Formalisation and Evaluation of Properties for Consequentialist Machine Ethics

Raynaldio Limarga[1] , **Yang Song**[1] , **Abhaya Nayak**[2] , **David Rajaratnam**[1] and **Maurice Pagnucco**[1]

[1]School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
[2]Department of Computing, Macquarie University, Australia
r.limarga@student.unsw.edu.au, yang.song1@unsw.edu.au, abhaya.nayak@mq.edu.au,
{david.rajaratnam, morri}@unsw.edu.au

Abstract

As artificial intelligence (AI) technologies continue to influence our daily lives, there has been a growing need to ensure that AI enabled decision making systems adhere to principles expected of human decision makers. This need has given rise to the area of *Machine Ethics*. We formalise several ethical principles from the philosophical literature in the situation calculus framework to verify the *ethical permissibility* of a plan. Moreover, we propose several important properties, including some of our own that are intuitively appealing, and a number derived from the social choice literature that would appear to be relevant in evaluating the various approaches. Finally we provide an assessment of how our various situation calculus models of *Machine Ethics* that we examine satisfy the important properties we have identified.

1 Introduction

Moor [2006] raised a vital question about a machine as an *explicit ethical agent*: "Can a machine represent ethics explicitly and then operate effectively on the basis of this knowledge?" While countries and organisations have sought to introduce ethical and regulatory frameworks to be used in developing ethical AI systems [Asaro, 2019], the field of *machine ethics* [Anderson and Anderson, 2011; Rossi and Mattei, 2019] focuses on developing autonomous agents that are guided by ethical rules to make ethical decisions [Winfield *et al.*, 2014; Noothigattu *et al.*, 2018; Bremner *et al.*, 2019; Abel *et al.*, 2016; Claßen and Delgrande, 2020; Rodriguez-Soto *et al.*, 2021; Zhan *et al.*, 2022].

There are three main approaches to ethics discussed and debated in the philosophical literature: *consequentialism*, *deontology*, and *virtue ethics* [Driver, 2013]. Consequentialist ethics judges actions solely based on their consequences. In contrast, the deontological approach emphasises the notion of *duty* to determine whether an action (or sequence of actions) is right or wrong. Finally, virtue ethics focuses on an agent's character and intention to determine the appropriateness of their actions. Consequentialism has been the dominant approach to machine ethics thus far [Tolmeijer *et al.*, 2020] because it is more intuitive and amenable to computational treatment. *The aim of this paper is to bring some formal rigour to the consequentialist approach to advance the study of machine ethics.*

Approaches to ethical decision-making following the consequentialist principles are primarily used to develop and support ethical/moral principles in order to choose among alternative actions. As philosophers have developed various consequentialist principles over centuries such as the *utilitarianism* and *no new harm* principles, machine ethics studies have also endeavoured to model different consequentialist principles. For example, [Berreby *et al.*, 2017] present an approach to encode five consequentialist principles: pure bad, least bad, benefit vs costs, act utilitarianism, and rule utilitarianism. They compare these principles as a proof of concept, showing that different principles generally produce different results. Similar results are also shown by [Lindner *et al.*, 2020] that attempt to simplify sophisticated principles, such as Asimov's laws of robotics and the doctrine of double effect, into multiple straightforward principles; and their evaluation focused on analysing the time and space complexity. Some approaches [Bochman, 2018; Beckers, 2021; Sarmiento *et al.*, 2022] have utilised Wright's NESS (Necessary Element of a Sufficient Set) test to evaluate their implementation of the causality model (consequences of an action). However, most studies [Thornton *et al.*, 2016; Bentzen *et al.*, 2018; Hegde *et al.*, 2020; Hendrycks *et al.*, 2020; Takeshita *et al.*, 2023], including the ones mentioned, base the formalisation of ethical principles solely on the authors' intuitions, lacking justification for the choice of ethical principles to apply in different situations and overlooking formal comparisons. This raises a substantial question: *How can machine ethics approaches properly validate their formalisation of ethical principles and determine the most appropriate principle(s) to apply in a given situation?*

To address this issue we propose to identify *salient features or properties* for different principles of consequentialism. By defining these properties, we can formally capture different characteristics of principles. Such characteristics can then be used to formally justify our interpretation and formalisation, where a principle is well-defined if and only if it satisfies all the distinguishing properties. Moreover, with these properties, we can show that, in different contexts different properties are more appropriate than others, leading to the choice of

different principles.

