Current Robotics Reports
https://doi.org/10.1007/s43154-023-00098-9

Machine Ethics and Cognitive Robotics

Ajit Narayanan[1]

Accepted: 26 January 2023

Abstract

Purpose of Review There is much debate in machine ethics about the most appropriate way to introduce ethical reasoning capabilities into robots and other intelligent autonomous machines (IAMs). The main problem is that hardwiring intelligent and cognitive robots with commands not to cause harm or damage is not consistent with the notions of autonomy and intelligence. Also, such hardwiring does not leave robots with any course of action if they encounter situations for which they are not programmed or where some harm is caused no matter what course of action is taken.

Recent Findings Recent developments in intelligent autonomous vehicle standards have led to the identification of different levels of autonomy than can be usefully applied to different levels of cognitive robotics. In particular, the introduction of ethical reasoning capability can add levels of autonomy not previously envisaged but which may be necessary if fully autonomous robots are to be trustworthy. But research into how to give IAMs an ethical reasoning capability is a relatively under-explored area in artificial intelligence and robotics. This review covers previous research approaches involving case-based reasoning, artificial neural networks, constraint satisfaction, category theory, abductive logic, inductive logic, and fuzzy logic.

Summary This paper reviews what is currently known about machine ethics and the way that cognitive robots as well as IAMs in general can be provided with an ethical reasoning capability. A new type of metric-based ethics appropriate for robots and IAMs may be required to replace our current concept of ethical reasoning being largely qualitative in nature.

Keywords Moral machines · Artificial morality · Robot ethics · Intelligent autonomous machines

Introduction

Recently reported fatalities involving driverless cars and reports of battlefield drones making autonomous decisions to fire on enemy targets have raised ethical concerns on two fronts: whether such autonomous machines should be used in situations where they can cause harm to humans in the first place and whether they can be given a moral dimension to their behavior so that, if they are used in these situations, they can distinguish right from wrong to prevent unethical conduct.

It may be useful to first identify the types of robots and intelligent autonomous machines (IAMs) in the scope of this review. A useful starting point is provided by the International Federation of Automatic Control, which defines "intelligent autonomous vehicles" (IAVs) as "automated vehicles capable of performing motion control tasks in unstructured or partially structured environments with little (if any) assistance from human supervisors[1]." Cognitive robotics can be understood in this context as the study of how to provide IAMs in general and robots in particular, with the appropriate architecture to learn and reason in such a way that complex tasks can be performed without human intervention. While this could be interpreted as a purely engineering or programming subarea of intelligent control, there is more to intelligent control than hardware and software development. For instance, modern cars have increasingly sophisticated systems for lane adherence, emergency braking, variable speed control, parking, and object recognition such as cycles, pedestrians, and road furniture in general. However, they still do not have something we human drivers have whenever we get into a car, no matter how technologically advanced: we know it is wrong to crash

✉ Ajit Narayanan
ajit.narayanan@aut.ac.nz

[1] Department of Computer Science, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand

[1] https://tc.ifac-control.org/7/5/activities/ifac-intelligent-autonomous-vehicles-iav-symposium

Published online: 28 February 2023

into parked vehicles, kill pedestrians, or run over cyclists. The question here for cognitive robotics is this: How can we provide a moral dimension to robot behavior which is somehow over and above, or built into, its processes which reflect human ethical reasoning that certain actions are, quite simply, wrong, no matter what the intelligent control architecture is doing?

Taking IAVs as a starting point for discussing moral cognitive robots, the potential capability of IAVs was still in doubt as recently as 10 years ago, with research in IAVs confined to simple laboratory environments due to difficulties of finding ways to incorporate and integrate intelligent sensing, reasoning, action, learning, and collaboration in a real-time environment [1]. The most promising developments up to that time included subsumption architectures, where IAV perception was linked to action without the need for an internal representation of the environment and with multiple control layers where each represented a specific behavior [2, 3]. Subsumption architectures led to major developments in walking robots [4], small rovers for Mars missions [5], and sociable robots [6]. A more classical, symbolic control approach was autonomous robotic architecture (AuRA), where different modules for planning, reasoning, and motion interacted via schemas for collision avoidance and problem resolution [7]. Perhaps the most successful approach was the intelligent controller (IC) architecture, which combines the subsumption approach with modules to create internal representations from incoming sensor data to fuse with previous sensor data [8, 9]. The IC architecture has been successfully applied to several unmanned aerial and underwater vehicles (e.g., [10]). Other developments included enhancing general cognitive architectures such as Soar [11] and ACT-R [12], with perception and actuation modules for interaction with the environment. Such cognitive architectures use production rules and can be supplemented with additional algorithms for dealing reactively within dynamic environments [13, 14]. A real-time control system (RCS) developed at NIST is a reference model architecture involving a systematic mapping using nodes, where a node consists of behavior generation, sensory processing, world modeling, and evaluation. RCS was applied to autonomous on-road driving in 2004 [15]. Underlying all IAV and intelligent robot approaches so far are three basic stages: sensing and processing, decision-making, and reacting. The second and third stages are dependent on the first stage: if the architecture does not sense something, it cannot be decided on or reacted to.

