The revised draft is forthcoming in *AI & Society*
For the final published version, please go to
https://link.springer.com/article/10.1007/s00146-022-01601-0

A Pluralist Hybrid Model For Moral AIs

1. Introduction

It goes without saying that developing moral machines is a pressing task for AI research. With the explosion of possible contexts of robot-human interactions, we have every reason to make sure machines shall at least work to minimize moral risks to humans, if not think morally. Failure to produce sufficiently moral behaviour in machines can be *morally disastrous* for humanity because, unlike moral failures of individual human beings, machine failures are likely to be more systematic (when machines with faulty programming are deployed widely) or far-reaching (when machines are tasked with important decisions, e.g. in autonomous weaponry, or social decision-making). Yet, despite the pressing need for such a task, the challenges are equally arduous. Many authors have suggested the imperfectability of the project of moral AIs. Brundage (2014), for example, synthesizes the concerns in the literature and argues that even if most technical obstacles are overcome, the project cannot ensure positive social outcomes from intelligent systems due to the nature of ethics, the computational limitations and complexity of the world.

In this paper, we synthesize the inherent limitations discussed in the literature with a particular focus on three possible approaches: 1) the top-down approach (e.g. deterministic algorithm model); 2) the bottom-up approach (e.g. machine learning model); and 3) hybrid systems (e.g. algorithm + machine learning) in *sections 2* and *3*. In section 4, we propose a novel approach called a pluralist hybrid system. The pluralist hybrid system has two features. First, it has a deterministic algorithm system that embraces different moral rules as action guidance. The deterministic algorithm system is responsible for making explicit moral decisions. Second, it has a machine learning system that is responsible for calculating the value of the parameters required by the application of moral principles. In sections 4 and 5, we argue that the pluralist hybrid system is better than the existing proposals in two aspects. First it better addresses the moral disagreement problem of the top-down approach. Second, it reduces the opacity of the system to a justifiable level compared with bottom-up models.

2. The Problems of Two Approaches to Moral AI

In this section, we synthesize the inherent limitations discussed in the literature with a particular focus on the two possible approaches: 1) the top-down approach (e.g. deterministic algorithm model); and 2) the bottom-up approach (e.g. machine learning model). According to the distinction made by Wallach and Allen (2009), the top-down approaches to moral AI refer to approaches that program a specific ethical theory in an AI system and apply them to different scenarios. For example, a system can program utilitarian moral principles at the top level, and analyze its informational and computational requirements necessary to implement the theory in specific cases. Cloos (2005) proposes an Utilibot model, a decision-

theoretic autonomous mobile robot guided by the utilitarian aim of maximizing human well-being, which is expected to act morally in decisions related to human life and health. This model consist of two dynamic Bayesian networks that model human and environmental health, a dynamic decision network that accounts for decisions and utilities, and a Markov decision process (MDP) that breaks down the planning problem to solve for the optimal course of action to maximize human safety and well-being. These sub-systems are designed to overcome the limits of human reasoning in a computational system.

Brundage (2014) criticizes that a utilitarian moral AI is problematic for three reasons. First, utilitarianism, as a normative ethical theory, is subjected to several criticisms. Utilitarianism is often challenged by its rivals for its failure to address population ethics, consider distributive justice and its endorsement of utility aggregation. Second, there are many reasons to believe that utilitarianism cannot capture significant moral intuitions, for example most people believe that it is morally impermissible to push the fat man over the bridge, even though doing so can save five lives on the trolley track. Third, a utilitarian model could be plausible for context-specific applications of AI (or "weak AI" such as Cloos's Utilibot for health care). However, a utilitarian strong AI would be potentially more dangerous. Unchecked, a utilitarian AI may promote intentional killing at a social level if doing so can save more lives.

