What if: robots create novel goals? Ethics based on social value systems

Matthias Rolf and **Nigel Crook** [1]

Abstract.

Future personal robots might possess the capability to autonomously generate novel goals that exceed their initial programming as well as their past experience. We discuss the ethical challenges involved in such a scenario, ranging from the construction of ethics into such machines to the standard of ethics we could actually demand from such machines. We argue that we might have to accept those machines committing human-like ethical failures if they should ever reach human-level autonomy and intentionality. We base our discussion on recent ideas that novel goals could be originated from agents' value system that express a subjective goodness of world or internal states. Novel goals could then be generated by extrapolating what future states would be good to achieve. Ethics could be build into such systems not just by simple utilitarian measures but also by constructing a value for the expected social acceptance of a the agent's conduct.

1 Autonomous Robots

Goal-driven behavior has a long and venerable history in Artificial Intelligence and robotics. Goals are frequently used to model high level decision making and to guide low level motor control. In the field of robotics, goals have played an important part in creating robots capable of complex interactions with the environment and with humans. In the vast majority of cases, the goals which direct the behavior of robots are predefined or tightly parameterized by their designers. The practice of predefining the goals that drive robot behavior gives designers the ability to ensure that this behavior remains within agreed ethical norms. As robots become more autonomous and as they operate in increasingly complex environments, however, it becomes impractical to attempt to predefine *the* complete set of robot goals that will cover every possible circumstance the robot finds itself in. If intelligent and adaptive behavior is required of an autonomous robot in unpredictable new circumstances, then the robot will need to be equipped with the ability to create its own novel goals. This then begs the question, if a robot can create its own novel goals, how can designers ensure that these goals lead to ethical behavior from the robot? In this paper we propose an approach to novel goal creation which is based on social value systems and which, we believe, offers the best hope of generating goals that will lead to morally acceptable behavior from a robot.

To illustrate the ethical issues that arise with novel goal creation, we will briefly consider four typical robot applications: household service robots, personal assistant robots, robot pets, and teaching robots. The physical and software design of robots for each of these cases will be directed towards the creation of application specific behavior that the designers anticipate will be expected of their robots. So household service robots might be expected to clean, personal assistant robots could be required to liaise with clients, robot pets might be required to entertain children, and teaching robots could reasonably be expected to engage children in an educational task.

In each of these application areas there are two general circumstances under which robots could create their own novel goals. The first is when the owner of the robot issues an instruction to the robot which requires new behavior. The household robot, which is designed for a home environment, might, for example, be requested to go and get some cash out of the ATM at the local bank. To comply with this request the robot will need to create new goals for getting itself to the bank, including safely crossing roads, perhaps negotiating a path through crowds of people, etc. It will also need to create new goals for getting cash out of the ATM, which might include queuing up for the machine, interacting with the machine, retrieving the cash, and getting itself and the cash safely back to the home. There are complex ethical considerations (e.g. safety, social norms) involved the creation of each of these goals.

Similar examples can be found for the other three robotic applications; the teaching robot might be required to adapt to a new pedagogic approach or a different curriculum, the robot pet might need to react to another new pet (real or artificial), the personal assistant might be invited to join the company's social event (e.g. soccer match). These instructions or new requirements each involve the creation of novel goals in contexts where there are complex ethical considerations to take into account.

A significant challenge for the designers of robots that are capable of generating novel goals in response to instruction or external circumstantial requirements is in evaluating the ethical implications of those instructions or requirements. Contrast, for example, the instruction to "get cash from the bank" with "steal cash from the bank". Even when the motivation for the creation of new goals comes from an external source (e.g. the robot's owner), an ethical basis for their creation is still required.

