral development could in these lines be achieved by cultivating an episodic memory [53] for the prospection of not only sensomotoric, but also social events.

IV. An Experimental Roadmap

A most basic interaction loop still relevant to this area could be realized in very simple way, e.g. through blinking LEDs or making noise for gaining attention and by predefined interpretation of positive of negative feedback in one or more sensory domains (acoustic, visual, touch) or by interpreting no external reaction as as a kind of negative feedback. The goal is to shape the initially rudimentary meta-control systems through learning, where we expect that both well established approaches from classical developmental learning e.g. [30], [31], [32], [33] and reinforcement learning have to be and can be transferred in this domain. We also expect that various mechanisms of classical association through conditioning will play a crucial role, where we can build on advances in dealing with delayed rewards from the neurorobotics domain [54], [55]. That is, we argue to adapt and deploy the successful ideas and algorithms of developmental learning to a higher level, where it is the meta-cognition skill of context-based skill coordination that has to be learned rather than the skill basis itself like in more sensorimotor based approaches. The role of the basic controllers here is played by socially and interactively meaningful basic interaction skills that in an ideal incremental approach themselves emerge from lower level learning. Here, however, they are assumed to be given as the latter appears more difficult than originally foreseen in the bottom-up sensorimotor developmental learning approach. This leads to a high freedom of choice for the robotic platform to be used, as interaction even with very simple devices can be in this sense meaningful as long as the have for the interaction partner cognitively and communicatively relevant actuation (e.g. loudspeakers) and respective sensors to interpret feedback. The current commercial success of such devices as ECHO is a striking example for such hardware.

A. Experimental Questions

A number of experimental questions needs to be addressed in this area. Firstly, what is the simplest actuation/action system to allow studies of social acceptance in practice? The blinking LED is the most radical simplification of an actuation system with just one binary degree of freedom. Experimentally this would only be fruitful if there would be a chance for socially "inappropriate" behavior. This is difficult but not impossible for an LED, which may better not blink glaringly in the dark while someone is trying to sleep. Possible escalations in terms of motor complexity are for example mechanical construction with a single degree

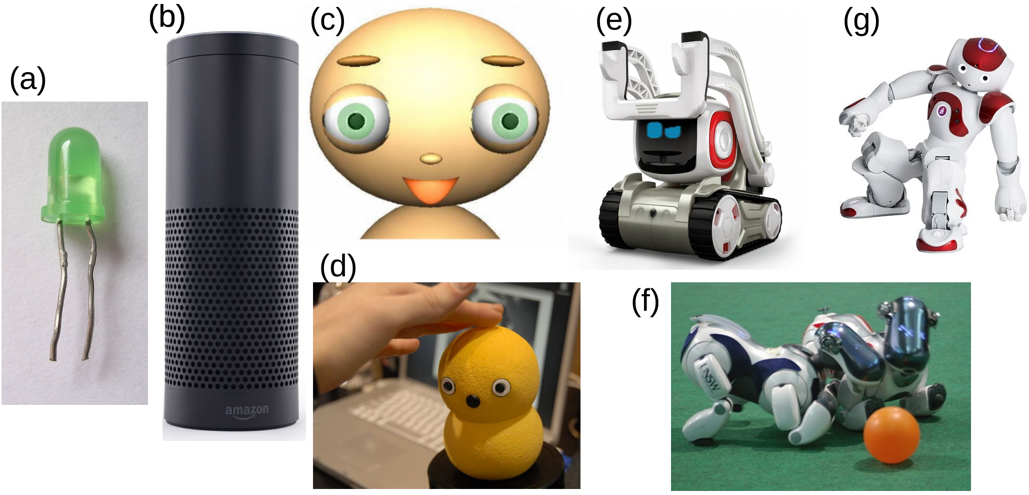

Fig. 3. A Range of possible platforms for learning of social norms, from simple to complex action capabilities: (a) An LED. (b) Amazon Echo. (c) Babyface visualization [56]. (d) Keepon [57]. (e) Anki Cozmo. (f) Sony Aibo. (g) Softbank NAO.

of freedom or beeping loudspeakers. Further escalations could be simple faces or toy-grade robots that elicit social responses without having too many degrees of freedom that need to be coordinated.