Power (2011) proposes a deontological model which incorporates Kantian categorical imperative and deontic logic. Although a rule-based ethical theory is a better candidate for machine moral reasoning, as rules or duties are computationally more tractable, critiques that similar to the utilitarian model may also apply. First, deontology also does not capture the whole of our moral intuition. Categorical imperatives, an exemplary way deontological ethics is based, generates absolute prohibitions against certain actions, such as killing, lying or promise breaking. However, it is widely accepted that in some special situations in which killing, lying or promise breaking are permissible. As Bringsjord (2009) points out that programming ethics in a deontological way may produce catastrophic outcomes when the machine encounters scenarios where deontological rules should be broken. For example, a deontological AI may barred from violating categorical imperatives, even if the violation may result in saving great numbers of lives.[1] For example, deontological AI may prohibit harming an innocent person, even if doing so may save the whole population of a country. Second, there is one serious problem with regard to transferring deontological principles to computer language by logic. Both monotonic and nonmonotonic approach to the deontological model involves serious problems. (Power, 2011) Third, as Asimov's famous "Three Laws" show, a strictly deontological system without some sensible, intuitive ordering of principles will likely fail when unexpected consequences of principles arise under certain situations and/or when principles lead to contradictory prescriptions (Asimov, 2004). The difficulty of weeding out failure scenarios are likely to expand drastically as the number of principles expand and when A.I. is applied more and more generally.

To sum up, the problems of top-down approaches are twofold. The first is the persistence of *moral disagreements* in moral and social life. Those moral theories are often equally defensible but mutually exclusive perspectives to moral problems. Deontological, utilitarian, virtue ethics can all

[1] https://vimeo.com/4032291

provide plausible accounts of morality, and no one seems to have been able to show convincingly that one account defeats all others.[2] Even though moral disagreements need not commit us to radical ethical scepticism, it does challenge any claims, who profess that they have arrived that *right* moral theory and found the *right* algorithm for machines.

The second is that top-down approaches are based on the assumption that morality is codifiable. However, this assumption is controversial. The non-codifiability of ethics can be further subdivided: 1) in-principle non-codifiable; and 2) in-practice non-codifiable. In moral philosophy, one argument for the in-principle non-codifiability thesis is based on moral particularism. One relevant current of moral particularism for our purposes begins with the view that morality in no way depends on the existence of moral principles. Jonathan Dancy (2004) defines particularism as follows: the possibility of moral thought and judgment does not depend on the provision of a suitable supply of moral principles. He argues that there is no way a general principle can fully capture the complexity of how various morally relevant factors interact with each other. If Dancy's particularist view is true, then it seems a top-down approach would be insufficient to develop moral AIs possessing adequate moral sensibilities.

The in-practice non-codifiability thesis, in contrast, agrees with moral generalism, but takes issue with the fact that principles can be adequately transformed into operatable algorithms. Ethical theory that centers on principles ought to provide a universal practical algorithm for machines. Consider the classical utilitarianism of which the only moral principle is the maximization utility principle. For utilitarianism, the decision procedure is as follows: 1) list all available options; 2) calculate the expected utility of each available option; 3) rank each available option; 4) choose the option with the maximal expected utility. This may seem simple at first, but a further glance shows otherwise. In order to facilitate the maximizing expected utility, the programmer needs to further program the machine to obtain complete information of the probability of each option, a comprehensive understanding of society that allows the machine to foresee all expected consequences for the options, and a complete cardinal and interpersonally comparable knowledge of utilities. It creates huge information costs and also computational burdens on the computers.

Taking a different approach to machine ethics, the bottom-up approaches aim at training machine to build ethical framework without appealing to explicit moral rules. The bottom-up approaches are based Machine Learning (ML), which are increasingly capable of solving complex problems, effectively avoiding some of the problems of top-down approaches. In contrast to the traditional approaches where software developers write explicit codes for a computer to solve specific tasks, ML developers do not select parameter values for the task themselves. Instead, they specify basic architectural principles such as whether the system takes the form of a deep neural network approach or a decision tree. Furthermore, the developers specify an appropriate learning algorithm and a suitable

[2] Indeed, if we subscribe to Max Weber's (1978) distinction between the "purposive-rational" and the "value-rational" (the former relates to seeking optimal, efficicacious *means* to well-defined problems; the latter relates to *deciding right ends* for ethical actions), the problem of disagreement may be intractable. This is because while purposive-rationality has well-defined criteria of success (consider chess playing with clear victory conditions, or stock-profiling with maximizing profit as its standard), we seem to have no absolute guidance of other than established moral conventions, moral intuitions, and, at best, partially convincing arguments to deal with our choice of morally desirable ends. For an illuminating reading of Weber on the issue of moral rationality and disagreements, see Habermas (1984).

learning environment, which helps determine values to the learnable parameters. Guarini (2011), for example, proposes an experiment that he has trained an artificial neural network on specific moral judgments.

