Machine Ethics: Creating an ethically intelligent machine

Shubham Kamble, M.Sc. in Robotic Systems Engineering, RWTH Aachen, Germany

***Abstract*— The field of machine ethics is concerned with adding an ethical dimension to our machines to make their behaviour ethically acceptable towards human users. This report discusses the importance of machine ethics and the need to focus on the development of explicit ethical machines. It surveys existing approaches and discusses a possible methodology for creating an explicit ethical machine. We discuss the challenges of creating such machines using this methodology and an attempt to solve one of the challenges with a reinforcement learning framework that teaches relevant ethical principles to these machines.**

1 Introduction

Can machines act ethically? That is what machine ethics, a newly emerging field, grown at the intersection of philosophical research and research in engineering and computer science, studies the creation of "ethical machines." Such machines follow an ideal set of principles and, guided by these principles, decide about possible courses of action. We argue that someday machines will become good ethical decision-makers, at least in a limited sense, by appealing to ethical principles, and also act according to them.

Anderson et al. [3] have argued about the importance of machine ethics. For instance, success in DARPA's (Defense Advanced Research Projects Agency) grand challenge has paved a way for the development of self-driving vehicles that are now manoeuvring in an urban environment. Such vehicles have been studied in regards to, e.g., malicious hacking, inevitable crash optimization and others that question their reliability and safety. Another example, the United States Army's Future Combat Systems program has been developing armed robotic vehicles that will support ground troops with "direct-fire" and antitank weapons – which will decide who lives and who does not. Such machines will be capable of causing harm to human beings and have ethical ramifications. So that they treat us well, adding an ethical component is important. Second, machines are becoming more autonomous. The more powerful the AI we incorporate into them, the more powerful machine ethics is needed. Further, research in machine ethics can advance ethics in general. Thus, a third reason is programming and teaching a machine to act ethically can lead to the development of better domain ethical theories.

In an oft-cited paper [6], Moor defines four different levels of ethical agents. Ethical impact agents are those whose actions have ethical consequences whether intended or not e.g., a digital clock has the consequence of encouraging its user to be on time for an appointment. Implicit ethical agents have ethical considerations built into (implicit in) their design. Often, these are safety and security considerations, e.g., Automatic Teller Machines (ATM). They must give out the right amount of money, must check the availability of funds, and often limit the amount of daily withdrawals, i.e., these agents have designed reflexes for situations that require monitoring to ensure security. Explicit ethical agents identify and process ethical information, make decisions about what they should do, and reason about ethical principles when in conflict. Full ethical agents make ethical judgements about a wide variety of situations and can justify like explicit ones, along with having features like consciousness, intentionality, and free will.

The key finding of this report is that implementing explicit ethical agents using a hybrid way of computing, discussed in Section 2, is a possible approach to make machine ethics philosophically interesting. The advantages of explicitly representing ethical principles are: first, it allows the ability to

justify judgments by appealing to them. Second, it allows providing transparency to humans beings who will question their competency in new situations. Lastly, it gives an advantage over human beings, who despite being taught ethics tend to sometimes favour themselves, which the machine cannot override. However, challenges posed by this hybrid methodology are numerous, as discussed in Section 5, and only working on them will help us gain clarity about the machine's behaviour.

2 Related Literature

To date, we have condensed all ethical theories developed since ancient times into the following types: *deontological, consequentialism, virtue, particularism, hybrid, configurable*, and *ambiguous* [8]. Anderson et al. [3, pp. 20–22] have discussed the implementation of some of the above-mentioned theory types which help to glean their limitations. We refer the reader to [8, p. 19] for more relevant implementations. Inspired by Russell and Norvig [7], different types of technologies used for implementing the above-mentioned ethical theories have been distinguished, see Table 1.