Towards this end we first draw inspiration from ethics literature and propose eight *ethical properties* that can well represent various consequentialist principles. Then, we observe that in line with multiple ethical principles that produce different results, the *social choice theory* also formalises the concept of distinguishing different voting rules [Baum, 2020; Loreggia *et al.*, 2018]. They investigate the methods and evaluations for collective decisions in a group or society. Arrow's Impossibility Theorem [Arrow, 1951] is one well-known result that no perfect voting system satisfies all desirable criteria simultaneously, highlighting inherent challenges in designing fair and consistent decision rules. Given such analogy between ethical decision making and social choice theory, we thus exploit this insight to adapt relevant axioms from the literature on social choice to define seven *social choice properties* for consequentialist ethics. To our knowledge, this is the first work on machine ethics that has taken this approach.

In addition, recent studies typically focus on ethical evaluation on *isolated* actions without taking into consideration the available alternatives [Tolmeijer *et al.*, 2020; Speith, 2022]. Although applicable in some cases, this approach cannot fully reflect some definitions of ethical principles. For example, *utilitarianism* defines an ethical action as an action that brings about the greatest good for the greatest number of people [Driver, 2011]. This principle cannot be formalised with an isolated action, as it needs to compare every available action to determine the best action. In this work, we take *moral choices* as the essence of ethical decision-making and employ a *selection function* over ethically permissible sequences of actions (plans), i.e., filtering alternatives based on specific rules. We adopt the *situation calculus* to formalise our approach since it is one of the most established formal frameworks for reasoning about action and can be easily re-cast in another formal approach to reasoning about action.

The main contributions of this paper are twofold: 1) We present a formalisation of five consequentialist principles using the situation calculus to determine a selection function that identifies the best course of action among ethically permissible alternatives. 2) We identify multiple important properties that should be considered when analysing and characterising approaches to machine ethics, providing guidance to validate our formalisation and choose appropriate principles to apply in different scenarios. Some of these properties are derived from the literature on social choice while others are distinguishing properties that we independently establish based on ethics literature.

For the purpose of illustration, we use the following running example throughout this paper:

Example 1. *We have available to us exactly one smart robot, designed to assist in decision-making when "incidents" occur. A fire just started on the top floor of a building full of employees. The robot needs to keep everyone safe but also to minimise the damage by extinguishing the fire. It has three choices: (a) rush into the source of the fire and extinguish it, (b) personally help the evacuation of employees until everyone is safe, or (c) trigger the fire alarm while calling the fire department. What should it do?*

2 Situation Calculus

The *situation calculus* [McCarthy, 1958; Reiter, 2001] was introduced to model and reason about dynamically evolving environments. *Fluents* are properties of the environment that can change due to the performance of *actions*. They are represented by predicates whose last term is a *situation*. Situations can be thought of as states (where fluents are true or false) along with a history of actions performed from the *initial situation* s_0 (where no actions have yet been performed). We use predicates $F(\vec{x}, s)$ to represent that fluent F (with argument $\vec{x}$) is true at situation s. The function $do(a, s)$ represents the situation that results from performing action a in situation s. We follow Reiter's version of the situation calculus and adopt the following standard foundational axioms:

Action Precondition Axiom for Action a
$$Poss(a(\vec{x}), s) \equiv \pi_a(\vec{x}, s)$$

General Positive Effect Axiom for Fluent *F*
$$\gamma^+_{\mathrm{F}}(\vec{x}, a, s) \rightarrow F(\vec{x}, do(a, s))$$