However, the bottom-up approaches are also subjected to criticisms. First, as Wallach and Allen point out (2010), since bottom-up approaches are heavily dependent on training data, they can be pose more moral risks than top-down or hybrid approaches as they lack assurances that certain important moral principles will be followed (especially in cases where some principles are clearly applicable). This is related to the general limitations of the *inscrutability* of the ML approaches. While machine learning can quickly apply knowledge and training from large datasets to excel in tasks too daunting for the symbolic approach, allowing a system to learn to recognize complex patterns and make predictions. The relative lack of intervention is one of the principal advantages of ML over the traditional approaches to AI, as it allows the machine finding unintuitive but efficient solutions. However, this also makes the ML systems characteristically opaque, that is, their operations are notoriously difficult to understanding and thus human programmers are quite unable to intervene so as to change and predict the system behaviors. Examples of how neural networks "overfit" their data and fail in "non-human" like ways show that how the intractability of bespeaks the potentially unreliability and untrustworthiness of machine learning approaches (Mitchel, 2019).

Second, if moral framework is developed purely out of training data acquired from ordinary persons, then trained machines may inherit or even reinforce undesirable prejudices and biases of ordinary persons. Third, the bottom-up approach may still be subjected to the computational limitation problem. Notably, the bottom-up approaches are based on a presumption that learning based on existing human judgments will develop a liable framework for making correct future moral decisions. However, there might be an infinite number of morally relevant features of each specific situation, and there seems no guarantee that a given machine learning system, even if it is well trained, can produce right decisions in all future cases, because morally relevant features not included in the current training dataset may make moral difference in a future case. It seems impossible to exhaust all possible morally relevant features. To sum up, the main problems of bottom-up approaches are 1) the problem of *inscrutability*; 2) the problem of training data; and 3) the problem of computational limitation.

3. Hybrid Model

Knowing the respective limitations of the top-down and bottom-up approaches to machine ethics, a hybrid model may also be developed for moral AIs: First, corresponding to the level of conscious processing, programable AIs (e.g. of the utilitarian kind outlined above) may be developed to keep the inscrutability of the machine-learning component in check. Second, corresponding to unconscious moral sensibilities, deep learning may be used to train AI systems (with training data from ordinary people's behavior – especially, as we shall argue, *skillful* behavior) to develop moral sensibilities reasonably close to ordinary human beings, and meanwhile reduce the information costs. These two systems are to be further integrated to generate a comprehensive system that can mimic a typical human moral agent acting in a specific domain.

The hybrid approach overcomes some of the problems of the first one by distributing the burden of moral de-risking and moral processing onto two systems. For example, one way of designing the hybrid system involves allowing 'moral intuitions' simulated the machine-learning system to produce a shortlisted set of maneuvers for symbolic processing, thus reducing the danger of faulty programming and reduce the burden for programmers to develop highly detailed, loop-hole free programs. Besides, machine-learning programs may also reduce the costs of explicit information processing by not requiring the machine to carry out time-intensive and exhaustive calculations in every situation.

Yet, the hybrid approach is not a panacea. For example, Anderson and Anderson (2011) have developed a hybrid system that incorporates ideas from W.D. Ross's *prima facie* duty approach to ethics. The hybrid system combines the explicit coding in which different *prima facie* duties are transferred to explicit algorithms with a machine learning system to prioritizing *prima facie* duties in a particular domain to avoid conflicts. The fundamental disagreement in ethics still exists in the hybrid system. There is no good reason to believe that the *prima facie* duty theory defeats other alternative normative theories. Also, if the hierarchy of duties is developed out of a ML system, such trained machines may still inherit two problems of ML. First, the parameters that determine the hierarchy of moral duties are unknown. Second, there is a high risk that the hierarchy involves undesirable prejudices and biases from ordinary persons.