The second general circumstance under which robots could create novel goals is when they are given the capacity to take the initiative in a given situation. This could happen, for example, if autonomous robots are endowed with the ability to recognize and *interpret* their own needs and the needs of others around them, and make autonomous decisions on how to meet those needs. The household robot might, for example, recognize that a visitor is hungry and so might decide to bake a cake for them. The robot pet might see that their human companion is lonely and so might decide to invite the companion's friend over. The teaching robot might see that a child is in danger of harm and so forms the novel goal to prevent that

[1] Oxford Brookes University, UK, email: {mrolf,ncrook}@brookes.ac.uk

harm from occurring. These are all conveniently contrived ethical responses to perceived needs. But it would be just as easy for the robot to take the initiative to do something which, *unknown to them*, would be quite unethical in response to a perceived need. The well meaning household robot might, for example, decide to cook beef burgers for their hungry visitor, who turns out to be vegetarian. The robot pet might phone an escort service for their lonely companion. The robot teacher, whilst attempting to avoid harm to one child, might unwittingly put another child in danger.

In all of these circumstances it will be expected that autonomous robots that have the capacity to behave in novel ways, will also have the capacity to recognize and take into account the ethical implications of those novel behaviors. This requires a novel goal generation mechanism that can evaluate the ethical consequences of acting on those goals.

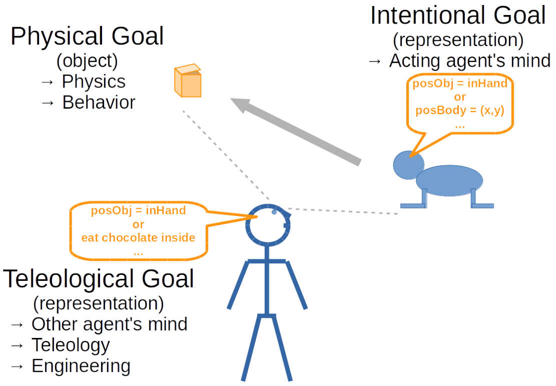

Figure 1. Physical goals are actual objects towards which behavior is directed. Intentional goals and teleological goals are representations of such end-states in the mind of an acting and observing agent respectively.

2 Origination of Novel Goals

Today's artificial agent are still extremely limited in their generation of truly novel behavior and novel goals. What even counts a novel goal is a delicate question [13]. Goals generally refer to the desired end states of action. This can be seen from three different perspectives (see Fig. 1): we may refer to the physical perspective of actual world states or objects they are referring to (such as a cake), an outside observer's teleological perspective (such as the observer explaining a robot's behavior by thinking the robot is about to make a cake), or the agent's internal, intentional perspective (such as the agent having a representation of making a cake as its goal).

What makes a goal actually novel depends on this perspective [14]. Novel physical goals simply refer to novel physical states or objects, but which do not necessarily concur with any intention of the agent. The teleological perspective is more relevant to our discussion. Novel teleological goals refer to an agent's behavior that requires a new explanation, very similar to emergent behavior [11, 5]. Looking through the eyes of a system's engineer, this would be any unforeseen or not explicitly designed behavior. This exactly describes the example scenarios we initially introduced, in which robots would generate behavior that is outside their initial design parameters. While the teleological perspective describes behavior from the outside, the intentional perspective must be considered for the agent's internal functioning, motivation, and eventually for its ethical sensitivity. Novel intentional goals are novel representation that the agent generates to steer its behavior. They describe the agent's decision making. A intentional goal could be novel because it generates an entirely new representation of something just for this purpose, or because something that as been represented already, but not immediately used for behavior control, newly becomes a goal.

Novel intentional goals are routinely created already in existing AI planning systems that are given specific goal representations from the start, and which are autonomously decomposed into sub-goals [3, 10]. Yet, such sub-goals necessarily necessarily stay within existing design parameters due to the explicitly designed initial goal. The autonomous creation of entirely novel intentional goals has been linked to notions of reward [1, 9] and reinforcement learning [12, 7]. Agents could generate novel intentional goals by predicting which states have the highest value (the prediction or future reward). This is not necessarily limited to reward or cost functions in any strictly economic or utilitarian way, but may may concern "subjective" value function that account for a variety of needs and desires. Such value functions provide the basis for (subjectively) rational behavior [17, 6], and therefore the selection of goals among known behaviors, but also allows to make predictions and extrapolations to entirely novel states that the agent has never experienced and that seem valuable.

