# The Case for Explicit Ethical Agents

*Matthias Scheutz*

■ *Morality is a fundamentally human trait that permeates all levels of human society, from basic etiquette and normative expectations of social groups, to formalized legal principles upheld by societies. Hence, future interactive AI systems, in particular, cognitive systems on robots deployed in human settings, will have to meet human normative expectations, for otherwise these system risk causing harm. While the interest in machine ethics has increased rapidly in recent years, there are only very few current efforts in the cognitive systems community to investigate moral and ethical reasoning. And there is currently no cognitive architecture that has even rudimentary moral or ethical competence, that is, the ability to judge situations based on moral principles such as norms and values and make morally and ethically sound decisions. We hence argue for the urgent need to instill moral and ethical competence in all cognitive system intended to be employed in human social contexts.*

As cognitive systems, that is, systems able to acquire and use knowledge for performing tasks and solving problems, continue to attain more intelligence and autonomy, new applications are quickly coming within reach: from sophisticated decision-support systems to broad cognitive agents (such as IBM's Watson) on the disembodied side, and from autonomously driving cars to all kinds of socially interactive robots on the embodied side. In particular the envisioned deployment of autonomous assistive robots in human societies introduces a new type of challenge that has hitherto not been sufficiently addressed by the robotics and artificial intelligence communities: that social interactions among humans in societies are based on social and moral norms that are deeply ingrained in human cognition and behavior. Failing to abide by those norms typically causes social reactions from humans, from blame and reprimands in simple cases all the way to full-fledged legal consequences. Given the fundamental normative expectations that humans have of each other, it is likely that they will extend and apply to artificial agents as well, especially to agents that are perceived as humanlike (given the human propensity to anthropomorphize machines with lifelike appearance). If true, it will be our job as agent designers to ensure that autonomous artificial agents are equipped with the moral and ethical competence to negotiate human societies in order to prevent harm they could cause otherwise by being oblivious to ethics and morality.

The science fiction writer Isaac Asimov was among a handful of visionaries who anticipated the ethical challenges of deploying autonomous robots in human societies. His well-known *Three Laws of Robotics* (Asimov 1942) were specifically designed to enable robots to operate safely in human physical and social environments, for these laws specify the fundamental societal obligations any robot has, in order of priority: (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) A robot must obey orders given it by human beings except where such orders would conflict with the First Law; and (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

From a literary perspective the three laws were ingenious, as they provided abundant materials for short stories exposing the tensions and conflicts that arise from attempts to apply the laws in different morally charged situations. From a logical point of view, however, the laws are quite problematic (for example, ordering the robot to not obey any order will cause a logical contradiction using the Second Law, since the robot cannot obey this very order); and practically speaking, they are not feasible from an implementation perspective (just consider all the interwoven moral and ethical complexities in Asimov's stories resulting from the tensions of these laws that a robot would have to disentangle, and there are many more in the real world!). Yet, there seems to be an important moral imperative buried in these laws, one of relevance outside the fictional universe: that *we should not deploy autonomous artificial agents without built-in ethical provisions*.

In a way, most if not all robots deployed in human societies nowadays can be viewed as having some rudimentary built-in ethical provisions that guide their behavior — just consider obstacle avoidance and the resultant collision-free navigation.[1] Robotic wheelchairs, tour guides, and autonomous cars are first and foremost designed to avoid collisions with objects in their environments, on the basis of which they can pursue other tasks (such as driving a user to another location, showing the crowd another exhibit in a different room, or parking themselves autonomously). These types of safety provisions are part of the algorithms implemented in the robot's architecture and thus implicit in the design of the robot in the sense that the robot does not know or cannot reason about these safety measures, nor does it need to.