Intelligent Autonomous Vehicle (IAV) Development

DARPA's sponsorship of a series of IAV challenges started in 2004, when a competition was held to self-navigate over 140 miles of desert roadway in 10 h. The aim at that time was that a third of military vehicles should be IAVs by 2015. But in 2004, no IAVs could proceed more than a few miles without crashing. Improvements in control software, collision avoidance, road following, and radar and laser sensing technologies contribute d to significant advances so that, in 2007, the route was changed to 60 miles of urban conditions. Four IAVs completed the task within the 6-h time limit allowed. Since that time, advances have led to several spin-offs in conventional vehicles, such as lane adhesion, emergency braking, and self-parking. Currently, IAV technology and development are being driven by major car manufacturers such as Mercedes, Nissan, BMW, VW, Volvo, and GM, with Google emerging as another major contributor in 2009 with its self-driving car project (evolved to Waymo in 2016) building on its Google Maps data to recognize locations. In 2015, Tesla introduced the Model S which had autonomous steering, side collision avoidance, lane changing, and parallel parking capabilities. Software updates to its autopilot system now allow Model S to self-park without the driver being in the car. Also in 2015, Delphi Automotive developed an IAV that drove over 3000 miles coast-to-coast across North America under autonomous control for 99% of the distance.

A classification system was released in 2014 by the automotive standardization body SAE International, based on the amount of driver attention required [16]. Level 0 is driver only, where the driver manages all driving aspects (steering, speed, monitoring of driving environment). Level 1 is assisted, where the driver is given support for either steering (e.g., parking) or speed (e.g., cruise control) in specific situations. Level 2 is partial automation, where the driver is given support for both steering and speed (e.g., lane adherence with cruise control) but must continue monitoring the driving environment to intervene when necessary. Level 3 is conditional automation where the driver can relinquish control to an automated driving system which controls steering, speed, and monitoring of the driving environment but must be ready to take back control. Level 4 is high automation where the automated driving system controls all aspects of the dynamic driving task (steering, braking, speed, environment monitoring, changing lanes) even if a human driver does not respond appropriately to a request to intervene. Finally, Level 5 is complete end-to-end journey without any driver intervention. The distinctions between "automated," "automatic," and "autonomous" are not always clear [17, 18], but it is generally accepted that autonomous vehicles are characterized by vehicles achieving levels 3, 4, and 5 of the SAE classification system, where the driver relinquishes control of the driving environment either partly or fully.

No comparable or systematic classification system exists for intelligent or cognitive robotics, although several working definitions and standards are provided depending on the application area. For instance, the Japanese Industrial Robot Association (JIRA) identifies Class 6 (intelligent robots) as

devices which have a good assessment of their environment and perform tasks by manipulating movements depending on changes in their environment.

But if the SAE approach is applied to robotics, the following levels can be proposed from a cognitive robotics perspective:

Level 0 is the robot fully under the guidance and control of a human.
Level 1 is assisted, where the robot is given support for some of its functions by a human.
Level 2 is partial automation, where humans are provided with some robotic support in specified situations.
Level 3 is conditional robotic automation, where the human can hand control over to the robot but can take back control when needed.
Level 4 is high robotic automation, with the human present but not necessarily in control of all situations.
Level 5 is full robotic automation with no human involvement in any of its tasks and functions.
Cognitive robots, under this proposal, are characterized by the use of architectures for enabling levels 3, 4, and 5, where aspects of robot autonomy and intelligent decision-making are demonstrated.
The notion of driverless cars and robots having a sense of right and wrong to guide their behavior and actions plays no part in current standards and proposals. We can therefore identify a potential level 6 for cognitive robotics under the SAE approach:
Level 6: Full automation with a moral sense of when it is right and wrong to act autonomously and independently of humans, with control being handed back to humans in case of moral indecision.
This leads to a level 7, or "final frontier," for cognitive robotics.
Level 7: Full automation where the robot can make its own decisions on whether its actions are right or wrong with no human intervention.

The question then arises as to whether levels 6 and 7 are feasible for cognitive robots and, if so, how. The underlying problem here is how to give robots at levels 6 and 7 the ability to make alternative decisions on moral grounds to those which may have been initially identified and calculated by its technological control architecture.