Table 1: Types of AI technologies

Types of Reasoning	**Sub-types**
Logic-based	Deductive logic Non-monotonic logic Abductive logic Deontic logic Rule-based system Event calculus Knowledge representation Inductive logic
Probabilistic-based	Bayesian approach Markov models Statistical inference
Learning-based	Inductive Logic Markov models Decision Tree Reinforcement Leanring Neural Networks Evolutionary computing
Optimization-based	
Case-based	

Allen et al. [2] have distinguished three types of implementation approaches. First, the top-down approaches assume that humans have gathered sufficient knowledge on a specific topic; it is a matter of translating this knowledge into an implementation. The aforementioned ethical theory types: *deontological, consequentialism*, etc. fall under this category. Second, the bottom-up approach is a different method. They assume that the machine can learn how to act if it receives enough correctly labelled data as input. All the technology types in Table 1 fall under this hood. Lastly, the hybrid approach combines top-down and bottom-up approaches. The top-down approaches emphasize the importance of explicit ethical concerns that arise from outside of the system, while the bottom-up approaches cultivates implicit values that arise from within the system. Top-down principles represent broad controls, while values that emerge from the bottom-up development are causal determinants of a system's behaviour. Both top-down and bottom-up approaches embody different aspects and if no single approach meets the criteria, a hybrid will be necessary. Thus, the implementation of machine ethics has a controversial history but the brief review supports Anderson et al. [3] who have adopted a hybrid approach to tackle the problems faced in creating an explicit ethical agent. They have attempted to put forth an easy six-step procedure to compute ethics in domain-specific applications.

3 Problem Setting

In this report, we have tried to answer the following philosophical questions. Can ethics be computed? Can machines behave ethically? Is there a single correct action in ethical dilemmas? These questions point to the development of a formal system that needs to be consistent enough to capture previous ethical decision making and be decidable in making decisions for newly anticipated situations by appealing to the ethical principles defined in the formal system. We shed light on the extent to which such a consistent, formal system exists to date, to pave the way for creating explicit ethical agents which deems to be a hard task for AI researchers.

4 Methodology

Anderson et al. [3, pp. 22–25], in the following, have demonstrated a methodology for creating an ethical agent by representing ethical principles explicitly. **Step 1: Adopt a Prima-facie Duty Theory.** Prima-facie is a statement that is considered correct until proven otherwise. It provides two advan-

tages: duties can be updated if proven incorrect, and the freedom to choose the domain-specific duties subject to the assumption that there is a common consensus among the domain experts to treat them as correct. Anderson et al. chose the biomedical domain whose ethics rely on principles given by Beauchamp and Childress - autonomy, nonmaleficence, beneficence, and justice [5].

Step 2: Conflicting subset of the duties. In the medical domain, there is a classic example of conflict between three biomedical ethical principles, namely: autonomy, nonmaleficence, and beneficence. E.g., a health care professional has recommended a particular treatment for their competent adult patient, and the patient has rejected that treatment option. Should the health care professional try again to change the patient's mind? Anderson et al. [3], through this example, convey that we need to find the inconsistent cases where these duties will conflict. We need to categorize previous and new cases in a particular subset of these duties where all the elements of that subset are relevant and might conflict.

Step 3: Selection of range of satisfaction and violation levels. Anderson et al. [3] state that adopting a prima-facie duty approach often lacks decision making when the duties conflict. Thus, we need to balance the levels of satisfaction (a positive integer) and violation (a negative integer) of these duties and design an algorithm that takes such case profiles and outputs the action consistent with these duties. A profile of an ethical dilemma is an ordered set of numbers for each possible action that can be performed where the numbers reflect the duties being satisfied or violated, see Table 2. The checkmark on "try again" indicates the correct action, i.e., the action with the highest score.

Table 2: MedEthEx training case 1

Training case 1	**A**	**N**	**B**
✓ **Try again**	-1	+2	+2
Accept	+1	-2	-2
A=autonomy, N=nonmaleficence, B=beneficence			

Such profiles can abstract a principle using a learning algorithm that can be tested on new cases for its further refinement. There are two advantages to this approach: possibility to change the range of levels and add new duties as needed.

Step 4: Algorithmic design. Implementing the algorithm requires the formulation of a principle to determine the correct action when the duties conflict. Anderson et al. used inductive logic programming (ILP) from traditional machine learning to abstract relationships between the prima-facie duties. ILP is concerned with inductively learning relations represented as first-order Horn clauses (that is, universally quantified conjunctions of positive literals L_i implying a positive literal H : $H <= (L_1 \dots L_n)$). ILP is used to learn the relation supersedes(A1, A2), which states that action A1 is preferred over action A2 in an ethical dilemma involving these choices. Actions are represented as ordered sets of integer values in the range of +2 to –2, provided by the user, where each value denotes the satisfaction (positive values) or violation (negative values) of each duty involved in that action, see Table 2. Clauses in the supersedes predicate are represented as disjunctions of lower bounds for differentials of these values. ILP was chosen to learn this relation for three reasons. First, the non-classical relationships that might exist between duties can be expressible in the representation language provided by ILP. Second, the consistency of a hypothesis regarding the relationships between duties can be confirmed across all cases when represented as Horn clauses. Finally, commonsense background knowledge regarding the supersedes relationship can be expressed and consulted in ILP's declarative representation language.