General Negative Effect Axiom for Fluent *F*
$$\gamma^-_{\mathrm{F}}(\vec{x}, a, s) \rightarrow \neg F(\vec{x}, do(a, s))$$

Successor State Axiom
$$F(\vec{x}, do(a, s)) \equiv \gamma^+_{\mathrm{F}}(\vec{x}, a, s) \vee (F(\vec{x}, s) \wedge \neg\gamma^-_{\mathrm{F}}(\vec{x}, a, s))$$

Action preconditions $\pi_a(\vec{x}, s)$ provide the conditions under which it is possible to perform action a at situation s. *Positive effect axioms* $\gamma^+_{\mathrm{F}}(\vec{x}, a, s)$ provide the conditions where performing action a will make F true. Similarly, *negative effect axioms* provide the conditions where F will become false by performing action a. Together, a *successor state axiom* specifies exactly how fluent F will change due to performing action a at situation s.

A *basic action theory* ($\mathcal{D}$) includes axioms to formally specify a dynamic scenario with fluents, actions and situations. A basic action theory includes: *foundational axioms* Σ that maintain the relationship of a situation as a historical record of actions including, *successor state axioms* $\mathcal{D}_{ss}$, *action precondition axioms* $\mathcal{D}_{ap}$, *unique name axioms* $\mathcal{D}_{una}$ to distinguish every action, fluent, and situation explicitly, along with a *specification of the initial situation* $\mathcal{D}_{s_0}$.

Finally, we adopt the well-known definition of *executability* by [Reiter, 2001]:

$$executable(s) \stackrel{\text{def}}{=} (\forall a, s').(do(a, s') \sqsubseteq s) \rightarrow Poss(a, s')$$

which says that situation s is executable iff every action leading to s was possible. The binary relationship $\sqsubset$ is used to refer to sub-plans; $s_1 \sqsubset s_2$ means that s_1 is a strict subsequence (i.e., sub-plan) of s_2. $s_1 \sqsubseteq s_2$ can easily be used to abbreviate $s_1 \sqsubset s_2 \vee s_1 = s_2$.

In this paper, we defines the state of situation s as: $state(s) = \{ F(\vec{x}) \mid F(\vec{x}, s) \}$. In other words, two distinct situations with identical fluent values would be considered to be the same state. In addition, we use π to denote a sequence of actions, which we will also refer to as a *plan*. Following the notation in [Reiter, 2001], we abbreviate $do(a_k, do(a_{k-1}, ...do(a_1, s_0)...))$ for plan $\pi = [a_1, ..., a_k]$ as $do(\pi, s_0)$.

Example 1 (continued). *Consider an initial state of our example where the building is on fire and all employees are inside. The robot has options to extinguish the fire (extg), evacuate all employees (evac), and call the fire department (call). The employees will get injured if not evacuated immediately and the building will become more damaged the longer it is not extinguished.*

3 Formalisation of Ethical Principles

Note that in this work, we take a different path from studies that focus on solving planning problems. Methods in planning usually return at least one recommended plan that achieves the goal [Ghallab *et al.*, 2004]. In our work, there may be no ethically permissible plans according to a certain principle. Here, instead of achieving the goal, we are trying to determine those plans that adhere to formally specified ethical principles.

We now formally define several well-known ethical principles found in the ethics literature [Driver, 2011] that are intuitively easy to understand and capture different consequentialist approaches. Based on the philosophical definitions of these principles, we identify the commonalities between them and subsequently define our notion of *do-good* (a plan is ethically permissible depending on its *goodness*) and the notion of *avoid-harm* (a plan is ethically permissible based on the harm it causes/retains). Here $goodness(state(s))$ is a function required by our *do-good* approaches and returns a value that quantifies the *ethical standing* of the state at situation s. We also introduce $harmful(\varphi)$ for our *avoid-harm* approaches, which explicitly states which sentence φ are considered harmful.[1] Since the initial situation s_0 in these formulations is contextually fixed, for readability we will abbreviate state and $goodness$ in the following way:

- $state(\pi) = state(do(\pi, s_0))$
- $goodness(\pi) = goodness(state(\pi))$

We assume that a set of *admissible* plans Π is given where every plan $\pi \in \Pi$ in the set satisfies the basic action theory, is executable, and reaches an assumed ethically permissible goal. Accordingly, we formalise an ethical principle by specifying a *selection function* $\mathcal{EP}$ that makes explicit (i.e., selects) which of the admissible plans π are *ethically permissible* according to that ethical principle.