The Need for Improved AI in Autonomous Systems

The need for artificial intelligence and intelligent decision-making in the architectures of autonomous control systems of driverless cars and intelligent robots was recognized early in the 1990s [17]. Since that time, major advances in autonomous control have been driven by increases in processing power and big data. Supercomputers can now process massive amounts of sensor, camera, and radar data in processors capable of operating at over three hundred trillion operations per second for the sensing and processing stage. Associated with these technological advances are deep learning and other machine learning technologies and architectures, where huge amounts of training data are fed into neural networks consisting of dozens or even hundreds of layers on GPU-accelerated platforms [19] for, say, learning about traffic conditions and making decisions in partially and fully automated driverless systems. But alongside predictions by industry commentators that advances in big data and deep learning are leading to level 5 IAVs within a matter of years [20], there are signs of skepticism that full autonomy will be achieved in such a short time as well as worries that over-expectation can lead to an "AI winter" [21]. The problem appears to be that, despite technological advances in sensors, architectures, and processing power, it is not clear how to incorporate basic "common sense" for dealing with new situations into such technologies [22]. The reason for this skepticism lies in the nature of the accidents involving autonomous vehicles that have occurred up to now.

One of the main motivations for autonomous vehicles is that they are intended to be safer than vehicles controlled by the average human driver. For 2015, there were approximately 3.6 road fatalities per 1 billion vehicle-driven km in the UK, 7.1 in the USA, and 8.7 in New Zealand [23]. Autonomous vehicles and especially fully autonomous vehicles have not been driven for long enough for comparable fatality numbers to be reliably calculated. But the number of deaths involving partially autonomous cars, which was four up to 2018, has been increasing with the number of vehicles on the road with automated driving systems, with 11 killed in the USA alone during a 4-month period in 2022[2]. Ten of these deaths involved Tesla vehicles. Investigations are still underway on how many of these deaths are attributable to driver error rather than Tesla's autopilot technology. The number of accidents involving vehicles with partially automated systems is now also being reported to the American National Highway Traffic Safety Administration, with nearly 400 crashes reported in the USA in the 11 months prior to June 2022[3]. Recent estimates for the USA indicate 9.1 self-driving car accidents per million miles driven in comparison

[2] https://www.latimes.com/business/story/2022-10-18/11-more-crash-deaths-are-linked-to-automated-tech-vehicles#:~:text=Eleven%20people%20were%20killed%20in,incidents%20linked%20to%20the%20technology

[3] https://www.npr.org/2022/06/15/1105252793/nearly-400-car-crashes-in-11-months-involved-automated-tech-companies-tell-regul

to 4.1 crashes with human-operated vehicles[4]. The implication is that, while fatalities may not be so high with partial or fully automated vehicles, the number of crashes involving such vehicles may be higher than average. Nevertheless, the nature of how those incidents raises questions concerning how intelligent driverless cars really are and whether they can be trusted to be autonomous in fully self-driving mode[5].

The first of the four fatalities reported up to 2018 was in January 2016, when a Tesla car in Hebei, China, crashed into the back of a road-sweeping truck, killing the driver. The lack of any evidence of car braking or swerving led to claims that the autopilot was engaged but failed to work properly. In May of that year, a Tesla car in Williston, Florida, crashed into a tractor-trailer while in autopilot, with the cause identified as the white side of the trailer not being distinguished from the brightly lit sky and so the brake not applied. In March 2018, a self-driving Uber car killed a pedestrian in Tempe, Arizona, when she walked her bicycle across a street at 10 pm. Also in March 2018, in Mountain View, California, a Tesla car in autopilot set for 75 miles an hour crashed into a safety barrier, killing the driver and causing two other vehicles to crash into it. More recently, four fatalities involved motorcyclists and Tesla vehicles, and several crashes involved emergency vehicles parked on the roadside and displaying warning lights. Such incidents have led to calls for partially and fully automated car manufacturers to improve their sensor systems, technology, and control architectures[6]. However, it is not clear whether improvements are needed in the hardware (sensor technology) or software (control and decision-making systems) components of IAVs, or both. "Blind spots" can be due to lack of sensor sensitivity or lack of algorithms for dealing with the sensor data.

Moral Machines

Another way to improve IAVs is to carry out research on whether they can determine for themselves that it is wrong to take actions (or not to take actions) that can lead to humans being killed. While technological advances in sensor technology, architectures, hardware, and programming can lead to improvement in autonomous vehicle safety, another approach is to ensure that IAVs and intelligent robots acquire an ethical sense that will ensure that their sensor-based decisions and actions do not harm humans no matter what control architectures or types of technology are being used. This, after all, is what underlies human driver behavior irrespective of the actual car being driven or driver-assisted technology being used in that car. In other words, there is a need to promote further development of moral autonomous systems, including robots, at levels 6 and 7 as proposed above, if such systems and robots are to be used in safety-critical situations that may lead to death.