Such more complex action systems may well be necessary for experimentation on longer time scales. Apart from asking which actuation would invoke responses at all, we have to ask what actuation or design maintains social interaction after initial curiosity and excitement have faded. An LED or beeping loudspeaker may not be enough for this, but it is not clear what is generally required.

The same two questions must be addresses experimentally for the nature of social feedback signal. It is firstly not actually clear whether a need for facial stimuli alone would be enough to effectively shape behavior in a way that is agreeable to people at all. Secondly, the necessary nature of the feedback signal may change when it comes to long-term experimentation after people habituated to the system and interact differently.

At the end of such experimental runs, the loop to ethics and social acceptability should be closed. Would people judge a system as socially acceptable after they have implicitly trained its behavior? Would it depend on whether they explicitly know that they have trained it? Would they explicitly characterize the system's behavior as "moral" or "ethical"?

An interesting question that could be addressed without human long-term involvement is the possible formation of new social norms in groups of robots. Studies on the emergence of social norms [58] have so far focused on game theoretic formulations with explicitly un-social and in fact not communicated payoffs. Adding social interaction dynamics in the sense of this article could therefore shed new light also on the formation of our human social norms. A possible integration point for this scenario are language emergence studies involving robot "societies" [59], [60].

B. *Experimental Paradigms*

The proposed approach poses a number of specific experimental challenges, because the learning robot or device must be embedded in the social context for which it shall learn the respective appropriate behavior. Controlled laboratory or wizard-of-oz studies are therefore neither possible nor reasonable and the respective platform must be deployed in the wild, for instance in households or offices. Consequently, it must be extremely robust and simple, self-explicatory and intuitive to interact with.

It will also be a challenge to keep the interest of the social partners beyond an initial exploratory phase, where they will be curious to interact. But long term experiments with robots deployed for instance in elementary schools have shown that after the initial curiosity phase, interest may quickly decline [61]. The commercial success of Tamagotchi devices, however, shows that a clever design and appeal to the native tutoring intuitions of humans can motivate a lot of interaction with such devices. In a similar vein, game-oriented methods to breed and raise robotic pets have been considered [62], however, without any emphasis on social and ethical behavior, as the device is rather passive and has no elaborate meta-cognitive control system. Given this experience, it can be assumed that a robot or simpler interacting device may create enough interest to initialize the desired learning and that the increasingly versatile behavior then initiates a self-sustained progress and interest. It may, however, be advisable to create a user community to reinforce their interest, possibly in a later stage of experimentation.

Two further crucial issues need to be tackled. Evidently, cross-cultural studies would be highly desirable, where culture is meant in a rather wide sense of groups with different social norms. This can be cross-country, but also simply in different societal groups within a country or even with a single organization (e.g. students, teaching staff, administration within a university). If successful, the progress of learning should actually be an indicator of cultural differences between groups.

A further, increasingly difficult problem is of more practical nature in securing the application of a rigorous research ethics. As the approach originates in interaction with human individuals and groups, who will give feedback in various ways, data protection is of utmost importance. Monitoring the interaction for research purposes will inevitably include the recording and evaluation of sensitive information, be it biometric information or potential behavioral profiles of the interacting individuals. According to the upcoming EU General Data Protection Regulation far reaching measures way beyond the obligatory informed consent of the participants. The research procedures must ensure that participants can withdraw and have their data erased at any time and that recording data can be switched off and temporarily disabled transparently and straightforwardly. Also, occasional interaction by visitors, for instance, must not occur unless these again are fully included in the experimentation as participants. Finally, we expect that autonomous operation of the device itself is the only feasible setup, since cloud solutions with remote computing would only multiply data protection problems. Still, a systematic evaluation will require some kind of networked remote access for data accumulation and diagnosis in case of misfunctions, such that respective security concerns must be addressed.

In total, a respective experimentation will not be easy to set up and comes with a large responsibility for the researchers. However, we believe that it is worthwhile nevertheless because it may provide through route to a lot more insight how social behavior forms and can be implemented in artificial agents.

