Machines That Know Right And Cannot Do Wrong: The Theory and Practice of Machine Ethics

Louise A. Dennis and Marija Slavkovik

***Abstract*—Machine ethics is an emerging discipline in Artificial Intelligence (AI) concerned with enabling autonomous intelligent systems to uphold the ethical, legal and societal norms of their environment. Why is machine ethics only developing as a field now, and what are its main goals and challenges? We tackle these questions and give a survey of state of the art implementations.**

"The fact that man knows right from wrong proves his intellectual superiority to the other creatures; but the fact that he can do wrong proves his moral inferiority to any creatures that cannot." – Mark Twain

I. Morality for machines

THE scenario is by now familiar: you are in a tunnel and your autonomous car has a break failure. There are workers on the road ahead. What should the car do? One option is to ram in the wall and possibly kill you, its owner and sole passenger. The other, to continue straight on its way and kill numerous road workers. Many questions are open regarding what the car should do, all subject of *machine ethics* [26].

The first challenge facing machine ethicists is what ethical conduct should an autonomous system exhibit and who gets to decide this. An equally important challenge is the one that we focus on here: How should an autonomous system be built and programmed so as to follow the ethical codex of choice? How can we do this in a way that allows a regulatory body to determine that the ethical behaviour described is the one exhibited? In summary, what does it mean to construct an artificial system that knows right from wrong and then ensure that it, unlike man in Mark Twain's quote, is unable to do wrong.

II. Why now?

AI has been an established research field since 1956 [31] but machine ethics, outside of science fiction, has emerged as a concern in the last decade. Why only now? At least two things have recently changed in how AI is used.

Powerful autonomous systems now share in our physical and e-space. Consider for example, industrial robots that have been in operation at least since the 80ies [27], and automated subway systems, which have been in operation for the past forty years[1]. Both of these types of machines have the capacity to seriously harm people and property, however they operate in a *work envelope*, a segregated space which only trained personel are allowed to enter. Machines that did share the space with people had no physical ability to do harm, such as automated pool cleaners. In contrast, machines like automated cars and assisted living devices have the ability do do harm and are not operating in a segregated environment.

Methods developed in AI have long been in use: *e.g.,* complex scheduling systems built using constraint satisfaction programming [33]. However, each of these AI systems have been domain and context specific. Any possible ethical, legal and societal issues that might arise from the use and deployment of the system could and had been handled during development. Today in contrast, particularly with machine learning applications, we see off-the shelf software and hardware available to any one to customize and deploy for an unpredictable variety of tasks in an unpredictable variety of contexts. Thus issues of machine "unethical behaviour" and impact can no longer be dealt with entirely in development.

III. Morality as a function of ability

Much has been said on whether an artificial agent, can be a moral agent, see *e.g.,* [17]. As with autonomy, we tend to refer to two different concepts: categorical morality for people, and degrees of morality for machines [35], [26].

Wallach and Allen [35, Chapter 2] distinguish between operational morality, functional morality, and full moral agency. An agent has operational morality when the moral significance of her actions are entirely scoped by the agent's designers. An agent has functional morality when the agent is able to make moral judgements when choosing an action, without direct human instructions.

Moor [26] distinguishes between agents with ethical impact, implicitly ethical, explicitly ethical and full moral agents. An agent has ethical impact if her operations increase or decrease the overall good in the world. A parallel can be drawn to [35]: implicitly ethical agents have operational morality, while explicitly ethical agents have functional morality. Dyrkolbotn et al. [15] further refine and formalise the concepts of implicitly and explicitly ethical agents by stipulating that implicitly ethical agents are those which do not use their autonomy to make moral judgements.

It is clearly better to build implicitly ethical artificial agents because their moral choices can be evaluated while the agent is built and assurances can be given about what the agent will do in a morally sensitive context. However, for agents whose context of operation is either unpredictable or too complex, explicit moral agency is the only design option [15].

L. Dennis was funded by EPSRC grants EP/L024845/1 (Verifiable Autonomy), EP/R026084/1 (Robotics and AI for Nuclear) and EP/R026092/1 (Future AI and Robotics Hub for Space).