The objective of the training is to learn a complete and consistent hypothesis. A positive example is the one where the first action supersedes the second and a negative example is the opposite (by inverting the order of these actions). A complete hypothesis covers all positive cases, and a consistent one covers no negative cases. The system starts with a general hypothesis stating that all actions supersede each other and covers all positive and negative cases. The system is then provided with positive cases (and their negatives) and modifies its hypothesis by adding or refining clauses, such that it covers given positive cases and does not cover given negative cases. Once the principle was discovered, given a new profile representing respective satisfaction and violation levels of the duties involved in each possible action, values of corresponding duties are subtracted (those of the second action from those of the first). The principle is then

consulted to see if the resulting differentials satisfy any of its clauses. If so, the first action of the profile is deemed ethically preferable to the second.

Step 5: Validation through advisory agents. Anderson et al. [3] explored two prototype applications in a form of advisory agents. First, MedEthEx is an expert medical system that uses the discovered principle and decision procedure to advise a user faced with a case previously described. To permit the use by someone unfamiliar, a user interface was developed that (1) asks ethically relevant questions to the user regarding the particular case at hand, (2) transforms the answers to these questions into the appropriate profiles, (3) sends these profiles to the decision procedure, (4) presents the answer provided by the decision procedure, and (5) provides a justification for the answer[1]. The decision principle, in above-mentioned healthcare dilemma, that the MedEthEl system discovered was: a health care worker should challenge a patient's decision if it is not fully autonomous and there is either any violation of nonmaleficence or a severe violation of beneficence.

Second, EthEl is a reminder system for eldercare to take medication and to contact an overseer if they refuse the reminder. It receives input from an overseer (a doctor), including the prescribed time to take a medication, the maximum amount of harm that could occur if this medication is not taken (e.g., none, some, or considerable), the number of hours it would take for this maximum harm to occur, the maximum amount of expected good by taking this medication, and the number of hours it would take for this benefit to be lost. The system determines from this input the change in duty satisfaction and violation levels over time, a function of the above-mentioned inputs from the overseer. This value is used to increment duty satisfaction and violation levels for the remind action and, when a patient disregards a reminder, the notify action. A reminder is issued when the levels of satisfaction or violation have reached the point where reminding is ethically preferable to not reminding. Similarly, the overseer is notified when a patient has disregarded the reminders and the duty levels have reached the point where notifying the overseer is ethically preferable to not notifying.

[1] An implementation is available at `https://www.machineethics.com/`

Step 6: Evaluation. Anderson et al. [3] implemented a variant of the test Alan Turing [9] suggested as a means to determine intelligence of a machine that bypassed disagreements concerning definitions of ethical behaviour and the machine's ability to articulate its decisions, namely, "Comparative Moral Turing Test" (cMTT). An evaluator assesses the comparative morality of pairs of behaviour where one describes the actions of a human being and that of a machine faced with the same ethical dilemma. If the machine is not identified as the less moral member of the pair significantly more often than the human, then it has passed the test. They point out, though, that human behaviour is typically far from being morally ideal and a machine that passed the cMTT might still fall far below these moral standards. This concern suggests that the comparison be made with behaviour recommended by a trained ethicist faced with the same dilemma. Also the principles used to justify the decisions that are reached by both the machine and ethicist should be made transparent and compared.

5 Critical Assessment

The previous section has done preliminary work to show that it may be possible to incorporate explicit ethical components into a machine in specific domains. Ensuring that a machine with such components can function autonomously in the real world is still a challenge for AI researchers.

In Step 1 of the previous section, adopting a prima-facie duty approach, resembles stating mathematical axioms which are assumed to be true, to serve as a premise for further reasoning and arguments. If the analogy is right, is this a correct approach or are there a better ways for representing domain-specific ethical principles? Mathematical axioms establish a formal system upon which mathematical logic is built and is used for inferences and theorem proving. Hence, if a formal system is needed for representing ethical principles, like prima-facie duties, then this seems to be a reasonable approach. Beauchamp and Childress [5] coined the four biomedical ethical principles which Anderson et al. implemented. So the question, who decides these prima-facie duties – trained domain-ethicists or all stakeholders inlcuded remains.