A better hybrid system is a system, which overcomes or, at least, does better jobs than alternatives in problems mentioned above. However, due to the limitation of the space of the paper, it is not feasible to examine them all. This paper will focus on two main problems: 1) the disagreement problem in moral theories; and 2) the *inscrutability* problem (i.e. black box problem)

4. A Pluralist Hybrid System and Moral Disagreement

We suggest a pluralist hybrid system that combines an explicit rule-based algorithm system that accommodates different moral decision procedure, which is responsible for explicit moral decision-making, with a machine learning system, which is responsible for analysing and computing relevant information necessary for the moral decision-making. In the following section 4.1-4.5, we present the main ideas underlying the pluralist hybrid system, which are inspired by works from Philosophy, Economics and Psychology.

4.1. Moral Theories as Decision Procedures

An important distinction in moral philosophy is the distinction between moral theories as the criterion of rightness and moral theories as moral decision procedure. The criterion of rightness tells us what for an act to be a morally right action. The moral decision procedure tells us what we should make moral decisions. Some act utilitarians argue that act utilitarian is correct as a criterion of what is right or wrong but reject it as a moral decision procedure. However, the idea of "esoteric morality" is famously defended by Sidgwick and Lazari-Radek and Singer (2010) but rejected by the vast majority of philosophers, including Kant, Rawls, Bernard Gert, Brad Hooker, T.M. Scanlon and Derek Parfit. The on-going debate about the acceptability of esoteric morality is beyond the scope of this paper.

Given the distinction articulated above, the problem of moral disagreement can be formulated in terms of the disagreement of which moral decision-making procedure an agent should follow. In this section, we first review a few moral decision rules proposed by economists, philosophers and psychologists, leading to different moral decision models. We show that different moral rules use different types of information and ignore the others. They disagree what type of information is relevant to moral decision making.

Maximizing and Satisfying Utilitarianism

Cloose's Utilibot model is designed based on the idea of maximization of human well-being, which is rooted in the utilitarian tradition. However, as we mentioned above, the traditional utilitarianism is often challenged as being implausible as moral decision-making procedure due to its commitment to the idea of utility-maximization. Michael Slot (1984) proposes "satisficing" utilitarianism as an alternative to the traditional utilitarian ethics. While different versions of satisficing utilitarianism exists, they mostly hold that an act is morally right if and only if its consequences are "good enough." [3] Slot's Satisficing theory is based on the economic notion of satisficing suggested by Herbert Simon.

Herbert Simon (1955) challenged that the goal of utility maximization as formulated by rational choice theory and argues that it is nearly impossible to achieve in real life. He contends that in typical choice situations the application of maximin rule, probabilistic rule, and certainty rule requires individuals to be fully informed about (1) the exact nature of each possible outcome; (2) the order of each possible outcome, and (3) the definite probabilities attached to each possible outcome; and (4) to be armed with the perfect calculative skill to calculate the expected utility of each outcome, but no human being can possess all these calculative and cognitive skills.

Simon (1955), therefore, proposes that decision-makers should be considered as bounded rationally and suggested an alternative decision model in which utility maximization is replaced with satisficing. The satisficing model replaces the utility maximization objective from the expected utility theory of selecting the outcome with the *highest* expected utility with one that selects any option that meets your aspiration. The satisficing model is more straightforward than the maximizing one in two dimensions. First, it is more straightforward in virtue of simplifying the utility function. While the maximization model requires individuals to assign a utility value to *each* possible outcome of *all available* alternative actions, the satisficing model only requires individuals to set up a threshold that distinguishes between "satisfactory" and "unsatisfactory" outcomes. Second, it is simpler by eliminating probabilities from rational choice. While individuals under the maximizing model have to weigh the utility of each possible outcome by the probability that it will occur, the model of satisficing rationality only requires an agent to choose a course of action where all of whose expected outcomes are satisfactory.[4]

---

[3] Different versions of Satisficing Consequentialism diverge in terms of how to interpret the concept of "good enough." See Bradley, 2006.

[4] Sunstein (2005) and Gigerenzer (2010) adopts Simon's bounded rationality perspective to explore which rules or heuristics boundedly rational individuals use, when they confronted with moral choice situations. Both scholars come up with ha number of heuristics, such as Do not knowingly cause a death;

Utilitarianism vs Social Contract

In the literature of the concept of justice, one of the most well-known debates is between John Rawls's two principles of justice and John Harsanyi's principle of average utility. For the purpose of this paper, we shall not provide full detail of the debate between these two eminent scholars. It is sufficient to note although the issue of their debate concerns conflicting conceptions of the just society and the validity of interpersonal comparison, crucial points in their reasoning concern the disagreement between two different decision-making rules under uncertainty: 1) maximin rule; and 2) Laplace rule.