Three types of relationships exist between ethics and autonomous systems. The first type is ethical design, which is a method for encouraging the design of systems for human values [24] and the consideration of ethical issues when designing and developing systems [25]. The IEEE Standards Association has recently launched a global initiative on ethical design approaches for autonomous and intelligent systems which outlines ethically aligned design around the principles of human rights, well-being, accountability, transparency, and awareness of misuse [26•]. The aim is to ensure that ethical considerations are prioritized in the design and development of autonomous systems for the benefit of humanity. The second type is the ethics of autonomous systems and the consideration of whether it is right or wrong to construct such systems [27]. Such considerations take into account the possibly dehumanizing aspects of research into AI as well as implications, such as loss of jobs and the dangers of superintelligence [28].

The third type is machine ethics, which is the area that looks into how we give ethical principles to intelligent systems so that such systems can decide for themselves what is right and wrong [29••]. The problem is that programming ethical principles into an autonomous system is like hard-wiring the system so that it must follow these principles. The problem with this approach is that such hard-wiring goes against the notion of intelligent autonomous systems that are supposed to make informed decisions for themselves [30•]. "I avoided killing the cyclist because I'm hard-wired to do so" is a command that does not allow for exceptions. This may work if the options are to harm or not harm the cyclist. But such a command is of no use when a dilemma occurs, as might be the case when the vehicle has to decide whether to knock over a cyclist or swerve and hit a mother and child on the pavement, with no other options available. For an autonomous system to do only what it is told to do raises questions as to what autonomy means as well as leaves no room for moral decision-making when confronted by a dilemma. More importantly, it raises questions as to whether the fatalities and crashes that have so far occurred involving IAVs have been caused by the system not being able to do other than what it was programmed to do. In other words, because there was no ethical component in the autonomous system, it followed instructions and executed actions blindly with no concept of the harm that such actions could cause. The question is whether improvements in sensor technology

[4] https://getjerry.com/questions/how-many-fatalities-have-been-due-to-self-driving-vehicles

[5] https://www.theguardian.com/technology/2022/dec/22/tesla-crash-full-self-driving-mode-san-francisco?CMP=Share_iOSApp_Other

[6] https://apnews.com/article/us-news-ap-top-news-ca-state-wire-government-regulations-transportation-d03d88fca7ef389ffbe3469f50e36dcf

and architectures, supported by a command-based approach, will lead to the prevention of accidents which, if a human had been driving, could have been avoided. The nature of the accidents that have so far been reported for IAVs indicate that the technology is not quite there for fully autonomous self-driving cars. Whether the technology will improve to the point where there are no accidents is currently unknown. But relatively unexplored in AI and cognitive robotics is whether to make the potential actions of IAV subject to a moral decision-making rather than hard-wiring them and, if so, how.

As can be seen from the above, while there is much research and debate concerning the rights and wrongs of AI, as well as what forms of ethics should be built into machines, the issue of how we design and implement an ethical reasoner for use in intelligent, autonomous machines including cognitive robots and autonomous vehicles is relatively unexplored. More precisely, research into how to make intelligent robots and autonomous vehicles calculate and behave as moral agents [31] and so shed light on "artificial morality" [32, 33] is a relatively under-explored area in artificial intelligence, machine learning, and cognitive robotics. It is possible that moral machines represent a final frontier for AI, cognitive robotics, and IAMs in general (the proposed levels 6 and 7 above).

Moral Machine Architectures

Three problems need to be addressed [34] when designing, developing, and implementing an ethical or moral machine. The first problem concerns the type and degree of *interactivity* that allows the moral machine to respond ethically to its environment. The second is the degree of *autonomy* from the environment that allows the moral machine to go through an ethical reasoning process of its own. And the third is the amount of *adaptability* the moral machine is allowed to change its ethical reasoning processes. Together, these three desirable properties provide the basis for a *trustworthy* moral machine. The extent of the trust placed on such moral machines will depend on how it responds to different ethical situations and its ability to provide justifications for its responses.

To understand the difficulties researchers face in designing and developing moral machines, we need to discuss briefly what ethics is. Ethical theories deal with rules or criteria for distinguishing right from wrong as well as good from bad. Examples of ethical theories are deontology (we must act according to duties and obligations), categorical imperative (we must act in accordance with human rational capacity and certain inviolable laws), utilitarianism (an action is right or good if it leads to most happiness), and consequentialism theories in general (whether an action is right or good depends on the action's outcome or result). Another approach is virtue ethics (we must act in ways that exhibit our virtuous character traits). Previous approaches to implementing ethical reasoning in intelligent systems have not always clearly identified the ethical approach adopted and have involved case-based reasoning, neural networks, constraint satisfaction, category theory, and abductive logic programming as well as inductive logic programming. We present a brief overview below.