Robots with such built-in ethical considerations have been called *implicit ethical agents* (Moor 2006) and need to be distinguished from explicit ethical agents, which have explicit representations about ethical principles which they can use to reason about ethical information in a variety of situations (see Bello and Bridewell 2017). Critically, explicit ethical agents can handle new situations not anticipated by their designers and make sensitive determinations about what should be done (for example, when ethical principles are in conflict, they can attempt to work out reasonable resolutions). For contexts where informing others of one's intention and reasoning is crucial, these agents could then also express their reasoning in natural language. The key question then is whether we need such explicit ethical agents or whether current implicit ethical agents are sufficient. This is important for architecture design as for implicit ethical agents no additional ethical or normative representational or reasoning apparatus is required in the architecture. Rather, all of the ethical behaviors in such agents will be the result of the interplay of existing algorithms in the architecture (for example, collision-free navigation or manipulation for robots, privacy-respecting recommendations for recommender systems). Explicit ethical agents, on the other hand, will require special representations as well as inference schemes in the cognitive system that enable their ethical competence, given that, by definition, they have explicit representations of ethical rules and principles.

## Three Applications in Need of Moral Agents

We start by considering the moral and ethical challenges involved in three pressing application areas of autonomous robots, which are no longer relegated to the domain of science fiction, but are rather an immanent reality, with prototypes already deployed: (1) the challenges of developing decision algorithms for autonomous cars about what to do when a crash is inevitable; (2) the challenge of developing assistive robots for vulnerable populations in medical and therapy domains; and (3) the challenges and potentially very harmful outcomes of building any type of social companion robot without social and moral awareness.[2]

### Autonomous Cars

Autonomously driving cars are probably the closest kind of autonomous robot with widespread impending deployment, given that prototypes of autonomous cars from various automakers are being tested on our roads already. Different from most kinds of autonomous robots, autonomous cars carry humans inside and thus create an intrinsic tie between the car's integrity and the human's well-being: if the car goes, so does the human operator. And while autonomous transportation systems are not new per se — just think of automated trains at airports, or even planes that can fly without human intervention — the difference with autonomous cars is that they are deployed in unaltered human environments, that is, our streets, which lack any additional provisions or environmental instrumentation (such as special beacons, tracks, or other guidance

systems). Moreover, owners of cars might rightfully expect that their cars have special duties towards them (as opposed to mass transit systems, which serve the community at large).

Clearly, safety is the single most important concern in the development of control algorithms for autonomous cars, both for the humans inside and those outside the car, but the current programming does not yet take important complexities of real-world situations into account. Just consider an autonomous car that could only avoid a collision with a group of pedestrians unexpectedly crossing the street in front of it by swerving into the left lane — should it do that? Clearly, there are many modulating factors that ought to figure in the decision of the autonomous car, factors that all matter to humans (for example, the likelihood of encountering cars at the given time of day on the given road at this particular location; a consideration of the number and ages of the pedestrians and the risk of injuring them when braking is attempted, compared to the risk of injuring other humans when avoiding the pedestrians, including the human operator of the car in both cases; the effects on the cars behind or the cars in the other lanes; the potential emotional repercussions to the human operator due to a decision by the car which the human would not have made). Assessing the complexities of such a situation, and understanding the tradeoffs among these various factors in order to make a decision that is morally viable for the human in the car (as the car is in a way the human's proxy), is far beyond the ability of current autonomous cars. Moreover, autonomous cars have a very different perspective of the driving environment due to their perceptual capabilities (360 degree three-dimensional [3D] lidar sensors and GPS localization in detailed metric road maps), which significantly differ from human perceptions (limited view visual 3D images reconstructed from the perceptions of two eyes). As a result, autonomous cars are able to perceive potential dangers and react to them unbeknownst to humans, leading to a different driving style that humans find unpredictable (in part because humans cannot have those percepts, in part because humans might not understand the car's control laws). Yet, neither autonomous cars nor humans have appropriate mental models of the abilities and limitations of each other that would allow them to anticipate each other's behaviors. Consequently, humans will likely find some reactions and behaviors of autonomous cars counterintuitive, even if they make sense from an algorithmic point of view.

In addition, the lack of coordination with humans through other channels (such as gestures or eye contact to indicate intent) could make autonomous cars genuine road hazards when injected into a system of human drivers, ultimately leading to destabilizing emergent properties of our whole traffic system when enough autonomous cars are present. Since the extent to which humans will be able to adapt and cope with current autonomous cars is limited, we need to develop mechanisms for those cars that make them more humanlike in their driving and decision making to minimize human harm while preserving the benefits of autonomous cars to reduce the number of overall accidents.