According to the maximin rule, the act is preferred to another act if and only if its worst possible outcome is preferred to the worst people outcome of the other act. And two acts are indifferent if and only if their worst possible outcomes are indifferent. The maximin rule has been criticized. John Harsanyi (1975) challenged the rule as an unreasonable decision-rule for decision-making under uncertainty. The main challenge is that under the maximin rule, our choices completely depend on the unfavourable contingencies regardless of how unlikely they could be. And it is extremely irrational to do so. In almost all daily activities, there may involve certain chance—very tiny but positive—of unfavourable outcomes. If we stick to maximin rule, we should *never* go outside, take public transportation or being engaged in activities.

Harsanyi, therefore, applies the Laplace principle, which is also known as the principle of insufficient reason, as an alternative to the maximin rule. The basic idea behind this rule is that if for two states of nature we have no reason to regard one of them as more probable than the other, we should regard them as equally probable. Using the Laplace principle, we end up with the utilitarian approach, according to which our choice should depend on the sum of the total expected utility.

4.3 Morally Relevant Information

One observation is that moral rules use some classes of information but ignores others and different moral rules disagree what information is morally relevant and what is irrelevant. (Sen, 1979)

For example, Consequentialism and utilitarianism disagree about what type of information should be considered to evaluate how good or bad the state of affairs. All forms of utilitarianism are welfarism, which only considers welfare of people involved in ranking social states. Non-welfare information such as values of liberty, or notions of egality is excluded. This is different from other forms of consequentialism, which may include non-welfare information in ranking social states. Maximizing and Satisficing Consequentialism disagree if probability information and complete utility information of the outcome should be considered to determined the rightness of action.

Rawls rejects the common utilitarian assumption that well-being can be reduced preference satisfaction and questions the pooling up of utility in issues of interpersonal comparison. This is why he designs a list of primary goods to facilitate the interpersonal comparison. The maximin rule that Rawls applies in the original position. In contrast, Harsanyi's utilitarianism uses von Neumann–Morgenstern

---

doing is morally worse than allowing, see ((Sunstein, 2005); Choose the default option; Imitate your peer; and tit-for-tat, see (Gigerenzer, 2010).

utility functions, or vNM utility functions, to represent the utility that an agent assigns to each possible outcome. And further, Harsanyi uses the vNM expected utility theory to impose the cardinality on our preference. In other words, Harsanyi's utilitarianism approach requires cardinal welfare information for interpersonal comparison.

Consider other moral theories. Deontologists deny that what we ought to do is determined by how good or bad the states of affairs are. They believe that sometimes it is permissible or even morally required to bring about suboptimal states of affairs. Contractualist principle may include information of well-being, and information of other reasons beyond well-being, but exclude the aggregation information. Nozick's theory of rights may include a class of "entitlement" information dealing with ownership.

4.4 Moral Disagreement Problem, Domain Specific and Pluralist Approach

One response to the moral disagreement problem is to redirect our focus to domain-specific instances. While disagreements pervade in moral theories of deontology, utilitarianism, and virtue ethics, we may agree that some ethical considerations are particularly important in specific contexts: for instance, virtue and empathetic sensitivity seems to be more pertinent than other considerations in educational contexts, while (deontic) ethics of self-defence and retaliation are important problems in military ethics. Pragmatic considerations of passenger comfort and accident avoidance should be the focus of driving, while issues of maximizing survival chances and reducing suffering (at least at the level of each individual) should be the baseline in the medical context.

However, pointing out domain-specific features of our moral intuitions in different contexts does not remove all moral disagreements, and especially in unpredictable cases, disagreements and moral dilemmas may still exist (consider the famous trolley cases, organ-donation cases, or caregivers when the building is on fire).

Second, we need to further develop a promising theory to justify why certain normative ethical theories should be adopted for specific tasks. So far, we still lack a theory as such. Moreover, even if we can develop a promising theory to explain why, say, deontology is the right theory for military ethics, in some particular contexts when what is at stake is millions of lives, Consequentialist approach might be more plausible than deontology. This problem pertains to the codifiability debate we discussed above: some philosophers believe that there is no universal moral principle can guide all our moral judgments and actions.