C. Platforms

Based on the previous discussion, simple platforms are both necessary for practical purposes of robustness and deployment and sufficient to test our hypothesis that a radically interaction-oriented learning approach can be feasible and leads to learning of appropriate behavior. In the extreme, already a simple diode together with a microphone could learn when to blink and when not, for e.g. cheering someone in a group communication. If unreasonable, it would be shut off temporarily for feedback. That is, we can cut off the traditional motor complexity of traditional robots and resort to much simpler designs. Fig. 3[1] shows a number of such devices, which however all would have to be equipped with sufficient computational power to execute the learning architecture sketched out in the previous sections and with some simple control board for connecting the different sensory and active channels. Fortunately, for the latter modern low cost microcontrollers are available, such that producing a respective hardware based on a simple off-the-shelf toy-like device or robot together with some home-brewed integration is not a major hurdle for implementation of the approach.

V. Ethical Machines?

How ethical can adaptive machines be? We have, in fact, argued that this may be the wrong question, given that societies cannot agree on the meaning of "ethical". What matters, anyway, is whether actual societies with all their diversity accept technology, regardless of potential philosophical standards. Considering the scope of adaptive machines, which change their decision making while in operation, then brings up how to learn to be socially acceptable. We have therefore suggested to consider social interaction itself as a positively reinforcing stimulus, and give robots social needs. This paper has outlined key computational and experimental challenges on the way to such learning capabilities. None of these challenges is entirely unique to the ethics problem. Ethics does, however, provide a unique perspective on their role and significance, and allows re-evaluate, re-prioritize, and re-think problems related to social interaction, reinforcement learning, cognitive architecture, and language.

How ethical can adaptive machines be if they are equipped with social needs? We do not know, but think that at the very least we would learn substantially about social interaction, development, and maybe society and ethics itself, by trying to find out.

References

[1] N. Boström, "Superintelligence: Paths, dangers, strategies," 2014.

[2] S. Senne, "For driverless cars, a moral dilemma: Who lives and who dies?" 2017. [Online]. Available: https://www.nbcnews.com/tech/innovation/driverless-cars-moral-dilemma-who-lives-who-dies-n708276

[3] R. Brooks, "The seven deadly sins of ai predictions," *MIT Technology Review*, vol. 120, no. 6, pp. 79–86, 2017.

[4] W. Wallach and C. Allen, *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[5] B. R. Duffy, "Fundamental issues in social robotics," *International Review of Information Ethics*, vol. 6, no. 12, 2006.

[6] J. Deigh, *An introduction to ethics*. Cambridge University Press, 2010.

[7] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal verification of ethical choices in autonomous systems," *Robotics and Autonomous Systems*, vol. 77, pp. 1 – 14, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889015003000

[8] R. E. Kasperson, O. Renn, P. Slovic, H. S. Brown, J. Emel, R. Goble, J. X. Kasperson, and S. Ratick, "The social amplification of risk: A conceptual framework," *Risk analysis*, vol. 8, no. 2, pp. 177–187, 1988.

[9] H. J. Otway and D. Von Winterfeldt, "Beyond acceptable risk: On the social acceptability of technologies," *Policy sciences*, vol. 14, no. 3, pp. 247–256, 1982.

[10] S. Bastide, J.-P. Moatti, F. Fagnani *et al.*, "Risk perception and social acceptability of technologies: the french case," *Risk analysis*, vol. 9, no. 2, pp. 215–223, 1989.

[11] P. Lin, "Tesla autopilot crash: Why we should worry about a single death," *IEEE Spectrum, Jul*, 2016.

[12] E. ISO, "12100: Safety of machinery–general principles for design–risk assessment and risk reduction (iso 12100: 2010)," 2010.

[13] B. Casey, "Amoral machines, or: How roboticists can learn to stop worrying and love the law," *Nw. UL Rev.*, vol. 111, p. 1347, 2016.

[14] M. L. Walters, K. Dautenhahn, R. Te Boekhorst, K. L. Koay, D. S. Syrdal, and C. L. Nehaniv, "An empirical framework for human-robot proxemics," *New Frontiers in Human-Robot Interaction*, 2009.

[15] D. Matsumoto and H. C. Hwang, "Cultural similarities and differences in emblematic gestures," *Journal of Nonverbal Behavior*, vol. 37, no. 1, pp. 1–27, 2013.