[1]http://www.railjournal.com/index.php/metros/uitp-forecasts-2200km-of-automated-metros-by-2025.html

Having chosen what kind of artificial moral agent one needs, one has a choice between a bottom-up, top-down or a hybrid approach [36], [10]. In a top-down approach an existing moral theory is chosen and the agent is implemented with an ability to use this theory. In a bottom-up approach, the artificial agent is presented with examples of desirable and undesirable choices and she develops an algorithm by which to make moral judgements in unfamiliar circumstances. A hybrid approach uses elements of both the top-down and bottom-up. All of these approaches have advantages and disadvantages [10].

IV. Ethical Theories for Machines

Moral philosophy is concerned with developing moral theories, which should guide moral judgements. However the theories so far developed have a human locus, so not all can be trivially adapted for use by artificial agents. How can virtue ethics [22] for example, be used for an agent that can choose her reward function? Alternatively one might consider developing a new moral theory, specifically for machines. A (perhaps bad) example of such a theory are the Three Laws of Robotics of Asimov [4].

Ethical theories considered for use by artificial agents are: utilitarianism [21], Ross's ethical theory [30], and Kantianism [16]. Utilitarianism stipulates that all choices can be evaluated by the amount of good or bad (utilities) that they bring about. A moral agent needs to maximise the utility sum of her actions. 'W.D. Ross [30] argues that no absolute moral theory can be developed and suggests instead that a set of principles, or *prima facie* duties is used whenever possible: fidelity, reparation, gratitude, non-injury, harm-prevention, beneficence, self-improvement and justice.

Kant suggests that a moral agent follows a set of categorical imperatives which are maxims that are sufficiently virtuous to be used by everyone at every context. Here the principle of double effect should also be mentioned [25]. According to this principle (or doctrine), unethical actions can be sometimes permissible as a side effect of pursuing a moral action. Those same "bad" actions would not be permissible when they are the *means* to accomplishing the same moral action. In general, these theories are ones in which the intentions of the actor are important in determining the ethics of an action. A variation of these theories are ones in which actions themselves have ethical force. Deontic logics [20] that specify the actions an agent is obliged to take or prohibited from taking are well studied and supported by a variety of programming frameworks which have been applied to normative reasoning in general not just ethical reasoning.

V. Giving Machines the Capacity to Know Right

All machine reasoning systems can be viewed as ethical reasoning systems at some level of abstraction. We survey the key contribution systems that are explicitly ethical [26].

A. GenEth

The GenEth system [1] has two purposes. Firstly, it demonstrates how input from professional ethicists can be used, via a process of machine learning, to create a *principle of ethical action preference*. GenEth analyses a situation in order to determine its ethical features (*e.g.,* that physical harm may befall someone). These features then give rise to *prima facie* duties (to minimize or maximize that feature). In this theoretical framework GenEth is explicitly adopting Ross' theory of *prima facie* duties.

The principle of ethical action preference is used to compare two options: each option is assigned a score for each ethical feature, the scores are then used by the principle to determine the appropriate course of action based on which, duties are of more importance given the other duties effected. *E.g.,* the system might prefer an action which had worse consequences for privacy on the grounds it was better for safety.

GenEth can "explain" its decisions in terms of its preferences over duties – so it can state how two options compared on the various ethical features and refer to the statement of the principle. It is important to emphasize this feature of *explainability* particularly since GenEth uses machine learning as part of the process by which its ethical behaviour is determined. Machine learning systems, in general, are not particularly transparent to users, but some can be made so.

B. $\mathcal{DCEC}_{CL}$

Bringsjord et al. have a body of work [8], [9], developing the *deontic cognitive event calculus,* $\mathcal{DCEC}_{CL}$, in which various ethical theories can be expressed. A key motivation is a belief that ethical reasoning must necessarily be implemented at the operating system level. Concepts in the $\mathcal{DCEC}_{CL}$ are expressed in explicitly deontological terms – *i.e.,* as obligations, permissions and prohibitions.

An illustrative example of the $\mathcal{DCEC}_{CL}$ approach is the Akratic robot [9]. This considers a scenario in which a robot charged with guarding a prisoner of war must choose whether or not to retaliate with violence to an attack. [9] argues that the underlying robot architecture, into which the modules for self-defence and detainee management have been embedded, must be capable of ethical reasoning in order to predict and prevent ethical conflicts.