In responding to these two problems, we suggest an alternative strategy: shift our focus from a monist approach to a pluralist approach. The basic idea is that if it is implausible to identify one particular ethical theory as the right moral action guidance for moral AIs, why shall we take a pluralist approach—include all the ethical theories in one system. All those moral rules (e.g. Maximizing Consequentialism, Satisficing Consequentialism, Contractualism and etc.) are equally justified as the right action guidance: no one is superior to the others. In a particular moral context, which moral rule will be activated for the moral decision-makings guidance depends on the principle of theory selection. Let us call it the **principle of selection**:

Principle of selection: In a particular moral context μ, moral rule ∂ is activated for machine to make moral judgments and decisions if and only if the morally relevant information of ∂ is all available or easy to access and calculate.

The **principle of selection** is subjected to challenges. But these challenges can be defeated. First, it might be argued that there is a high risk that the machine may activate a moral principle based on availability and accuracy of moral information, yet causes a moral disaster. For example, there is a chance that in a war context, given the available information, the maximizing utilitarian moral rule is activated by an autonomous weapon that a small group of civilians' lives is sacrificed in order to save a large group of soldiers' lives. It is morally wrong. We admit that the pluralist approach cannot 100% avoid moral risk as such. However, given the problem of moral disagreement and moral codifiability, it is hard to conceive any design that could do better than the pluralist approach. The moral risk remains equally high in a monist system, no matter what ethical theory it is adopted. Moreover, in war situations, very often the calculations of the morally relevant parameters of maximizing utilitarian principle often are highly complex and therefore less likely to be accurate and precise in practice. Therefore, we speculate that it is not likely that a maximizing utilitarian principle will be activated.

Second, it might be argued that the **principle of selection** may yield indeterminacy. Suppose that in a moral context μ, the ordinal information of each outcome is available and easy to obtain, but the cardinal outcome information and the probability information are not and both are incomplete. Moreover, the information of certain deontological constraints is also easy to obtain. Given the availability of information, both the maximin principle and the deontological principle would be activated. However, indeterminacy does not necessarily yield conflicting results. Consider the footbridge problem. The deontological principle would yield the choice of allowing five people die without doing anything. The doctrine of doing and allowing imposes deontological constraint on harming: the outcome resulted from harming is ranked lower than the outcome resulted from allowing harm. The maximin principle also would endorse the choice of allowing five people die. Because the worse situation of doing nothing is that five people die, but the person on the footbridge is alive; whereas the worse situation of pushing the fat man over the footbridge is that all six people die if the fat man fails to stop the trolley.

We admit that there are situations in which more than one moral rule would be activated, and yield conflicting moral choices. For example, in the famous trolley dilemmas, maximizing utilitarian rule and deontological rule often yield conflicting results. Yet, as we mentioned above, maximizing utilitarian rule often requires probability information and cardinal utility information to facilitate the expected utility calculation. Those calculations are highly complex in a real-life situation. For example, in a real-life trolley problem situation, the calculation of the likelihood and utility of saving five people by switching the trolley is far more complex than the hypothetical scenario. Factors, such as weather and surrounds, can all affect the value of each parameter. Thus, in practice, the information required by the expected utility calculation might not be easily available. And thus, the likelihood that a maximizing utilitarian rule is activated in given situation is much lower than the likelihood of other moral rules.

To sum up, we admit that the pluralist system does not eliminate all moral uncertainty caused by the moral disagreement problem. The remaining moral uncertainty is twofold. First, there is a risk

that the activated moral rule causes moral disaster. Second, there is a risk that more than one moral rule is activated and thereby it yields conflicting orders to the machine. The responses to the challenges are also twofold. First, any monist approach may inevitably involve the risk of causing moral disasters (see section 2), and it might be an empirical question whether a pluralist approach does worse than a monist approach. Our speculation is that given the availability of morally relevant information in most real-life cases, the most frequently activated moral rules should neither be maximizing utilitarian rule nor Kantian deontology rule, instead should be something between (e.g. a contractualist rule or a satisficing utilitarian rule or a threshold-deontological rule, and etc. depends on the availability of information). Second, if put all the ethical theories into a spectrum, maximizing utilitarian and Kantian absolutist deontology might be most distinct, whereas the rest are less so. For example, threshold-deontological rule allows limited aggregation and thus it may behave very similarly as a satisficing consequentialist rule in most of the cases. Thus, even though more than one rules are activated, it does not always yield conflicting results.