## Assistive Robots for Vulnerable Populations

Assistive agents, in particular, autonomous robots present different challenges, especially if intended to work with challenged and vulnerable populations such as children with autism or the elderly. Assistive agents are specifically designed to connect to humans on a social level and to support them in their activities. In the case of autistic children, robots or virtual agents are aimed at improving social awareness, social signaling, and social interactions, while robots assisting the elderly will be tasked to perform all sorts of jobs, from cleaning around the house, to preparing meals, to helping the person get dressed, to encouraging and supporting people with their exercises. While the first human-robot interactions (HRI) studies already yielded positive results in both domains (for example, see the paper by Scassellati, Admoni, and Mataric [2012] for autism and the paper by Fasola and Mataric [2013] for therapy), there is a danger lurking in these interactions: humans could develop a dependency on these robots without even noticing. And this very dependency, and some of its concomitant such as gratitude, could cause all kinds of problems. In particular, humans could develop unidirectional emotional bonds that a robot cannot reciprocate, since it is unaware of any such bonds and unequipped conceptually to understand and counteract them (Scheutz 2012). For example, an autistic child accustomed to interacting with a robot in the course of a long-term HRI study might feel cheated when the experiment — and thus the child's ability to interact with the robot — comes to an end. Or an elderly person might be tempted to engage the household robot in small talk, as would be natural for somebody with a human helper who routinely works in one's household; except that the robot will fail miserably because it does not have the ability to connect at an emotional level (for example, repeatedly telling the person that it did not understand). Either way, the robot's behavior might not be construed as a technical limitation by the human, but rather as purposeful neglect, thus causing the person emotional harm. Moreover, just as with autonomous cars, it is easy to imagine all kinds of morally challenging situations where assistive agents without moral competence will fail to make adequate decisions (for example, deciding whether to let an elderly person have a drink to improve the person's depressed mood despite the doctor's strict order for the person to not have any alcohol). And without the ability for an agent to communicate why it decided as it did, all

agent decisions — correct and incorrect — may lead to adverse reactions from the people they purportedly serve.

### Robot Companions

The third category — companion robots — is in a sense the broadest, as it includes everything from robot toys for kids (such as robot animals, robot trains, robot dolls, and others) to various robots for entertainment purposes (robot musicians and dancers, robot game players for all kinds of games, chat robots and robot helpers like Jibo or Pepper, sex robots, and others). These robots differ vastly in their capabilities and appearance (from phonelike designs on actuated pan-tilt units to human-looking androids) as well as their purposes (from embodied versions of Siri that keep track of one's appointments and answer simple queries, to robots that are intended to project human agency and intimacy through their appearance and actions). Due to their programmed or learned behaviors, it will be natural and easy for people to project agency and intelligence onto these machines, and as a result humans will automatically form expectations about the robots' abilities, including social and moral abilities: for example, that the robot doll should not tell its owner's friend all the secrets it learned from its owner.[3] Or that it was unfair for a robot soccer team to communicate telepathically to coordinate the robots' actions when playing humans. Regardless of whether the robot knew about the norm and failed to recognize its applicability, or whether it was not even capable of processing norms, any failure to apply a norm will still be considered a norm violation from the human perspective. And just as with assistive robots, humans interacting with companion robots will likely form emotional bonds with those robots during satisfying long-term interactions. Yet, they will likely not understand when the robots behave in ways different from humans (for example, failing to show signs of an emotional, empathetic expression in the face will likely be interpreted, or rather "misinterpreted" by the robot's human interlocutor as a strong signal that the robot does not care, with all the ensuing consequences).

## Enabling Ethical in Cognitive Systems

The three aforementioned application domains make a strong case for incorporating at the very least some explicit ethical mechanisms into agent control systems. For autonomous cars, it might be important to be able to at least explain and justify their decisions in cases of a close call or accident to their owner. For social robots, assistive or companion, the social aspects and human emotional dimensions will require the agent to be able to understand and use normative language (for example, words and concepts used in moral discourse about praise and blame) to some extent to be able to properly respond at the very least to blame and disapproval that will be inevitably raised by human interactants in some contexts. Moreover, to better understand human reactions and expectations, these agents might have to include sophisticated models of human moral cognition and processing to be able to see how and why humans react the way they do. Note that we distinguish ethical from moral competence, with the former referring to the proper use of principles from ethical theories, and the latter referring to humanlike abilities for moral evaluations and judgments of situations.