$\mathcal{DCEC}_{CL}$ uses automated reasoning to deduce ethical courses of action by reasoning explicitly about its obligations, prohibitions and so on. Automated reasoning, also referred to as automated theorem proving, has a long history in AI [29], with particular attention paid to implementations with high degrees of assurance. As a result automated reasoning with $\mathcal{DCEC}_{CL}$ can be considered *correct by virtue of the reasoning process* so long as the concepts supplied correctly capture the values of the community the system is designed to serve.

C. Ethical Governors

Arkin et. al [2], [3] outline the architecture for an *ethical governor* for automated targeting systems. This governor is charged with ensuring that any use of lethal force is governed by the "Law of War", the "Rules of Engagement". This initial work on was then re-implemented in a new setting of healthcare [32]. The governor is implemented as a separate module that intercepts signals from the underlying deliberative system and, where these signals involve lethality, engages in

a process of *evidential reasoning* which amasses information about the situation in a logical form and then reasons using prohibitions and obligations. If any prohibitions are violated or obligations unfulfilled then the proposed action is vetoed.

The authors note that "it is a major assumption of this research that accurate target discrimination with associated uncertainty measures can be achieved despite the fog of war". It should be noted that throughout the literature on machine ethics there is an assumption seldom explicitly stated as it is in Arkin's work that complex, sometimes highly nuanced, information is available to the ethical reasoning system in order for it to make a determination. A key open area of research in machine ethics would seem to be the development of techniques for *ethical situation awareness*. The explicit use of evidential reasoning is an important step towards developing such techniques but only part of the story.

Unlike $\mathcal{DCEC}_{CL}$, the reasoning used by Arkin's ethical governors is not grounded in a formal logical theory. Ad-hoc reasoning techniques are therefore used rather than ones derived from automated theorem proving – as such deductions can not be assumed correct by virtue of the reasoning process.

D. Ethical Consequence Engines

Winfield et. al [34] have investigated systems based on the concept of an *Ethical Consequence Engine*. Ethical consequence engines are grounded in consequentialist theories of ethics, particularly utilitarianism. Like ethical governors, ethical consequence engines, pay attention to the ethical information upon which reasoning is based. Given they are using utilitarian ethics the question becomes one of generating appropriate utilities for each action.

The consequence engines use simulation to evaluate the impact of actions on the environment. In particular they simulate *not just* the actions of a robot itself but the activity of other agents in the environment. This allows the robot to determine not only if its actions have directly negative consequences (*e.g.,* colliding with a person) but if they have indirectly negative consequences *e.g.,* failing to intercept a person who might come into danger). The ethics implemented in each system thus has a distinctly Asimovian flavour, as directly acknowledged in [34]. The implemented ethical system can be seen as a combination of utilitarianism and Asimov's Laws.

E. ETHAN

The ETHAN system [13] was developed to investigate ethical decision making in exceptional circumstances. In ETHAN a *rational agent* [28] reasons about the ethical risks of plans proposed by an underlying planning system. The operation of reasoning in normal circumstances is assumed to be ethical by default (*i.e.,* that the agent is implicitly ethical), but in exceptional circumstances the system might need to make use of techniques such as planning or learning whose behaviour is difficult to analyse in advance.

[13] considers the case of a planning system that returns candidate plans to the agent which are annotated with context specific ethical concerns. These concerns are then reasoned about using a priority-based context specific *ethical policy* that prefers plans violating lower priority concerns to plans violating higher priority concerns and, where two plans violate concerns of the same priority, prefers the plan violating the fewest concerns. As with GENETH, ETHAN's ethics are based on Ross's *prima facie* duties [30] and ETHAN's ethical principles can be considered broadly similar to GENETH's ethical features.

F. HERA

The hybrid ethical reasoning agent (HERA) system [24] uses a model theoretic approach to investigate the implementation of different ethical theories. Its primary focus has been constructing a rich framework that can express both Utilitarian and Kantian/Deontological systems – in particular the categorical imperative [6] and the principle of double effect [5].