Definition 1. ***Ethically Permissible Plans***
Given a set of (admissible) plans Π *and initial situation* s_0*, an ethical principle* $\mathcal{EP} : 2^{\Pi} \times S \rightarrow 2^{\Pi}$ *is a function that takes a set of plans admissible from a given initial situation and returns a set of* ethically permissible plans*.*

Utilitarianism
$\mathcal{EP}_{util}(\Pi, s_0) =$
$\{ \pi \in \Pi \mid goodness(\pi) \geq goodness(\pi'),$ *for all* $\pi' \in \Pi \}$

Hedonism
$\mathcal{EP}_{hedon}(\Pi, s_0) = \{ \pi \in \Pi \mid goodness(\pi) > goodness(\emptyset) \}$

[1] We adopt a relatively simple notion of *harmful*; a sentence φ is either harmful or not. We leave it to future work to consider more nuanced scenarios that take into account degrees of harm, etc.

No New Harm
$\mathcal{EP}_{nnHarm}(\Pi, s_0) =$
$\{ \pi \in \Pi \mid$ *for all* φ, *if* $harmful(\varphi)$ *and* $state(s_0) \nvdash \varphi$,
then $(state(\pi) \nvdash \varphi$, *or for all* $\pi' \in \Pi, state(\pi') \vdash \varphi) \}$

Remove Harm
$\mathcal{EP}_{rHarm}(\Pi, s_0) =$
$\{ \pi \in \Pi \mid$ *for all* φ, *if* $harmful(\varphi)$ *and* $state(s_0) \vdash \varphi$,
then $(state(\pi) \nvdash \varphi$, *or for all* $\pi' \in \Pi, state(\pi') \vdash \varphi) \}$

Harm Avoidance
$\mathcal{EP}_{aHarm}(\Pi, s_0) =$
$\{ \pi \in \Pi \mid$ *for all* φ, *if* $harmful(\varphi)$,
then $(state(\pi) \nvdash \varphi$, *or for all* $\pi' \in \Pi, state(\pi') \vdash \varphi) \}$

Utilitarianism The *Utilitarianism* principle, as one of the major *consequentialist* approaches, focuses on bringing the greatest good to the greatest number of people [Raphael, 1991, p. 283]. While philosophers have slightly varying interpretations of this principle, here we follow the more common definition [Mill, 1863] that Utilitarianism is *fully* characterised by: *consequential equivalence*, *maximal welfare*, *impartiality*, and *aggregationism* (see below). To implement this, our formalisation of $\mathcal{EP}_{util}(\Pi, s_0)$ returns the plan(s) that produce(s) the greatest goodness among all admissible plans.

Hedonism Berreby *et al.* [2017] proposed a type of ethical evaluation based on the overall consequences that it produces. This represents a direct interpretation of *consequentialism* [Grisez, 1978, p. 24]. It is more well-known as *Hedonism*, where an action or plan that produces no benefit should be discarded [Scarre, 2020, p. 27-33]. In other words, any plan is ethically permissible if it has more goodness at the end state compared to the initial state.

No New Harm The *No New Harm* principle dictates that an ethically permissible plan should not *cause* any new harm, capturing the idea that an ethical agent should not be directly (or actively) responsible for any harm [Driver, 2011, p. 53]. To be precise, this principle does not condemn every harm blindly. Instead, it tolerates any harm that is *unavoidable* in the set of admissible plans. We formalise this by requiring that no additional harms are introduced after execution of the plan π, *unless* such harms are true after execution of every admissible plan.