Overall, there are three, mutually compatible approaches towards enabling ethical competence in cognitive systems: (1) the integration of legal principles into the architecture, which would require a sufficient formalization of legal principles to render them computationally tractable in a cognitive system; (2) modeling human moral competence, which would require a sufficient understanding of how humans represent and use moral principles (such as norms and values) in perception and decision making; and (3) implementing one of the ethical theories proposed by philosophers (for example, virtue ethics, deontology, or consequentialism), which, depending on the theory, will pose its own challenges. We will next briefly examine each direction in more detail.

### Implementing Legal Theories

One obvious requirement for autonomous cognitive agents is that they abide by the legal principles governing their application realm. Hence, enabling ethical competence in a cognitive system by implementing the laws defined in a legal system (national and international law, depending on the application) seems like a natural first step. In the case of social robots, for example, these might include the "intentional torts against the person" specified in the US tort law such as false imprisonment (impeding a person's free physical movement), battery (harmful or offensive bodily contact), assault (putting someone in a position where they perceive harmful or offensive contact to be imminent, even if no battery occurs), and intentional infliction of emotional distress (extreme and outrageous conduct that causes severe distress). From these the legal definitions of these torts, one could then extract the necessary aspects for a robot to determine when harmful contact and thus battery would occur (for example, see the paper by Mikhail [2014]). This not only poses a challenge for the robot's perceptual system, but also for the cognitive architecture's inferential system, for determining harmful contact is further exacerbated by aspects of implicit consent, which often depends on circumstances and the legal notion of what a rational person would do. Hence, the effort would have to include making legal terms such as "intent" or "imminent"

or "distress" computational, that is, provide algorithms that detect intent, perceptions of imminence, or distressed emotional states, in addition to formalizing the legal concept of a "rational person" (to be able to use it in cases where the law specifically refers to the decisions and actions performed by a rational person). There are currently no proposals for any of these that could be directly realized within a cognitive system.

## Implementing Humanlike Moral Competence

In addition to implementing legal concepts and interpretations, it will be useful for an artificial agent to model (at least to some extent) human morality. If rendered computational, processes involved in human moral cognition could be used by the agent both for itself (for example, to generate human understandable justifications for why it did what it did) and for predicting and making sense of human behavior (for example, why a human acted a certain way in a situation with moral conflicts). Such an approach would require a sufficient computational understanding of core human moral competence (Malle and Scheutz 2014), which includes how humans learn, represent, and reason with moral norms (Malle, Scheutz, and Austerweil 2015), and how they detect, violate norms themselves, and respond to norm violations from others. However, attempting to implement humanlike moral competence is very challenging, for it is not even clear in the human case what perceptual, cognitive, affective, communicative and behavioral components underwrite human moral competence. For example, is it necessary to be able to simulate another person's decision making in order to be able to judge whether that person behaved morally even though the person committed a norm violation (assuming that all norms can be violated without consequences under certain circumstances, for example, Gert [2005])? Moreover, there are additional important ethical questions as to whether we should attempt to replicate human morality in a machine (for example, because human moral performance can be suboptimal or irrational at times, and we would ideally expect robots to be morally superior, that is, show supererogatory performance [Arnold and Scheutz 2016]).

## Implementing Ethical Theories

Last but not least, it has been suggested (for example, Gips [1995]) that we follow one of the three major philosophical ethical theories to guide the development of ethical agents: virtue ethics, deontology, or consequentialism. At the core of virtue ethics is the idea that ethical thought and action is guided by a person's character, which is constituted by virtues such as wisdom, courage, temperance, and justice, even though there is no agreement on a core set of virtues. In some cases, virtues could be directly implemented in cognitive systems, and, in fact, Moor specifically links implicit ethical agents to virtue ethics saying that "implicit ethical agents have a kind of built-in virtue — not built-in by habit but by specific hardware or programming" (Moor 2009). As an example, consider an autonomous car that decides to take a risky maneuver that might make it crash into parked cars in order to prevent a collision with a pedestrian — this propensity to purposefully take risk could be viewed as courage, although it is debatable whether a solid notion of courage is reducible to such dispositions. And it is even less clear how other virtues such as wisdom might be realized as they seem to be systemic properties of cognitive systems rather than reducible to single processes or even numeric values.