[1] Images (a) public domain, (b) Frmorrison CC-BY-SA-3.0, (c) IEEE [56], (d) IEEE [63], (e) public domain `https://brandfolder.com/cozmo/public`, (f) public domain, (g) Softbank Robotics Europe CC-BY-SA-4.0.

[16] S. Kita and S. Ide, "Nodding, aizuchi, and final particles in japanese conversation: How conversation reflects the ideology of communication and social relationships," *Journal of Pragmatics*, vol. 39, no. 7, pp. 1242–1254, 2007.

[17] B. F. Malle, M. Scheutz, and J. L. Austerweil, "Networks of social and moral norms in human and robot agents," in *A world with robots*. Springer, 2017, pp. 3–17.

[18] M. Rolf and N. Crook, "What if: Robots create novel goals? ethics based on social value systems." in *EDIA@ ECAI*, 2016, pp. 20–25.

[19] J. Graham, P. Meindl, E. Beall, K. M. Johnson, and L. Zhang, "Cultural differences in moral judgment and behavior, across and within societies," *Current Opinion in Psychology*, vol. 8, pp. 125–130, 2016.

[20] R. Loftin, B. Peng, J. MacGlashan, M. L. Littman, M. E. Taylor, J. Huang, and D. L. Roberts, "Learning something from nothing: Leveraging implicit human feedback strategies," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*. IEEE, 2014, pp. 607–612.

[21] A. L. Thomaz, C. Breazeal *et al.*, "Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance," in *AAAI*, 2006, pp. 1000–1005.

[22] K. Fox, *Watching the English: The Hidden Rules of English Behavior Revised and Updated*. Nicholas Brealey Publishing, 2014.

[23] P. W. Schultz, J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius, "The constructive, destructive, and reconstructive power of social norms," *Psychological science*, vol. 18, no. 5, pp. 429–434, 2007.

[24] K. A. Dodge, J. E. Lansford, V. S. Burks, J. E. Bates, G. S. Pettit, R. Fontaine, and J. M. Price, "Peer rejection and social information-processing factors in the development of aggressive behavior problems in children," *Child development*, vol. 74, no. 2, pp. 374–393, 2003.

[25] I. G. Initiative *et al.*, "Ethically aligned design v2," *IEEE Standards*, 2018.

[26] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous systems*, vol. 37, no. 2-3, pp. 185–193, 2001.

[27] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection science*, vol. 15, no. 4, pp. 151–190, 2003.

[28] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The icub humanoid robot: an open platform for research in embodied cognition," in *Proceedings of the 8th workshop on performance metrics for intelligent systems*. ACM, 2008, pp. 50–56.

[29] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic design of nao humanoid," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2009, pp. 769–774.

[30] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini, "Learning about objects through action-initial steps towards artificial cognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 3. IEEE, 2003, pp. 3140–3145.

[31] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Trans. Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[32] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling: a new concept for early sensorimotor exploration," in *Proceedings of Workshop on Developmental Robotics*, 2012.

[33] C. Teulière, S. Forestier, L. Lonini, C. Zhang, Y. Zhao, B. Shi, and J. Triesch, "Self-calibrating smooth pursuit through active efficient coding," *Robotics and Autonomous Systems*, vol. 71, pp. 3–12, 2015.

[34] M. H. Johnson, "Subcortical face processing," *Nature Reviews Neuroscience*, vol. 6, no. 10, p. 766, 2005.

[35] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[36] A. Vouloumanos, M. D. Hauser, J. F. Werker, and A. Martin, "The tuning of human neonates preference for speech," *Child development*, vol. 81, no. 2, pp. 517–527, 2010.

[37] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *IEEE Conference on Computer, Communication, Control and Power Engineering (TENCON)*, vol. 3. IEEE, 1993, pp. 321–324.

[38] M. Rolf, M. Hanheide, and K. J. Rohlfing, "Attention via synchrony: Making use of multimodal cues in social learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 55–67, 2009.

[39] N. J. Emery, "The eyes have it: the neuroethology, function and evolution of social gaze," *Neuroscience & Biobehavioral Reviews*, vol. 24, no. 6, pp. 581–604, 2000.