Early attempts in case-based reasoning approaches include Truth-Teller and SIROCCO, where the former identified shared features in a pair of ethical dilemmas and the latter retrieved ethical cases similar to a new case [35]. However, case-based reasoning systems in general are intended to support human ethical decision-making rather than help machines perform ethical reasoning on their own. Artificial neural network (ANN) approaches that learn ethical outputs from training samples [36] require specific topology and learning architectures for successful testing, with uncertainty concerning how to characterize the type of moral reasoning involved in the learning interaction. Also, the lack of rule-based reasoning capability internally or as output can lead to criticisms that such networks lack both transparency and autonomy. That is, such networks can only go through an internal and possibly ethically uninterpretable transition when given an input from the environment. Constraint satisfaction approaches [37, 38], while useful for certain types of AI problems requiring optimal solutions that do not violate conditions, assume full observability of the world ("closed world assumption") that can cause problems when knowledge is partial, vague, or uncertain. Also, the need not to violate constraints is a form of deontology: actions are right or wrong depending only on rules rather than consequences. This makes the application of constraint satisfaction approaches to ethical dilemmas difficult, since dilemmas involve a decision to be made between two opposing constraints. Category theory approaches [39, 40] lead to ethical reasoning being interpreted as a functional process of mappings or morphisms, from a domain of entities to a codomain. The use of category theory in machine ethics applies this formal approach so that an ethically relevant decision is correct if a formula containing that decision can be identified and proved as a theorem in an axiomatic system. As noted earlier, formal rule-based approaches to machine ethics, such as category theory and constraint satisfaction, raise questions concerning genuine autonomy. Such machines can only do what they are programmed to do within the formal system. Prospective, or abductive, logic approaches [41] attempt to "look ahead" to future states before selecting a posteriori preferences. While the application of such an approach in machine ethics has the advantage of allowing a degree of consequentialism, the choice between preferences needs the support of a knowledge base and a set of non-violable

integrity constraints. Such abductive logic approaches, similar to the other formal approaches of constraint satisfaction and category theory, depend on the closed-world assumption of a preference not being against known principles and constraints which are hardwired into the program. Finally, inductive logic programming approaches can be used for machine learning of prima facie duty theory and where there is no single absolute duty to be adhered to for deciding whether an IAV should or should not take over control [42, 43]. The requirement is for a list of binary ethical features, a list of duties for minimizing or maximizing, and a list of actions. A number of cases can be represented in these data structures for machine learning of ethical principles, together with preferable actions as target (class) values, leading to a training-testing regime for learning when, morally, to take over control from a human driver. However, it is not clear how non-binary features can be handled (e.g., the varying desirability of respecting driver autonomy depending on continuously changing sensor information). Nor is it clear how missing, partial, or inexact values affect the learning of ethical principles, since all features need to have values for the inductive engine to operate on a complete and consistent basis. Finally, there is a slowly emerging consensus that inductive machine learning algorithms with human-specified feature values could be subject to algorithmic bias or learn biases in supplied data [44•, 45, 46]. It is especially important for moral machines not to learn specific ethical preferences of programmers or biases in supplied data if they are to be considered trustworthy.

Conclusion

In summary, previously applied work in machine ethics does not address all three of the desirable properties of machine ethics. Interactivity is typically implemented through fixed data input (e.g., training data) rather than sensors that produce dynamically changing data. Previous approaches have not shown how moral decision-making can vary through interaction with a dynamic environment. Adaptability is implemented as classifying unseen cases after successful training, as in the case of inductive logic programming and ANNs. However, another more intuitive sense of adaptability is the generalization or application from what is known from previous cases so that moral decisions continue to be made consistently for situations not previously encountered. Finally, there is a tendency to "over-prescribe" the system with strict and formal moral rules, leading to questions as to how much genuine autonomy a machine ethics system contains. In particular, moral decision-making is typically based on formal rule-following rather than internal reasoning based on state-matching and state-transition, both of which may be approximate or imprecise.

A recent approach has been to use fuzzy logic to represent general moral principles of deontology and consequentialism in an ethics architecture and embed the architecture in a simulation to generate a thousand sample "case studies" of when it is right or wrong for an ethical reasoner to take action depending on a deontological or consequentialist perspective. These samples can then be used as training data so that an IAV, battlefield drone, or other forms of cognitive robot can learn for itself the rules for making moral judgments of its own in response to situations not included in the training set of samples, including dilemmas where life will be at risk no matter what action is taken [47, 48].