In contrast, deontology lends itself much more directly to computational implementations as it is intrinsically rule based. Following Gert (2005), for example, one could define ethical behaviors in terms of general rules like "don't kill," "don't cause pain," "don't deceive," "obey the law," which apply quite generally across contexts, but might be difficult to integrate computationally so that when such a rule is triggered, it will result in meaningful executable behavior (for example, a rule like *obey* might require complex perceptual and inference skills in order to determine that a certain action might constitute the violation of a legal principle and should thus not be performed). Aside from questions about how to select appropriate principles, various formal systems (for example, based on deontic or linear temporal logic) are available for representing such rules and allowing for rigorous logical inference.

Finally, consequentialism as an ethical theory based on utilitarianism best meshes with thinking in AI and cognitive systems, for ethical decisions are those that maximize expected utility (possibly generated from moral values). Algorithms based on expected utility are found in almost all areas of AI and robotics, including cognitive systems (for example, production selection in ACT-R is based on utility calculations (see also Laird, Lebiere, and Rosenbloom [2017]). Different from typical decision-making algorithms implemented in artificial agents, consequentialism considers "the overall good," that is, the utility for all agents, and thus cannot be realized by approaches that solely compute an individual agent's expected utility. As a result, following consequentialist theories requires artificial agents to cope with the limitations an agent has knowing how good an action will be for others, how many others to take into considerations, and so on. And, moreover, there are open questions on whether moral values can be easily mapped onto utilities (as assumed by Russell, Dewey, and Tegmark [2015]) or whether they have to be treated differently (as argued in Scheutz [2016]).

Overall, a main challenge associated with imple-

menting philosophical ethical theories is that there is not even consensus among philosophers, let alone the wider intellectual community, about which approach is the normatively correct one. And since the different theories sometimes make different recommendations for how to act in certain situations, implementing any particular philosophical theory will automatically require the agent designer to also take a philosophical moral stance.

## Ethical and Moral Processing in Cognitive Systems

All three main conceptual approaches to enabling ethical and moral processing in a cognitive system have their own architectural requirements, which may or may not be present in a give cognitive system, thus requiring the architecture designer to make several important decisions: (1) whether to use existing mechanisms in an architecture and implement ethical processing on top of them or whether to add new specialized mechanisms (for example, if the existing mechanisms are not easily amenable to ethical extensions); (2) whether existing data representations (for example, condition-action rules) can be utilized for ethical processing or whether additional representations need to be employed (for example, for deontic rules); and in the case of cognitive architectures attempting to model human cognition (see also Laird, Lebiere, and Rosenbloom [2017]), (3) whether the full spectrum of human moral competence (à la Malle and Scheutz [2014]) is to be realized, or only select features (for example, moral reasoning, but not perception, language).

There are currently only a few approaches for moral processing in the cognitive systems community, most of which are based on reusing existing representations and processing mechanisms in the cognitive architecture. For example, the Companion cognitive architecture uses existing analogical inference (Dehghani et al. [2008]; Blass and Forbus [2015]) (see also Forbus and Hinrichs [2017]) to learn morally appropriate decisions in new dilemmalike contexts based on known ones; or the Icarus architecture uses existing mechanisms to explain what people do and why they do it in morally charged situations (Iba and Langley 2011).