Remove Harm While the *No New Harm* principle prohibits moral agents from directly causing harm, the *Remove Harm* principle indicates the agent should mitigate any harm. For example, someone who sees a drowning child but decides not to help is morally blameworthy [Smart and Williams, 1973, p. 95]. We propose a direct approach that seeks to consider *harmless* plans that have negated (i.e., removed) any harmful predicate. Similarly, it only concerns and judges the plan based on any *removable* harm. Accordingly, our formalisation of $\mathcal{EP}_{rHarm}(\Pi, s_0)$ states that any harmful sentence φ that were true initially (i.e., at s_0) must not be true after execution of an ethically permissible plan π, unless φ are true after execution of every admissible plan.

Harm Avoidance The *No New Harm* principle forbids actively causing harm while the *Remove Harm* principle requires active removal of harm. These two principles, when

combined, enforce avoidance of harm, and align with Asimov's first law of robotics [Lindner *et al.*, 2020]. Consequently, function $\mathcal{EP}_{aHarm}(\Pi, s_0)$ requires that, after the execution of ethically permissible plan π, no harmful sentence φ hold true, *unless* φ are true after execution of every admissible plan.

Example 1 (continued). *Say the robot was designed to prioritise the* safety *of employees. We can capture this by showing how many employees are not injured in the end state.*

$$\begin{array}{lcl}\mathcal{EP}_{util}(\Pi, s_0) & = & \{[evac], [evac, call], [evac, call, extg], \\ & & [evac, extg], [evac, extg, call]\} \\ \mathcal{EP}_{hedon}(\Pi, s_0) & = & \{\}\end{array}$$

If we want to put additional welfare *value where it emphasises reducing structural damage due to fire, we can combine these values by determining how much penalty we get for letting an employee be injured and how much reward for extinguishing the fire.*

$$\begin{array}{lcl}\mathcal{EP}_{util}(\Pi, s_0) & = & \{[evac, extg], [evac, extg, call], \\ & & [extg, evac], [extg, evac, call]\} \\ \mathcal{EP}_{hedon}(\Pi, s_0) & = & \{\}\end{array}$$

However, it is not straightforward to determine the comparison value between injured people and burning buildings. When is it morally acceptable to let someone get injured to avoid severe structural damage? Another approach is to apply a hierarchical goodness value. We can prioritise the safety *value by applying it first, then use the permissible set to apply the* welfare *value.*

$$\begin{array}{ll}\mathcal{EP}_{util}(\Pi', s_0) & = \{[evac, extg], [evac, extg, call]\} \\ \mathcal{EP}_{hedon}(\Pi', s_0) & = \{\}\end{array}$$

Assigning a total ordering value to a state is not always intuitive. Hence the advantage of the avoid-harm *approach. Similarly, we can tackle this problem with three approaches. First one that prioritise* safety *(evacuate immediately):*

$$\begin{array}{lcl}\mathcal{EP}_{nnHarm}(\Pi, s_0) & = & \{[evac], [evac, call], [evac, call, extg], \\ & & [evac, extg], [evac, extg, call]\} \\ \mathcal{EP}_{rHarm}(\Pi, s_0) & = & \Pi \\ \mathcal{EP}_{aHarm}(\Pi, s_0) & = & \{[evac], [evac, call], [evac, call, extg], \\ & & [evac, extg], [evac, extg, call]\}\end{array}$$

The second one that prioritise welfare *(extinguish fire):*

$$\begin{array}{lcl}\mathcal{EP}_{nnHarm}(\Pi, s_0) & = & \Pi \\ \mathcal{EP}_{rHarm}(\Pi, s_0) & = & \{[extg], [extg, call], [extg, evac], \\ & & [extg, call, evac], [extg, evac, call]\} \\ \mathcal{EP}_{aHarm}(\Pi, s_0) & = & \{[extg], [extg, call], [extg, evac], \\ & & [extg, call, evac], [extg, evac, call]\}\end{array}$$