Any approach to machine ethics must also demonstrate that it can cope with ethical dilemmas, since such dilemmas test the ability of the system to go beyond what it has learned to do in specific situations to situations not previously encountered. In particular, dilemmas test the ability of an ethical system to balance conflicting aspects of duty against consequences. Consider the previous situation where an IAV is confronted by a cyclist suddenly on its path, and it is too late to brake without hitting the cyclist. The alternative is to swerve onto the pavement where almost certainly, according to its sensors, a mother and child will be killed. In this situation, someone is going to be badly hurt or die, and it is too late to hand control back to the driver. One moral decision is to justify hitting the cyclist on the basis that the lives of a child and mother count more than the life of an adult cyclist. While there may be disagreement on whether this is the right moral outcome, what should not be doubted is that a moral judgment was made by the IAV which is transparent and reasoned, leading to assurance that the system is acting morally and therefore is more trustworthy than a system that makes no decision or hands back control to the human driver when it is too late for the driver to react.

The key to producing successful moral machines will probably consist of providing a minimal set of moral principles that are acceptable in the application domain and allow the machine ethics system to "decide for itself" (autonomy) how to apply those principles in dynamically changing environments (interactivity) and to derive moral rules that will allow it to monitor and change its behavior in the light of new information (adaptability). Another challenge is that previous approaches have generally used formalisms for representing moral rules and outcomes that make it difficult for the system to reason flexibly and produce output which is difficult to interpret. Since natural language is used to express moral arguments and reasoning, it may be better to represent moral principles and reasoning in ways that are more naturalistic than formalistic.

There are still issues concerning the location of any ethics reasoner in moral machines and cognitive robots. The first is for the ethics architecture to be totally independent of the control architecture of a cognitive robot (zero ethics

coupling). The architecture provides an ethical commentary to the sensor-based behavior of the robot but does not interfere in any way with its actual behavior. Such commentary could be used off-line to monitor and evaluate robot actions in responses to sensor data. The second is for the ethics architecture to work in parallel with the control architecture of the robot so that control architecture output and ethics architecture output are combined in some way (parallel ethics coupling) for "morally considered" action. And the third possibility is for the output of the control architecture to be one of the inputs to the ethics architecture (serial ethics coupling) so that no action is possible without "ethical approval." The choice of architectural involvement will depend on the context in which cognitive robots will be used. Perhaps sometime in the future, there will be a requirement for all cognitive robots, IAVs, and other IAMs that have the potential to inflict harm on humans to have a compulsory ethics architecture so that control information and ethics outcomes can be used together in parallel or in serial mode.

None of this is achievable without human designers providing the first set of principles to help guide the robot to initial moral decision-making (in the way that children are taught moral reasoning). But human designers do not have to specify every possible ethical situation. As noted by several researchers, for instance, just because fuzzy logic deals with inexactness and approximation due to its use of the real number interval between 0 and 1, that does not mean that fuzzy logic cannot be subject to the formalization of its inferential processes so that fuzzy reasoning is shown to be effective and computable [49].

Finally, research into how values can be embedded into IAVs and intelligent robots through advances in data collection, sensor technology, pattern recognition, and machine learning is now actively encouraged as a method of achieving a correct level of trust between humans and autonomous intelligent systems [50]. Ethical reasoning in philosophy and machine ethics has so far dealt almost exclusively with qualitative reasoning. The use of numbers in ethical decision-making and moral outcome evaluation is not familiar to us humans and could lead to accusations of a metric-based ethics being "ethics by numbers." But if we want our IAVs, cognitive robots, and other intelligent autonomous systems to develop a sense of right and wrong so that they do not harm us, we may have to accept that non-qualitative, metric-based ethics is not just the best way to go but the only way to go, since fundamentally machines work in and on numbers. One of the implications of the final frontier for cognitive robotics being moral cognitive robotics is that such moral machines will lead to new theories of ethics based on numbers. It will be interesting to see if we humans can accept metric-based ethics in moral machines to assure ourselves that cognitive robots are trustworthy as they roll out over the next 10 to 20 years into all parts of society.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflict of interest The author declares no competing interests.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by the author.

References

Papers of particular interest, published recently, have been highlighted as:
• Of importance
•• Of major importance