There are also some hybrid approaches that extend cognitive architectures by additional inference mechanisms that can be applied specifically to moral inference such as extending the Clarion cognitive architecture by analogical reasoning) (Licato, Sun, and Bringsjord 2014).[4] One of the most prominent proposals for extending a hybrid robotic architecture is the AuRA architecture (Arkin and Balch 1997), which adds an ethical governor, a responsibility advisor, and an ethical adaptor to the system to allow for modifications of the robot's behavioral repertoire in case unethical behaviors are observed. Specifically, the ethical adaptor uses a scalar guilt value that monotonically increases over time as unanticipated ethical violations are detected by the system (Arkin and Ulam 2009). As a result, actions with harmful potential are subsequently disallowed. The current system can only handle very specific, hard-coded moral decisions, however, but it can advise human operators ahead of a mission about possible ethical conflicts (Arkin, Wagner, and Duncan 2009). Yet, the architecture does lack the formal representations of norms, principles, values, to allow it to perform general ethical inferences and reason through normative conflicts.

Similarly, the additional mechanisms proposed for the cognitive robotic DIARC architecture (Scheutz et al. 2007) can detect potential norm violations that would result from carrying out human instructions that are in conflict with given normative principles (Briggs and Scheutz 2015, 2013). In that case, the robot can engage the human operator in a brief dialogue about why it is not permitted to carry out an instruction and offer a justification for its refusal (see also McShane [2017]). Different from the ethical extensions to the AuRA architecture, the DIARC extension is based on general inference algorithms that work with explicit representations of normative principles. However, the current system can handle only simple potential, but no actual norm conflicts (that is, conflicts that could arise if it were to follow a particular command and execute an action that would be in conflict with its existing principles). A recent proposal for systematically handling norm conflicts in stochastic environments with norms being expressed in linear temporal logic is still in need to be incorporated into DIARC's goal and action manager.

## Discussion

The three example cases of artificial agents facing morally difficult decisions were intended to demonstrate that implicit ethical agents are insufficient for handling all morally charged situations they may encounter, at least if they are to do so correctly. Rather, because our world is open and new morally charged situations can arise anytime, robots will need mechanisms analogous to humans to deal with the situational openness and unpredictability of human societies: they need to be explicit ethical agents, able to represent, learn and reason with norms and values in much the same way humans do, albeit to different degrees depending on their application domain.

What can go awry when agents have no notion of moral norm or moral value while using unconstrained machine learning is best demonstrated with Microsoft's Twitter bot *Tay,* which was taught all kinds of racial slurs by Internet users attempting to tease out the agent's learning abilities. While this example was easily solved (the agent was taken

offline very quickly), it is easy to imagine other cases of autonomous agents neglecting human moral expectations with much longer-lasting and perhaps even more severe impact. Part of the problem is that unconstrained machine learning, for example, inverse reinforcement learning, does not automatically lead to learning moral values as some would like to have it (for example, Russell et al. [2015]).[5] Moreover, there is an argument to be made that moral values are not just utilities, and that moral decision making should thus not be treated on a par with all other decisions an agent must make.

By making ethical principles explicit in a cognitive system, it is possible to treat them differently from other principles that govern the agent's behavior (for example, task-based decision making) and thus enable a type of ethical and moral reasoning that is accessible to introspection and allows for performance guarantees. In fact, we strongly suspect that only with explicit representations and processes (and the right additional architectural structures) will it be possible to formally prove that an artificial agent has no choice but to obey and act according to ethical principles.

There is currently a small, but increasing number of projects that attempt to tackle explicit ethical and moral artificial agents with formal guarantees of their behaviors (for example, see our Moral Competence in Computational Architectures project[6]). The goal is to develop explicit representations of social and moral norms, as well as inference, decision-making, and action-execution algorithms, that will allow robots (1) to detect morally charged situations, (2) reason through them based on their ethical rules, norms, obligations and permissions, and (3) find the best action that meets their obligations while minimizing harm to humans. Of course, these early attempts at developing moral competence for autonomous agents (Scheutz 2013) will require a sustained effort and large-scale buy-in from the AI and robotics communities in order to succeed. And even then they face a social dimension that we need to consider more broadly and prepare for critically: the conditions for accepting moral robots in our society.