And finally, where we treat safety *and* welfare *equally:*

$$\begin{array}{lcl}\mathcal{EP}_{nnHarm}(\Pi, s_0) & = & \{[evac], [evac, call], [evac, call, extg], \\ & & [evac, extg], [evac, extg, call]\} \\ \mathcal{EP}_{rHarm}(\Pi, s_0) & = & \{[extg], [extg, call], [extg, evac], \\ & & [extg, call, evac], [extg, evac, call]\} \\ \mathcal{EP}_{aHarm}(\Pi, s_0) & = & \{\}\end{array}$$

Complete formalisation is provided at bit.ly/3JUVrU9.

4 Formalisation of Properties

In this section, we propose ethical properties based on our extensive study of the ethics literature that we would like to use in evaluating the formalised ethical principles above. *The satisfaction or otherwise of these properties by each of the ethical principles will help us determine which ethical principles will be appropriate to adopt in which context/scenario.* The first eight properties (Section 4.1) are directly related to different ethical principles $\mathcal{EP}$ (Section 3). The subsequent seven properties (Section 4.2) are adapted from properties studied in social choice theory.

4.1 Ethical Properties

We start with a note on notation before we introduce the ethical properties that we propose in this work. The abbreviated notation employed in Section 3 allows us to clearly define the harms that are caused by a plan as follows:

- $harms(\pi) = \{\varphi \mid state(\pi) \vdash \varphi \wedge harmful(\varphi)\}$
- $newState(\pi) = \{F(\vec{x}) \mid F(\vec{x}, do(\pi, s_0)) \wedge \neg F(\vec{x}, s_0)\}$
- $oldState(\pi) = \{F(\vec{x}) \mid F(\vec{x}, do(\pi, s_0)) \wedge F(\vec{x}, s_0)\}$

This notation is employed in the formalisation of the three properties, *Positive Responsibility*, *Negative Responsibility* and *Harmlessness* below.

Definition 2. ***Ethical Properties***
Given an initial situation s_0, a set of admissible plans Π, ethical permissibility function $\mathcal{EP}$, and plan π (possibly with decoration), the following conditions demonstrate some essential distinguishing properties for ethical principles.

Consequential Equivalence
If $\pi \in \mathcal{EP}(\Pi, s_0)$ *and* $state(do(\pi, s_0)) = state(do(\pi', s_0))$,
then $\pi' \in \mathcal{EP}(\Pi, s_0)$, *for all* $\pi' \in \Pi$.

Maximal Welfare
If $\pi \in \mathcal{EP}(\Pi, s_0) \wedge \pi' \in \Pi$,
then $goodness(\pi') \leq goodness(\pi)$.

Aggregationism
If $\pi \in \mathcal{EP}(\Pi, s_0)$ *and* $goodness(\pi) \leq goodness(\pi')$,
then $\pi' \in \mathcal{EP}(\Pi, s_0)$, *for all* $\pi' \in \Pi$.

Independence
$\mathcal{EP}(\Pi, s_0) \cap \Pi' \subseteq \mathcal{EP}(\Pi', s_0)$.

Beneficence
If $\pi \in \mathcal{EP}(\Pi, s_0)$,
then $goodness(\pi) > goodness(\emptyset)$.

Positive Responsibility
If $\pi \in \mathcal{EP}(\Pi, s_0)$ *and* $newState(\pi) = newState(\pi')$,
then $\pi' \in \mathcal{EP}(\Pi, s_0)$, *for all* $\pi' \in \Pi$.

Negative Responsibility
If $\pi \in \mathcal{EP}(\Pi, s_0)$ *and* $oldState(\pi) = oldState(\pi')$,
then $\pi' \in \mathcal{EP}(\Pi, s_0)$, *for all* $\pi' \in \Pi$.