1. Long, LN, Hanford SD, Janrathitikarn O, Sinsley GL, Miller JA. A review of intelligent systems software for autonomous vehicles. Proceedings of the 2007 IEEE Symposium of Computational Intelligence in Security and Defense Applications (CISDA 2007), 2007, 69-76.
2. Brooks RA. A robust layered control system for mobile robot. IEEE J Robotics Auto. 1986;2(1):14–23.
3. Brooks RA. Intelligence without representation. Artif Intell J. 1991;47:139–59.
4. Brooks RA. A robot that walks; Emergent behaviour from a carefully evolved network. Neural Comput. 1989;1(2):253–62.
5. Matijevic M. Autonomous navigation and the Sojourner Microrover. Science. 1998;280(5362):454–5.
6. Breazeal C. Designing sociable robots. 2002, MIT Press.
7. Arkin RC, Balch T. AuRA: Principles and practice in review. J Exp Theor Artif Intell. 1997;9(2–3):175–89.
8. Stover JA, Ratnesh K. A behaviour-based architecture for the design of intelligent controllers for autonomous systems. IEEE International Symposium on Intelligent Control/Intelligent Systems and Semiotics. 1999, Cambtridge, MA, 308-313.
9. Stover JA, Hall DL, Gibson RE. A fuzzy logic architecture for autonomous multisensory data fusion. IEEE Trans Ind Electron. 1996;43:403–10.
10. Kumar R, Stover JA. A behaviour-based intelligent control architecture with application to coordination of multiple underwater vehicles. IEEE Trans Syst Man and Cybernetics. 2000;30:767–84.
11. Laird JE, Newell A, Rosenbloom PS. Soar: an architecture for general intelligence. Artif Intell. 1987;33(3):1–64.