For each action available to it, HERA builds a model depicting the overall utility of the action, as well as whose utilities are affected (positively or negatively) and which agents are *ends* of the action and which are affected as *means* to those ends. These models have a formal basis allowing automated reasoning to determine whether some logical formula is satisfied by the model, so again this reasoning can be considered correct by virtue of the reasoning process.

In the case of utilitarianism HERA compares all models and selects the one with the highest overall utility. In the case of the categorical imperative and principle of double-effect it constructs a logical formula expressing the ethical constraints and then vetoes models which do not satisfy the formula.

VI. ENSURING A MACHINE CAN NOT DO WRONG

Formal verification is the process of assessing whether a formal specification is satisfied on a particular formal description of a system. For a specific logical property, φ, there are many different approaches to this [18], [12], [7], ranging from deductive verification against a logical description of the system ψ_S (*i.e.,*$\vdash \psi_S \rightarrow \varphi$) to the algorithmic verification of the property against a model of the system, M (*i.e.,*$M \models \varphi$). The latter has been extremely successful in Computer Science and AI, primarily through the *model checking* approach [11]. This takes a model of the system in question, defining all the model's possible executions, and then checks a logical property against this model.

The approach most often applied to the verification of machine ethics is a model-checking approach for the verification of agent-based autonomous systems outlined in [19] which considers the decision taken by the system given any combination of incoming information. This methodology adapts well if we can implement an ethical decision agent on top of an underlying autonomous system which accepts processed ethical information as input. We note that this is the architecture adopted in most of the systems we have described. A model-checker can then verify that such a system always chooses options that align with a given code of ethics based on the information that it has. This approach has been applied both to the verification of ETHAN programs [13] and to the verification of ethical consequence engines [14].

In ETHAN programs the emergency planning system was replaced by a random component that generated plans annotated as violating some combination of ethical concerns. The model-checking process then ensured that all such combinations were considered. Given a ranking of concerns according to some ethical policy the verification was able to show that a plan was only selected by the system if all other plans were annotated as violating some more serious ethical concern. In [14] a simplified model of the ethical consequence engine was constructed on a 5x5 grid. This was used to check the decision making as in the ETHAN system. In an extension, a probabilistic model of the human behaviour was also created in order to use a probabilistic model-checker (PRISM [23]) to generate probabilities that the robot would successfully "rescue" a human given any combination of "human" movement on the grid. The results of this verification differed greatly from the probabilities generated through experimental work in a large part because the model used in verification differed significantly, in terms of the environment in which the robot operated to the environment used experimentally.

HERA and $\mathcal{DCEC}_{CL}$ use formal logical reasoning in order to make ethical choices – model checking in HERA and theorem proving in $\mathcal{DCEC}_{CL}$. For simple models/formulae it is easy to rely on the correctness of this reasoning to yield correct results but we note that for more complex models this is more challenging. Even in systems that perform ethical reasoning that is correct by virtue of the reasoning process, it may be necessary to verify some "sanity" properties.

VII. CONCLUSIONS

We here attempted to survey the current state of the art in the implementation and verification of machine ethics having noted that, unlike human reasoning, we require machine ethical reasoners not only to know which is the correct action, but also then act in accordance with that knowledge. We have restricted ourselves to explicitly ethical systems which reason about ethical concepts as part of the system operation. While the field of practical machine ethics is still in its infancy, it is thus possible to see some clear convergence in approaches to implementation and consensus about the need for strong assurances of correct reasoning.

REFERENCES

[1] M. Anderson and S. Leigh Anderson. Geneth: A general ethical dilemma analyzer. In *Proceedings of the 28th AAAI Conference on AI, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 253–261, 2014.

[2] R.C. Arkin, P. Ulam, and B. Duncan. An Ethical Governor for Constraining Lethal Action in an Autonomous System. Technical report, Mobile Robot Laboratory, College of Computing, Georgia Tech., 2009.

[3] R.C. Arkin, P. Ulam, and A. R. Wagner. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proc. of the IEEE*, 100(3):571–589, 2012.

[4] I. Asimov. *I, Robot*. Gnome Press, 1950.

[5] M.M. Bentzen. *The principle of double effect applied to ethical dilemmas of social robots*, pages 268–279. IOS Press, 2016.