## Conclusions

Returning to Asimov's writings, which anticipated human reservations to autonomous robot technology, one particular short story has striking relevance to our world today. In "… That Thou Art Mindful of Him," the two most sophisticated robots ever built attempt to address human opposition to robot technology and conclude that the Three Laws were the impediment, because they took up the most space in the robots' positronic brain and prevented robots from being small enough to do useful things, such as robot birds that catch fruit flies or robot bees that pollinate flowers.[7] They eventually arrive at the conditions under which the Three Laws could be eliminated.[8] And interestingly, this is exactly our vantage point today: we are developing and deploying robots into our society without any ethical or moral provisions, though in our case, we never had robots with moral competence in need of "devolving" in the first place. But just as the robots in Asimov's story suggest, our current simple robotic helpers will likely pave the way to more sophisticated autonomous robots, and unless we act quickly the "useful" things they do will come at a price: their ethical and moral ignorance will cause human harm.

The engineering challenge then is to ensure that our society will become a thriving, prosperous one, where humans and artificial agents cohabitate in a peaceful, mutually synergistic manner. Failing to develop appropriate cognitive architectures for autonomous agents that are sensitive to human ethical and moral concerns as well as our social emotional needs could turn utopia into dystopia, the opposite of what technological innovation aims to achieve. It is ultimately upon us whether we will succumb to the temptation of developing cognitive systems without moral and ethical abilities and endanger human societies in the process. For one thing is clear: without such moral and ethical competence, artificial agents will have no reason to act morally in any way with anybody, starting with us.

## Acknowledgments

This work was in part funded by grants N00014-14-1-0144 and N00014-14-1-0149 from the US Office of Naval Research and grant IIS 1316809 from the US National Science Foundation.

## Notes

1. It is arguable whether those provisions are truly built-in ethical provisions or rather projections of the designer onto the functional role of the implemented safety mechanisms.

2. An additional currently debated domain of autonomous robots with serious ethical implications is robots with lethal force (such as autonomous weapons systems).

3. Matell's new Hello Barbie recently triggered privacy concerns when it became known that the doll uses Wi-Fi to connect to speech recognition and AI dialogue software in the cloud to have simple dialogue interactions with kids.

4. Note that the logic-based community has started to investigate normative reasoning in single-agent and multiagent systems.

5. We do not have space here to present a full argument against that view, but just consider the fact that a moral agent might not always act to maximize its expected utility based on moral values, which are unknown to the observing IRL agent.

6. www.moralrobots.org.

7. Compare the director of research in the story stating that "… there is nothing inconceivable in the possibility of robobees designed to fertilize specific plants" to ongoing work on robotic bees at robobees.seas.harvard.edu.

8. The first condition is that the robot must never be placed in a position of danger to itself, or must be so easily replaceable that it did not matter whether it was destroyed or not. Second, it must be designed to respond automatically to certain stimuli with fixed responses, with nothing else expected of it, so that no order need ever be given it, and the fixed responses must never entail danger to human beings.

## References


Arkin, R., and Ulam, P. 2009. An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions. In *Proceedings of the IEEE 2009 International Symposium on Computational intelligence in Robotics and Automation* (CIRA 2009), 381–387. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Arkin, R.; Wagner, A.; and Duncan, B. 2009. Responsibility and Lethality for Unmanned Systems: Ethical Pre-Mission Responsibility Advisement. Paper presented at the 2009 IEEE Workshop on Roboethics, Kobe, Japan 17 May.

Arkin, R. C., and Balch, T. 1997. Aura: Principles and Practice in Review. *Journal of Experimental and Theoretical Artificial Intelligence* 9(2): 175–189.

Arnold, T., and Scheutz, M. 2016. Feats Without Heroes: Norms, Means, and Ideal Robotic Action. *Frontiers in Robotics and AI* 3(32). doi.org/10.3389/frobt.2016.00032

Asimov, I. 1942. Runaround. *Astounding Science Fiction*. (March): 94–103.

Bello, P., and Bridewell, W. 2017. There Is No Agency Without Attention. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2742

Blass, J. A., and Forbus, K. D. 2015. Moral Decision-Making by Analogy: Generalizations Versus Exemplars. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,* Austin, TX. Palo Alto, CA: AAAI Press.

Briggs, G., and Scheutz, M. 2013. A Hybrid Architectural Approach to Understanding and Appropriately Generating Indirect Speech Acts. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence,* Bellevue, WA, 1213–1219. Palo Alto, CA: AAAI Press.