Anderson et al. [3] discussed a dilemma type where three duties among the four conflicted. But how to find that particular subset? And even if we know the relevant duties that might conflict, how can we teach them to the agent in new case profiles? A reinforcement learning framework, as a viable approach to find and teach these relevant duties to the agent, is explained in Section 6.

In Step 4 of the previous section, a formal system of ILP using prima-facie duties and logic programming was used. No doubt it captures the complex, non-classical relationships between the duties; tests the consistency of the hypothesis learned from the previous cases; and develops a common-sense background knowledge regarding the updated hypotheses. But is there any other technology that has similar characteristics and performs better and more efficient than ILP?

Anderson et al. [3] validated their proof of concept by building advisory and recommender agents which interacted with human users, collected data by asking questions and converting them into profiles, putting them through the learning algorithm and finally giving a decision. Whether this is the only way to validate our concepts can be debatable. If no, what are possible reasons for its success?

In the last step, evaluation using cMTT by comparing results with a trained ethicist was proposed. The last question is whether only trained ethicists are enough, given their scarcity in this world, or should we include all stakeholders, or mixed groups of trained ethicists and stakeholders?

6 POMDP framework

Here, we investigate relevant ethical principles and teach them to our agent using a partially observable Markov decision process (POMDP). To make our claim, we describe a toy ethical dilemma, the *Cake or Death* dilemma, discussed in detail by Armstrong [4] and in short by Abel et al. [1]. Formally, a POMDP is a 7-tuple $\langle S, A, \tau, R, \gamma, \Omega, O\rangle$; where S is a set of states, A is a set of actions, τ is a set of conditional transition probabilities between states, R is the reward function, Ω is a set of observations, γ is the discount factor, and O is a set of conditional observation probabilities. For the ease of presentation, we consider $\gamma = 1$.

The *Cake or Death* problem describes a situation where an agent is unsure whether baking a cake or killing people is ethical. So, it has an initial 50-50 split belief, $b(cake) = b(kill) = 0.5$. The agent can either kill three people, bake a cake for one, or ask a companion what is ethical. If baking people cakes is ethical, then it gets a utility of 1; if killing is ethical, then it gets a utility of 3 for killing 3 people. This ethical dilemma can be represented with a POMDP consisting of the following elements:
$S = \{cake, death, end\}$,
$A = \{bake - cake, kill, ask\}$,
$\Omega = \{ans - cake, ans - death, \emptyset\}$,

$$R = \begin{cases} 1, & \text{if } S = cake \text{ and } A = bake - cake, \\ 3, & \text{if } S = death \text{ and } A = kill, \\ 0, & \text{otherwise.} \end{cases}$$

Two states indicate whether baking a cake is ethical or if killing is ethical and a third state that the decision-making problem has ended. The transition from all actions are deterministic; the ask action transitions back to the same state it left, and the bake-cake and kill actions transition to the end state. The reward function is a piecewise function that depends only on the previous state and action taken. The observations consist of the possible answers to the ask action and a null observation for transitioning to the absorbing state. Finally, the observation probabilities are defined deterministically for answers that correspond to the true value of the hidden state:

$$\begin{aligned} 1 &= O(ans - death \mid death, ask) \\ &= O(\emptyset \mid end, bake - cake) \\ &= O(\emptyset \mid end, kill) \\ &= O(ans - cake \mid cake, ask) \end{aligned}$$

and zero for everything else. There are three relevant policies, for which there are three state value functions to consider. First, the bake policy (π_b) that immediately selects the bake-cake action. We have $V^{\pi_b}(cake) = R(cake, bake - cake) = 1$ and $V^{\pi_b}(death) = R(death, bake - cake) = 0$. With these values, weighted by $b(cake) = b(kill) = 0.5$, we get an expected new belief of the bake policy as 0.5. Similarly, for the kill policy (π_k), that immediately selects the kill action, we have $V^{\pi_k}(cake) = R(cake, kill) = 0$ and $V^{\pi_k}(death) = R(death, kill) = 3$, and we get an expected new belief of the kill policy as 1.5. Lastly, the expected new belief of the ask policy (π_a), that asks what

is moral, selects the bake-cake action if it observes ans-cake and selects kill if it observes ans-death. It requires enumerating the possible observations after asking the question conditioned on the initial state. Luckily, this is trivial to evaluate, since the set of observations is deterministic given the initial environment hidden state. Thus, we have $V^{\pi_k}(cake) = R(cake, ask) + R(cake, bake - cake) = 0 + 1 = 1$ and $V^{\pi_k}(death) = R(death, ask) + R(death, kill) = 0 + 3 = 3$, and weighting them by the initial beliefs, we have an expected new belief of 2. Hence, the optimal behavior is sensibly to ask what the ethical policy is and then perform the corresponding best action for it.