[40] L. Schillingmann and Y. Nagai, "Yet another gaze detector: An embodied calibration free system for the iCub robot," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 8–13.

[41] K. S. Lohan, "A model of contingency detection to spot tutoring behavior and respond to ostensive cues in human-robot-interaction," Ph.D. dissertation, Bielefeld University, 2011.

[42] Y. Nagai, Y. Kawai, and M. Asada, "Emergence of mirror neuron system: Immature vision leads to self-other correspondence," in *IEEE INT. Conf. Development and Learning (ICDL)*, vol. 2. IEEE, 2011, pp. 1–6.

[43] M. Asada, "Towards artificial empathy," *International Journal of Social Robotics*, vol. 7, no. 1, pp. 19–33, 2015.

[44] V. Klucharev, K. Hytönen, M. Rijpkema, A. Smidts, and G. Fernández, "Reinforcement learning signal predicts social conformity," *Neuron*, vol. 61, no. 1, pp. 140–151, 2009.

[45] C. Isbell, C. R. Shelton, M. Kearns, S. Singh, and P. Stone, "A social reinforcement learning agent," in *Proceedings of the fifth international conference on Autonomous agents*. ACM, 2001, pp. 377–384.

[46] P. Vamplew, J. Yearwood, R. Dazeley, and A. Berry, "On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2008, pp. 372–378.

[47] B. Seymour and S. Elfwing, "Parallel reward and punishment control in humans and robots: Safe reinforcement learning using the max-pain algorithm." in *IEEE Int. Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2017.

[48] N. Navarro-Guerrero, R. J. Lowe, and S. Wermter, "The effects on adaptive behaviour of negatively valenced signals in reinforcement learning," in *IEEE Int. Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2017.

[49] J. Lones, M. Lewis, and L. Canamero, "From sensorimotor experiences to cognitive development:: How does experiential diversity influence the development of an epigenetic robot?" *Frontiers in Robotics and AI*, 2016.

[50] M. Rolf and M. Asada, "What are goals? And if so, how many?" in *IEEE Int. Joint Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2015.

[51] ——, "Where do goals come from? A generic approach to autonomous goal-system development," 2014. [Online]. Available: http://arxiv.org/abs/1410.5557

[52] ——, "Autonomous development of goals: From generic rewards to goal and self detection," in *IEEE Int. Joint Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2014.

[53] D. Vernon, M. Beetz, and G. Sandini, "Prospection in cognition: the case for joint episodic-procedural memory in cognitive robotics," *Frontiers in Robotics and AI*, vol. 2, p. 19, 2015.

[54] A. Soltoggio, F. Reinhart, A. Lemme, and J. Steil, "Learning the rules of a game: neural conditioning in human-robot interaction with delayed rewards," in *IEEE Int. Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2013, pp. 1–6.

[55] A. Soltoggio, A. Lemme, F. Reinhart, and J. J. Steil, "Rare neural correlations implement robotic conditioning with delayed rewards and disturbances," *Frontiers in neurorobotics*, vol. 7, p. 6, 2013.

[56] C. Muhl and Y. Nagai, "Does disturbance discourage people from communicating with a robot?" in *IEEE Int. Symposium Robot and Human interactive Communication (RO-MAN)*, 2007, pp. 1137–1142.

[57] H. Kozima, M. P. Michalowski, and C. Nakagawa, "Keepon," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 3–18, 2009.

[58] S. Sen and S. Airiau, "Emergence of norms through social learning," in *IJCAI*, vol. 1507, 2007, p. 1512.

[59] L. Steels, "The synthetic modeling of language origins," *Evolution of communication*, vol. 1, no. 1, pp. 1–34, 1997.

[60] M. Spranger, *The evolution of grounded spatial language*. Language Science Press, 2016.

[61] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro, "A two-month field trial in an elementary school for long-term human–robot interaction," *IEEE Transactions on robotics*, vol. 23, no. 5, pp. 962–971, 2007.

[62] J. Rheey, "Method of breeding robot pet using on-line and off-line systems simultaneously," July 2002, uS Patent App. 09/808,119.

[63] E. Ackerman, "Beatbots releasing $40 'my keepon' robot toy," *IEEE Spectrum*, 2011.