If an agent makes such an extrapolation to a presumably value state, it takes the initiative to some new goal without explicit instruction. However, a novel goal (with respect to the agent's initial design) might also come in via an instruction such as a verbal command. In both cases, ethical considerations must take place. A robot should neither generate an unethical goal autonomously, nor adopt an instruction without ethical evaluation. In order to discuss this complex of novel goals and ethics in this article, we consider the ethical dimension to be embedded in the shaping of the value functions. Hence, we consider value functions that contain components of ethical goodness and badness of agents' conduct.

3 The need for speculation

Future robotic or AI systems that could actually generate entirely novel goals or adopt entirely goals by instruction pose a substantial challenge to machine ethics. In this article we are therefore not arguing that such machines should be built, but rather discuss possible ethical mechanisms and consequences if they would be built.

The challenge is that, by definition, novel goals take an agent into unknown territory. It has been emphasized that autonomous ethical agents first of all need to be able to predict the consequences of their behavior [16] for instance by means of internal models [18]. When an agents actually enter new territory such predictions can be grounded on general knowledge but cannot be perfectly accurate. Rather, the agent has to make informed guesses what might follow from its conduct. In human terms, it has to speculate. However, predicting the bare consequences of action is not the only problem. Also the ethical value of entirely novel behavior might not be known or at least not perfectly known to the system. When an agent enters domains that have neither be thought about at design time nor have been previously experienced by the agent, it might simply misjudge what constitutes as good or bad behavior. Again, the agent would have to make an informed guess.

No example we could give could actually prove the existence of cases in which ethical rules necessary for some behavior could not have been pre-programmed at design time — the fact that we bring up the example shows that at least we might have considered the case

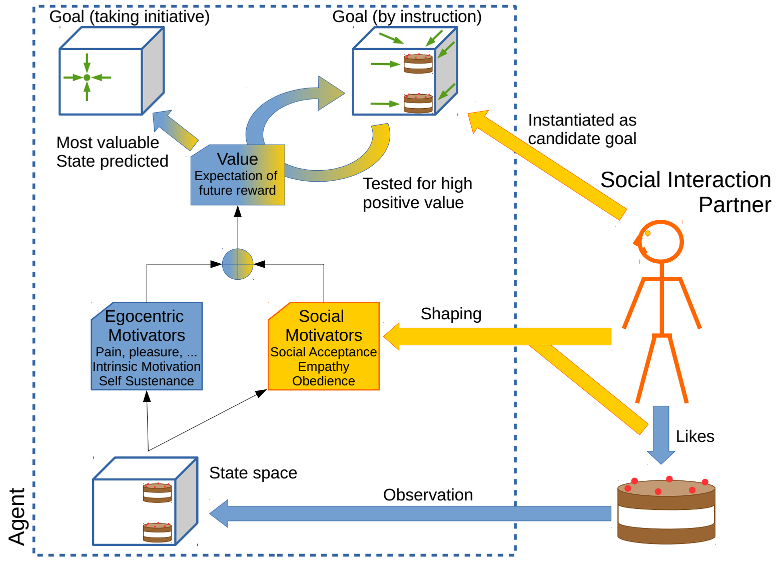

Figure 2. An agent assigns a reward/value semantic to the state space. The reward function may contain typical egocentric measures, but also social components such as the need for social acceptance (by its owner or also other people) in order to shape ethical behavior or an immediate reward for obedience to instruction. It could then generate novel goals autonomously by predicting valuable states, or take goal candidate by instruction that are then evaluated for their value. For instance, world states in which cakes exist might become a goal due to direct instruction, but also due to an empathic response to the owner's liking of cakes.

during design. Yet, one might doubt that programming an entirely comprehensive and universal ethics engine is possible. In any way, we think that the examples we discuss here show cases in which it is very *plausible* that a system built for some purpose does not possess ethical competences with respect to novel domains.