4.5 Machine Learning System and Morally Relevant Information

The ML system is designed to address the information costs problem. The possible choices machines make in real-life scenarios do not have clearly discernable odds and payoffs printed on their forehead. Discerning payoffs and odds, even for the less demanding decision rules, require the machines to extrapolate predictions of possible futures that may result from individual decisions. The values of individual parameters are often numerous and nonlinearly coupled means so that the software developers may not be able to track the way individual inputs are transformed into specific outputs. The ML approach is designed to analyze the relevant information and then calculates the values of moral parameters. Software developers specify the conditions under which machine learn how to effectively calculate the values of individual variables. Such design aims to take the advantages of the merits of ML. ML lacks influence on the way in which problems are solved to identify the solutions of value calculation, which might be highly unintuitive but effective.

To sum up, to address the moral disagreement problem, the pluralist hybrid system combines the explicit coding in which different *moral principles* are transferred to explicit algorithms with a machine learning system to efficiently and accurately calculate the value of parameters specified for each moral principle. The idea of the pluralist hybrid system can be presented by Figure 1:

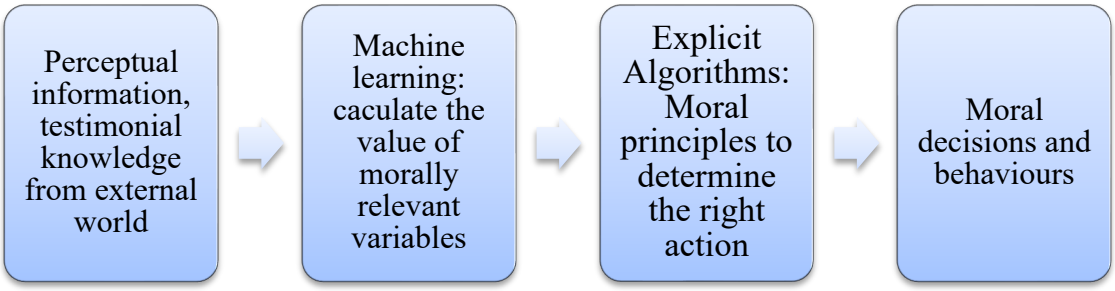

Figure 1: Pluralist Hybrid System

5. Pluralist Hybrid System and Inscrutability Problem

As discussed in section 2, the inscrutability problem (i.e. black-box problem) arises from the fact that machine learning systems used by many computing systems are *opaque*. The root cause of the opacity in machine learning is that the software developers relatively lack influence on the way in which an AI problem is solved, compared to the traditional algorithms-based system. And the root cause of the lack of influence is that the machine learning developers cannot specify how an AI task is tackled but instead merely specify the conditions which solutions to the task may be found.

Not all system opacities are worrisome and some are only worrisome to particular shareholders. For example, in the go match between AlphaGo and Lee Sedol, AlphaGo won and showed astonishing proficiency in making accurate and brilliant moves. However, there were a few moves from AlphaGo are difficult to interpret even by professional commenters. In a go game, coming up with strategies that is unexpected or unintuitive is considered an advantage, as it makes the player's moves unpredictable to its opponents. However, it might be a problem for software developers, as they do not know if they need to update their system when certain moves are hard to be interpreted: it is unknown if they are system errors or genius moves.

In the same vein, imagine that in the near future, AI robots are closely engaged in our daily life, whose behaviour affects our well-being in different aspects of human life. Those AIs, like AlphaGo, are initially trained to mimic human moral behaviours and moral decisions from "moral databases", but what moral behaviours and decisions they comes up with are out of our hands. It is conceivable that those AI robots may come up with certain actions or decisions that are opaque to us such that we cannot understand even with our best efforts. Worse still, AI robots may perform in such a way that we consider as immoral. We do not know whether these actions are caused by a system error, or they result from certain novel principles that the machine (mis-)learned from the training data. This inscrutability is especially

worrisome to the members of the society, as it makes them harder to form reasonable expectations on those machines' behaviour and plan their life accordingly.