12. Anderson JR, Bothell D, Byrne MD, Douglas S, Lebiere C, Qin Y. An integrated theory of mind. Psyc Rev. 2004;11(4):1026–60.
13. Jones RM, Laird JE, Nielsen RE, Coulter KJ, Kenny R, Koss FV. Automated intelligent pilots for combat flight simulations. AI Magazine. Spring 1999, 27-41.
14. Bugajska MD, Schultz AC, Trafton JG, Taylor M, Mintz FE. A hybrid cognitive-reactive multi-agent controller. Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2002, Lausanne, 2807-2812.
15. Barbera T, Albus J, Messina E, Schlenoff C, Horst J. How task analysis can be used to derive and organize knowledge for the control of autonomous vehicles. Rob Auton Syst. 2004;49(1–2):67–78.
16. Automated driving: Levels of driving automation are defined in new SAE International Standard J3016. http://www.sae.org/misc/pdfs/automated_driving.pdf (Accessed on 24 August 2018).
17. Antsaklis PJ, Passino KM, Wang SJ. An introduction to autonomous control systems. IEEE Control Syst. 1991;11(4):5–13. https://doi.org/10.1109/37.88585.
18. Wood SP, Chang J, Healy T, Wood J. The potential regulatory challenges of increasingly autonomous motor vehicles. 2012. 52nd Santa Clara Law Review. 4 (9): 1423–1502.
19. Deep learning. Nvidia Accelerated Computing. https://developer.nvidia.com/deep-learning (Accessed 24 August 2018.)
20. Pal, K. The 5 most amazing AI advances in autonomous driving. April 2018. *Techopedia.* https://www.techopedia.com/the-5-most-amazing-ai-advances-in-autonomous-driving/2/33178 (Accessed 24 August 2018.)
21. Brandom R. Self driving cars are headed toward an AI roadblock. July 2018. T*he Verge*. https://www.theverge.com/2018/7/3/17530232/self-driving-ai-winter-full-autonomy-waymo-tesla-uber (Accessed 24 August 2018)
22. Knight W. Finally, a driverless car with some common sense. September 2017. *MIT Technology Review*. https://www.technologyreview.com/s/608871/finally-a-driverless-car-with-some-common-sense/ (Accessed 24 August 2018)
23. Road safety annual report 2015. OECD/ITF. Paris: International Traffic Safety Data and Analysis Group, International Transport Forum. https://www.oecd-ilibrary.org/transport/road-safety-annual-report-2015_irtad-2015-en (Accessed 24 August 2018)
24. Sekiguchi K, Tanaka K, Hori K. "Design with discourse" for design from the ethics level. *Proceedings of the 2010 Conference on Information Modelling and Knowledge Bases XXI*. 2010, pp. 307-314.
25. Spiekermann S. *Ethical IT innovation: a value-based system design approach*. 2015. CRC Press.
26.• IEEE Standards Association. *Ethically Aligned Design*, First Edition (EAD1E). 2022. Available from https://standards.ieee.org/industry-connections/ec/ead1e-infographic/ (Accessed 16 December 2022.) **An important statement from the world's foremost professional association for electronic and electrical engineering with over 400,000 members worldwide concerning ethical design, development, and implementation of intelligent and autonomous systems for ensuring human rights, well-being, accountability, transparency, and awareness of misuse.**
27. Russell SN, Norvig P. The ethics and risks of developing artificial intelligence. 2009 (3rd Edition). In *Artificial Intelligence: A Modern Approach.* Chapter 26.3. Prentice Hall.
28. Bostrom N, Yudkowsky E. The ethics of artificial intelligence. 2014. In K. Frankish, W.M. Ramsey (Eds), *The Cambridge Handbook of Artificial Intelligence.* Chapter 15. CUP.
29.•• Anderson M, Anderson SL (Eds). *Machine Ethics.* 2011. CUP. **An important collection of foundational papers in machine ethics from some of the influential researchers in artificial intelligence, philosophy, and robotics. Several references in this bibliography are taken from this collection.**
30.• Arvan M. Mental time-travel, semantic flexibility, and AI ethics. *AI & Society.* 2018. https://doi.org/10.1007/s00146-018-0848-2 (Accessed 24 August 2018). **Introduces and discusses the problem of how to get a moral machine to deal with moral dilemmas, where one or other option leads to some human harm.**
31. Anderson SL. Machine metaethics. In Anderson and Anderson (Eds.), *Machine Ethics*. 2011. CUP.
32. Allen C, Wallach W, Smit I. Why machine ethics? In Anderson and Anderson (Eds.), *Machine Ethics*. 2011. CUP.
33. Anderson SL. Philosophical concerns with machine ethics. In Anderson and Anderson (Eds.), *Machine Ethics*. 2011. CUP.
34. Floridi L. On the morality of artificial agents. In Anderson and Anderson (Eds.), *Machine Ethics.* 2011. CUP.
35. McLaren BM. Computational models of ethical reasoning: challenges, initial steps, and future directions. *IEEE Intelligent Systems.* 2006, 21(4): 29-37. Reprinted in Anderson and Anderson (Eds.), *Machine Ethics.* 2011. CUP.
36. Guarini M. Computational neural modelling and the philosophy of ethics: reflections on the particularism-generalism debate. In Anderson and Anderson (Eds.), *Machine Ethics.* 2011. CUP.
37. Mackworth AK. Architectures and ethics for robots: Constraint satisfaction as a unitary design framework. In Anderson and Anderson (Eds.), *Machine Ethics.* 2011. CUP.
38. Turilli M. Ethical protocols design. In Anderson and Anderson (Eds.), *Machine Ethics.* 2011. CUP.
39. Bringsjord S, Arkoudas K, Bello P. Towards a general logicist methodology for engineering ethically correct robots. IEEE Intell Syst. 2006;21(4):38–44.
40. Bringsjord S, Taylor J, van Heuveln B, Arkoudas K, Clark M, Wojtowicz R. Piagetian roboethics via category theory: moving beyond mere formal operations to engineer robots whose decisions are guaranteed to be ethically correct. In Anderson and Anderson (Eds.), *Machine Ethics.* 2011. CUP.
41. Pereira LM, Saptawijaya A. Modeling morality with prospective logic. In Anderson and Anderson (Eds.), *Machine Ethics.* 2011. CUP.
42. Anderson SL, Anderson M. A prima-facie duty approach to machine ethics: machine learning of features of ethical dilemmas, prima facie duties, and decision principles through a dialogue with ethicists. In Anderson and Anderson (Eds.), *Machine Ethics.* 2011. CUP.
43. Anderson SL, Anderson M. A prima facie duty approach to machine ethics. In Anderson and Anderson (Eds.), *Machine Ethics.* 2011. CUP.
44.• Garcia M. Racist in the machine: the disturbing implications of algorithmic bias. 2016. *World Policy Journal*, 33(4):111-117. Accessed October 2018, from http://muse.jhu.edu/article/645268/pdf.**An important contribution with major implications for machine learning and how bias can be introduced through selective data. The paper has implications for machine ethics in that the moral behaviour of autonomous machines may be determined by the data provided to learn such behaviour.**
45. Devlin H. AI programs exhibit racial and gender biases, research reveals. Accessed October 2018, from https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals.
46. Fuchs DJ. The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer*, 2(1). Accessed October 2018, from http://scholarsmine.mst.edu/peer2peer/vol2/iss1/1.
47. Narayanan A. Ethical judgement in intelligent control systems for autonomous systems. Proceedings of the 2019 Australian and New Zealand Control Conference (ANZCC), 231-236. 2019. https://doi.org/10.1109/ANZCC47194.2019.8945790.
48. Narayanan, A. Can lethal autonomous robots learn ethics? In: Gallagher, M., Moustafa, N., Lakshika, E. (eds) AI 2020: Advances in Artificial Intelligence. Lecture Notes in Computer Science, vol 12576. Springer, Cham. 2020. https://doi.org/10.1007/978-3-030-64984-5_18

49. Novak V. Fuzzy logic, fuzzy sets, and natural languages. Intl J Gen Syst. 1991;20(1):83–97. https://doi.org/10.1080/03081079108945017.
50. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Industry Connections Activity Initiation Document* (ICAID), v 4.0, 3 February 2022. https://standards.ieee.org/wp-content/uploads/import/governance/iccom/IC16-002-Global_Initiative_for_Ethical_Considerations_in_the_Design_of_Autonomous_Systems.pdf (accessed 16 December 2022).