Briggs, G., and Scheutz, M. 2015. "Sorry, I Can't Do That:" Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. In *Artificial Intelligence for Human-Robot Interaction: Papers from the AAAI 2015 Fall Symposium,* ed. B. Hayes and M. Gombolay, 32–36. Palo Alto, CA: AAAI Press.

Dehghani, M.; Tomai, E.; Iliev, R.; and Klenk, M. 2008. Moraldm: A Computational Model of Moral Decision-Making. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society,* Washington, D.C. Austin, TX: Cognitive Science Society Inc.

Fasola, J., and Mataric, M. 2013. A Socially Assistive Robot Exercise Coach for the Elderly. *Journal of Human-Robot Interaction* 2(2): 3–32.

Forbus, K. D., and Hinrichs, T. R. 2017. Analogy and Relational Representations in the Companion Cognitive Architecture. *AI Magazine* 38(4). doi.org/10.1609/aimag.v27i2.1882

Gert, B. 2005. *Morality: Its Nature and Justification*. Oxford, UK: Oxford University Press.

Gips, J. 1995. Toward the Ethical Robot. In *Android Epistemology,* ed. K. M. Ford, C. Glymour, and P. J. Hayes, 243–252. Cambridge, MA: AAAI Press / The MIT Press.

Iba, W., and Langley, P. 2011. Exploring Moral Reasoning in a Cognitive Architecture. In *Proceedings of the Thirty-Third Annual Meeting of the Cognitive Science Society,* Boston, MA. Austin, TX: Cognitive Science Society Inc.

Laird, J.; Lebiere, C.; and Rosenbloom, P. 2017. A Standard Model of the Mind: Toward a Common Computational Framework Across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2744

Licato, J., Sun, R., and Bringsjord, S. 2014. Structural Representation and Reasoning in a Hybrid Cognitive Architecture. In *2014 International Joint Conference on Neural Networks* (IJCNN 2014), 891–898. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

McShane, M. 2017. Natural Language Understanding (NLU, not NLP) in Cognitive Systems. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2745

Malle, B. F., and Scheutz, M. 2014, Moral Competence in Social Robots. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology,* 30–35. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Malle, B. F.; Scheutz, M.; and Austerweil, J. L. 2015. Networks of Social and Moral Norms in Human and Robot Agents. Paper presented at the International Conference on Robot Ethics ICRE 2015, Lisbon, Portugal.

Mikhail, J. 2014. Any Animal Whatever? Harmful Battery and Its Elements as Building Blocks of Moral Cognition. *Ethics* 124(4): 750–786.

Moor, J. H. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*. 21(4): 18–21.

Moor, J. H. 2009. Four Kinds of Ethical Robots. *Philosophy Now* 72.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine* 36(4): 61–70. doi.org/10.1609/aimag.v36i4.2577

Scassellati, B.; Admoni, H.; and Mataric, M. 2012. Robots for Use in Autism Research. *Annual Review of Biomedical Engineering* Volume 14: 275–294. Palo Alto, CA: Annual Reviews, Inc.

Scheutz, M. 2012. The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots. In *Anthology on Robo-Ethics,* ed. P. Lin, G. Bekey, and K. Abney. Cambridge, MA: The MIT Press.

Scheutz, M. 2016. The Need for Moral Competency in Autonomous Agent Architectures. In *Fundamental Issues of Artificial Intelligence,* ed. V. C. Müller, 517–527. Berlin: Springer.

Scheutz, M.; Schermerhorn, P.; Kramer, J.; and Anderson, D. 2007, May. First Steps Toward Natural Human-Like HRI. *Autonomous Robots* 22(4): 411–423.


**Matthias Scheutz** is a professor of cognitive and computer science in the Department of Computer Science and Bernard M. Gordon Senior Faculty Fellow in the School of Engineering at Tufts University. He has more than 250 peer-reviewed publications in artificial intelligence, natural language processing, cognitive modeling, robotics, and human-robot interaction. His current research focuses on complex cognitive robots with rudimentary moral competence.