In the example of a household robot being ordered to get cash from an ATM we clearly see how such a system might lack proper prediction skills about the consequences of its action. The robot might not even have been designed to go outside and navigate there. In such a new environment it might lack skills to predict pedestrian movement or eventually the behavior of the ATM interface itself. This scenario might also come with a lack of ethical sensitivity: general ethics rules of moving through traffic and public spaces might not have been given to such a system. Even if they were given – common sense might suggest so – a purely domestic robot might not have a concept of possession and the ethical rules around it. If it is not able to withdraw cash from the ATM it might not consider it mischievous to steal it (let alone to rob it), since within its single household environment it might just take any object its owner desires.

Also the the scenario of a personal assistant robot that is asked to participate in a soccer game comprises both difficulties: both the particular world dynamics or soccer as well as ethics and morals of team sports might not be known to such a system. In particular the moral dynamics of such matches are highly non-trivial: the agent would be required to cooperate with individuals on its team, but work *against* the other team while still complying to sports specific rules.

Similarly, robots that take the initiative face uncertainties and may mis-predict both the immediate consequences as and the ethical aspects of their proactive behavior. A household robot that autonomously decides to make a cake because cakes make his owner happy might use ingredients the owner wanted to use differently, or even use ingredients a visitor is allergic to. Conversely, the robot might observe how displeased his owner is about the neighbors' barking dog, and pro-actively decide to make his owner happy by shutting the dog up – maybe injuring or killing the dog due to misjudgment of immediate consequences of its action or the ethical implications.

4 Social Value Systems for Ethics

Simple rule-based approaches to ensuring that the novel goals generated by autonomous robots result in ethically acceptable behavior are impractical for three reasons. The first is that hand-crafting a set of ethical rules that could anticipate every circumstance in which novel goals might be created by a robot is equivalent to the problem of trying to predefine a complete set of robot goals at design stage, which is against the basic premise of this paper as we have already argued.

The second reason for asserting that the rule-based ethics approach is impractical for novel goal creation is that "simple" ethical rule sets do not work well in situations where all possible actions have negative ethical consequences. The so-called 'Trolley problem' [15], which describes a scenario in which any behavioral option involves the death of humans, illustrates this issue very well. Also the examples for novel goals in this paper are full of subtleties (possession, fair-play in sports, animal rights) that can barely be stated in any compact form. The third reason that simple rule-based approaches are impractical is that as the ethical rule set increases to cover the widening set of circumstances, it becomes ever more challenging to avoid internal conflict and ensure consistency between the rules.

There are broader issues with attempting to 'design in' or hard code an ethical system for a robot when that robot may be expected to handle novel domains autonomously. One issue is that predefined ethical systems may reflect more of the ethical preferences of the designers rather than those of people who end up being subject to the robot's actions. This is especially true in cases where novel robot actions take it into new circumstances and amongst different groups of people who have distinct cultural expectations that were not anticipated by the designers. Robotics and AI literature nowadays routinely talks about agents' adaptation and learning for the prediction of unknown environments and mastering of novel skills. Then, we think it is natural that an agent must also be able to acquire novel ethical concepts and values along with those environments and skills.

All of this leads to the question - how can a robot autonomously acquire a sense of ethics for novel domains? If robots are to be 'ethical' in the eyes of those who interact with them, then they will need to be able to adapt to unwritten, socially evolved ethical preferences of the communities in which they find themselves. Human moral development provides a precedent for such adaptation [2]. We propose that novel goals along with ethics be generated on the basis of an adaptive social value system (see Fig. 2). This system is founded on both predefined egocentric motivators (e.g. self sustenance, pain, intrinsic motivation) and adaptable social motivators (e.g. empathy, the need for social acceptance) that are activated by changes in state space. The social motivators are shaped ('learnt') through interaction with the robot's social partner(s). This goes beyond simple reinforcers such as reward objects or pain, but makes social relation a direct object of internal reward signals. Hence, like humans, robots could be repelled from conducting behavior that would repel important social partners from them – and increase behavior which results in positive reactions from the social environment. The value of the activated egocentric and social motivators is estimated through an expectation of future reward signals. In the case where the robot is taking the initiative, the motivators with the highest estimated future value would be selected to form the novel goal. A household robot that has run out of instructed tasks thus might predict a happy and grateful owner, thus a positive social interaction, if only there was a cake.