[6] M.M. Bentzen and F. Lindner. A formalization of kant's second formulation of the categorical imperative. *CoRR*, abs/1801.03160, 2018.

[7] R. S. Boyer and J. Strother Moore, editors. *The Correctness Problem in Computer Science*. Academic Press, London, 1981.

[8] S. Bringsjord, K. Arkoudas, and P. Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4):38–44, 2008.

[9] S. Bringsjord, N Sundar, D. Thero, and M. Si. Akratic robots and the computational logic thereof. In *Proc. of the IEEE 2014 Int. Symposium on Ethics in Engineering, Science, and Technology*, pages 7:1–7:8, Piscataway, NJ, USA, 2014.

[10] V. Charisi, L.A. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A.F.T. Winfield, and R. Yampolskiy. Towards moral autonomous systems. *CoRR*, abs/1703.04741, 2017.

[11] E. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, 1999.

[12] R. A. DeMillo, R. J. Lipton, and A.J. Perlis. Social Processes and Proofs of Theorems of Programs. *ACM Communications*, 22(5):271–280, 1979.

[13] L. A. Dennis, M. Fisher, M. Slavkovik, and M. P. Webster. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

[14] L. A. Dennis, M. Fisher, and A. F. T. Winfield. Towards Verifiably Ethical Robot Behaviour. In *Proceedings of AAAI Workshop on AI and Ethics*, 2015.

[15] S. Dyrkolbotn, T. Pedersen, and M. Slavkovik. On the distinction between implicit and explicit ethical agency. In *AAAI/ACM Conference on AI, Ethics and Society*, New Orleans, USA, 2018.

[16] J. W. Ellington. *Translation of: Grounding for the Metaphysics of Morals: with On a Supposed Right to Lie because of Philanthropic Concerns by Kant, I. [1785]*. Hackett Publishing Company, 1993.

[17] A. Etzioni and O. Etzioni. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, pages 1–16, 2017.

[18] J. H. Fetzer. Program Verification: The Very Idea. *ACM Communications*, 31(9):1048–1063, 1988.

[19] M. Fisher, L. Dennis, and M. Webster. Verifying Autonomous Systems. *ACM Communications*, 56(9):84–93, 2013.

[20] D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors. *Handbook of Deontic Logic and Normative Systems*. College Publications, London, UK, 2013.

[21] J.C. Harsanyi. Rule utilitarianism and decision theory. *Erkenntnis (1975-)*, 11(1):25–53, 1977.

[22] R. Hursthouse and G. Pettigrove. Virtue ethics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[23] M. Kwiatkowska, G. Norman, and D. Parker. PRISM: Probabilistic Symbolic Model Checker. In *Proc. 12th Int. Conf. Modelling Techniques and Tools for Computer Performance Evaluation (TOOLS)*, volume 2324 of *LNCS*, 2002.

[24] F. Lindner and M.M. Bentzen. The hybrid ethical reasoning agent IMMANUEL. In *Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*, pages 187–188, 2017.

[25] A. McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter edition, 2014.

[26] J. H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.

[27] S.Y. Nof. *Handbook of Industrial Robotics*. Number v. 1 in Electrical and electronic engineering. Wiley, 1999.

[28] A. S. Rao and M. P. Georgeff. BDI Agents: From Theory to Practice. *Proc. of the First International Conference on Multiagent Systems*, 95:312–319, 1995.

[29] A. Robinson and A. Voronkov, editors. *Handbook of Automated Reasoning*. Elsevier Science Publishers B. V., 2001.

[30] W.D. Ross. *The Right and the Good*. Oxford University Press, 1930.

[31] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[32] J. Shim and R. C. Arkin. An Intervening Ethical Governor for a Robot Mediator in Patient-Caregiver Relationships. In M. I. Aldinhas Ferreira et al., editor, *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, pages 77–91. Springer Int. Publishing, 2017.

[33] H. Simonis. Constraints in computational logics. chapter Building Industrial Applications with Constraint Programming, pages 271–309. Springer-Verlag New York, Inc., 2001.

[34] D. Vanderelst and A. Winfield. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 2017.

[35] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2008.

[36] W. Wallach, C. Allen, and I. Smit. Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI Society*, 22(4):565–582, 2008.