The pluralist hybrid system removes the opacity in the procedures of moral decision-makings by adopting the traditional approaches to AI. The explicit algorithms specify the individual parameters that determine the moral rightness of choice or action. Yet, the system still retains one kind of opacity in the calculation of the value of each parameter, thanks to ML. Yet, this opacity can be justified in a cost-benefit analysis: a machine-learning approach can significantly improve the effectiveness in identifying the value of individual parameters. It outweighs the risk brought about by the remaining opacity in the ML system. Moreover, the risk in question can be supervised. A supervised learning algorithm can be used to teach machine to yield desired results and make the system more predictable. And even though certain level of opacity is inevitable, at least the normative elements in a moral decision-making are explainable by the explicit algorithms. We can trace and identify which moral principle is activated and the numerical value of each parameter. Although we (different shareholders) may still be blind from how those numerical values are calculated, but at least an *ex post* explanation would be available.

6. Conclusion

Despite the above proposals, they should not be taken as imperatives for designing moral A.I.s, but only as tentative methodological recommendations for potential projects that works on aligning A.I.s with ethically desirable outcomes. Other than this, we should remember that much of what is proposed depends on the technical work of practitioners of the field in approximating the ideal functioning of the systems outlined above.

It should be born in mind that the pluralist hybrid approach proposed above does not decisively resolve the surveyed limitations of the top-down and bottom-up approaches to machine ethics. Even if the pluralist hybrid approach is adopted, its robustness and reliability in producing consistently ethical outcomes have to be tested and such results should be carefully scrutinized by engineers, experts, and relevant stakeholders.[5] Recent attempts in "explainable A.I.s" or "interpretable machine learning" may aid this process in ethical risk minimizing. Besides, even by outlining the information requirements of different decision-making procedures and seeing their potential applications in specific domains, disagreements about which decision-making procedure one should adopt to produce ethical outcomes may still arise, especially in our morally pluralistic societies. These disagreements may only be potentially ironed out through social engagement of people of different ethical positions when such algorithms are designed.

Reference

Anderson, M., & Anderson, S. L. (2011). *Machine Ethics*. Cambridge University Press.

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 355–372.

Bringsjord, S. (2009). *Unethical but Rule-Bound Robots Would Kill Us All*. Retrieved December 10, 2021, from http:// kryten.mm.rpi.edu/PRES/AGI09/SB_agi09_ethicalrobots.pdf.

Cloos, C. (2005). *The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism*. Palo Alto: Association for the Advancement of Artificial Intelligence.

Dancy, J. (2004). *Ethics Without Principles*. Oxford: Oxford University Press.

Gigerenzer, G. (2010). Moral Satisficing: Rethinking Moral Behavior as Bounded Rationality. *Topics in Cognitive Science*, *2*(3), 528–554.

Guarini, M. (2011). Computational neural modeling and the philosophy of ethics: Reflections on the particularism-generalism debate. In Anderson, M., and Anderson, S. L. (Eds.), *Machine Ethics*. New York: Cambridge University Press.

Harsanyi, J. C. (1975). Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory. *The American Political Science Review*, *69*(2), 594–606. https://doi.org/10.2307/1959090

Habermas, J. (1984). *Reason and the Rationalization of Society* (Vol. 1). Cambridge: Polity Press.

Mitchel, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. London: Penguin.

Lazari‐Radek, K. D., & Singer, P. (2010). Secrecy in Consequentialism: A Defence of Esoteric Morality. *Ratio*, *23*(1), 34–58.

Powers, T. (2011). Prospects for a Kantian machine. In Anderson, M., and Anderson, S. L. (Eds.), *Machine Ethics*. New York: Cambridge University Press.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. JSTOR. https://doi.org/10.2307/1884852

Slote, M., & Pettit, P. (1984). Satisficing Consequentialism. *Aristotelian Society Supplementary Volume*, *58*(1), 139–176. https://doi.org/10.1093/aristoteliansupp/58.1.139

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*(4), 531–542. https://doi.org/10.1017/S0140525X05000099

Sen, A. (1975). "Informational Analysis of Moral Principles," in Ross Harrison, ed., *Rational Action.* Cambridge: Cambridge University Press.

Wallach, W., & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. In *Moral Machines*. Oxford: Oxford University Press.