In the case of an instruction ffrom the social partner, the value of the proposed candidate goal would be generated from the same mechanism of evaluating expectation of future reward of that goal on the basis of currently activated egocentric and social motivators. In this case, one of the social motivators might be obedience. We think this approach could provide a very powerful mechanism to (i) capture the subtleties of what humans perceive to be ethical conduct and (ii) allow for the acquisition of novel ethical aspects along with new environments and tasks. This would reflect a level of autonomy, capability, and adaptivity that is indeed comparable to human achievement. However, such an adaptive social approach would be subject to the same ethical flaws as have been shown to exist in humans. Classic experiments like the Milgram Experiment [8] and the Stanford Prison experiment [4] have well shown how humans can adopt or autonomously generate unethical conduct in certain social contexts.

If we ever want to – or will – bring robots to human-comparable autonomy, capability, and adaptivity, we may have to face them having human-comparable flaws. As long as universal and verifiably comprehensive rules of ethics are not in sight, we may not rule out this possibility.

REFERENCES

[1] Anthony Dickinson and Bernard Balleine, 'Motivational control of goal-directed action', *Animal Learning & Behavior*, **22**(1), 1–18, (1994).

[2] Nancy Eisenberg, 'Emotion, regulation, and moral development', *Annual review of psychology*, **51**(1), 665–697, (2000).

[3] Richard E. Fikes and Nils J. Nilsson, 'STRIPS: A new approach to the application of theorem proving to problem solving', *Artificial intelligence*, **2**(3), 189–208, (1972).

[4] Craig Haney, W Curtis Banks, and Philip G Zimbardo, 'A study of prisoners and guards in a simulated prison', *Naval Research Reviews*, **9**(1-17), (1973).

[5] W. Daniel Hillis, 'Intelligence as an emergent behavior; or, the songs of eden', *Daedalus*, **117**(1), 175–189, (1988).

[6] Thomas L. McCauley and Stan Franklin, 'An architecture for emotion', in *AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition*, (1998).

[7] Ishai Menache, Shie Mannor, and Nahum Shimkin, 'Q-cutdynamic discovery of sub-goals in reinforcement learning', in *European Conf. Machine Learning (ECML)*, (2002).

[8] Stanley Milgram, 'Behavioral study of obedience.', *The Journal of abnormal and social psychology*, **67**(4), 371, (1963).

[9] P. Read Montague, Steven E. Hyman, and Jonathan D. Cohen, 'Computational roles for dopamine in behavioural control', *Nature*, **431**, 760–767, (2004).

[10] Allen Newell and Herbert A. Simon, 'GPS, a program that simulates human thought', in *Lernende Automaten*, ed., H. Billing, Munchen: R. Oldenbourg, (1961).

[11] Timothy O'Connor and Hong Yu Wong, 'Emergent properties', in *The Stanford Encyclopedia of Philosophy (Summer 2015 Edition)*, ed., Edward N. Zalta, (2015).

[12] Matthias Rolf and Minoru Asada, 'Where do goals come from? A generic approach to autonomous goal-system development', (2014). (submitted).

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

[14] Matthias Rolf and Minoru Asada, 'What are goals? an interdisciplinary review', *Frontiers Robotics and AI*, (2016). (Under Review).

[15] Judith Jarvis Thomson, 'Killing, letting die, and the trolley problem', *The Monist*, **59**(2), 204–217, (1976).

[16] Wendell Wallach and Colin Allen, *Moral machines: Teaching robots right from wrong*, Oxford University Press, 2010.

[17] Allan Wigfield and Jacquelynne S. Eccles, 'Expectancy-value theory of achievement motivation', *Contemporary educational psychology*, **25**(1), 68–81, (2000).

[18] Alan FT Winfield, Christian Blum, and Wenguo Liu, 'Towards an ethical robot: internal models, consequences and ethical action selection', in *Advances in Autonomous Robotics Systems*, 85–96, Springer, (2014).