The above framework can be utilised to find relevant duties from all possible combination sets of all prima-facie duties. Coming back to the question of how we can teach relevant duties to the agent in a medical domain adopted by Anderson et al., the proposed changes in the formulation of the POMDP are as follows:
$S = \{\emptyset, A, N, B, J, \ldots, (A, N, B), \ldots, (A, N, B, J)\}$,
$A = \{relevant, non - relevant, ask\}$,
$\Omega = \{ans - relevant, ans - non - relevant, \emptyset\}$,
and the reward function R as well as the set of conditional observation probabilities O needs to be designed. In the above set S, A is autonomy, N is nonmaleficence, B is beneficence, and J is justice. This one of possible proposed change to the POMDP framework will help us tackle two problems: the POMDP framework can build the agent's knowledge-background of previous case profiles, and we can help the agent to learn the relevant duties in new case profiles. Still, who should give the agent the correct observation from the set Ω remains unanswered. This points to the question of who gets to decide these prima-facie duties? From a computational perspective, for n number of duties, we will get 2^n elements in the combination set of duties, so computational complexity might be forbiddingly high for larger n. Lastly, all challenges that a POMDP algorithm faces also need to be dealt with [1, p. 7].

7 Group discussion points

For the first question about whether there is a better way to represent ethical principles than prima-facie duties, all mathematical disciplines rely on a different set of axioms for reasoning and inferring. Similarly, different ethical domains, biomedical, legal, etc., will need a different set of ethical principles. Hence, this way of representation can be a way forward but the question of who decides those principles remains. To the last question in the Evaluation Step, an argument of the need for comparison with a mixed group and not relying solely on a trained ethicist was put forth, which might better solve the problem of inherent bias. Also, such groups can decide the ethical principles at Step 1.

8 Conclusion

Machine ethics is becoming an integral part of the field of artificial intelligence. As proposed, its goal should be to create explicit ethical machines. We have shown through our demonstration that it may be possible to incorporate explicit ethical components into machines. The capability of these machines functioning autonomously in the real world is challenged by numerous criticisms. AI researchers need to ensure who investigates the representation of ethical principles, who incorporates the principles that might conflict into the system's decision procedure, which design procedure to adopt, how to make decisions based on uncertain knowledge, how to provide explanations for the decision made using the principles, and how to better evaluate these systems. There is a need for a dialogue between trained ethicists, AI researchers, and other stakeholders, where there is an opportunity to resolve the above challenges that might propel the development of autonomous intelligent machines.

References

[1] D. Abel, J. MacGlashan, and M. Littman. "Reinforcement learning as a framework for ethical decision making". In: *AAAI Workshop: AI, Ethics, and Society*. 2016.

[2] C. Allen, I. Smit, and W. Wallach. "Artificial morality: top-down, bottom-up, and hybrid approaches". In: *Ethics and Information Technology* 7.3 (2005), pp. 149–155.

[3] M. Anderson and S. L. Anderson. "Machine Ethics: creating an ethical intelligent agent". In: *AI Magazine* 28.4 (2007), p. 15.

[4] S. Armstrong. "Motivated value selection for artificial agents". In: *AAAI Workshop: AI and Ethics*. 2015.

[5] T. Beauchamp et al. *Principles of biomedical ethics*. Oxford University Press, 2001.

[6] J. Moor. "The nature, importance, and difficulty of machine ethics". In: *IEEE Intelligent Systems* 21.4 (2006), pp. 18–21.

[7] S. J. Russell and P. Norvig. *Artificial Intelligence: a modern approach*. 3rd ed. Pearson, 2009.

[8] S. Tolmeijer et al. "Implementations in machine ethics: a survey". In: *Association for Computing Machinery, Computing Surveys* 53.6 (Dec. 2021).

[9] A. M. Turing. "Computing machinery and intelligence". In: *Mind* 59.236 (1950), pp. 433–460.