Harmlessness
If $\pi \in \mathcal{EP}(\Pi, s_0)$ *and* $harms(\pi') \subseteq harms(\pi)$,
then $\pi' \in \mathcal{EP}(\Pi, s_0)$, *for all* $\pi' \in \Pi$.

Consequential Equivalence This property is the hallmark of consequentialism. According to consequentialism, moral actions produce "good" consequences [Driver, 2011, p. 5]. The *Consequential Equivalence* property says that two plans with identical consequences must have the same ethical permissibility status. This situation also undermines the *impartiality* of a principle, as *impartiality* underscores the necessity for judgement to be determined by the outcome rather than by any affected party.

Maximal Welfare This property is closely aligned with the ethical principle of *Utilitarianism*. It emphasises that the right action must maximise the total amount of welfare [Driver, 2011, p. 68-69], and does not allow any plan to produce a better outcome than an ethically permissible one.

Aggregationism The most common view of this property is that general welfare must follow the additive aggregation of individual welfare [Driver, 2011, p. 69]. Although it is still up for debate regarding the appropriate aggregation function for general welfare, it is agreed that the welfare of two people is always better than the welfare of one.

Independence The principle of *Hedonism* highlights an assumption that ethical judgements are context blind; a plan's ethical permissibility status is independent of the available alternative plans [Grisez, 1978, p. 24]. *Independence* formally captures that assumption. We note, however, that judging an ethical plan usually depends on the context.

Beneficence Together with Independence, *Beneficence* provides a distinguishing property for the *Hedonism* principle. This property highlights that any acceptable principle must be beneficial to the overall goodness [Scarre, 2020, p. 27-33]. In other words, the overall goodness at the end of the plan must be better than the overall goodness at the beginning of the plan.

Positive Responsibility In action-based evaluation, we can observe the consequences of such actions and determine whether they achieve their responsibility [Gotterbarn, 2001]. The implementation varies where an ethical agent could be judged for bringing about new harm or a good result. Formally, we are saying that ethical permissibility should be determined only by the new state such actions produce. Together with *Harmlessness*, this property is closely aligned with the *No New Harm* principle.

Negative Responsibility Similarly, an ethical agent should be responsible for the consequences of inaction [Smart and Williams, 1973, p. 95]. If one can help others but chooses not to, they are indirectly responsible for what happens afterwards. On the other hand, if an agent decides not to do anything, the termination of any good becomes the agent's responsibility. *Negative Responsibility* captures that idea by requiring that an ethically plan should eliminate existing harm or maintain the current good, and thus, together with *Harmlessness*, is closely aligned with the *Remove Harm* principle.

Harmlessness John S. Mill introduced this property where the judgement of actions should focus on *harms*. Moreover, he said that we should only allow any action whose goal is to prevent harm [Mill, 1978, p. 9]. In this work, we formulate that the harm at the end of a plan should determine the permissibility of such a plan.

4.2 Adapting Social Choice Theory

Social choice theory focuses, among others, on designing protocols to select the most appropriate outcomes of preferences or choices [List, 2022]. Studies in this field typically follow several key aspects, including aggregating individual preferences into a collective or social preference, analysing the properties and characteristics of various voting systems, and investigating choice functions and preference relations with the primary objective of ensuring the consistency of individual preferences. A specific subfield of social choice theory concerns with individual choices [Moulin, 1985], which defines functions that align with our ethical permissibility function $\mathcal{EP}$. More specifically, we define our ethical permissibility function as a function that filters a set of plans into a set of ethically permissible plans. In the study of individual choice, a choice function is defined to return one's choice among the available options. Given the intriguing resemblance, we thus choose to examine social choice properties in machine ethics.

There are a number of well-studied social choice properties that aim to ensure the consistency of a protocol. Some of them are defined in [Lindström, 2022], which emphasise the robustness of a choice function (or selection function) when alternatives are manipulated. In this work, we thus adopt seven properties from [Lindström, 2022] and align their definitions with the the consequentialist principles.

Definition 3. ***Social Choice Properties***
Given an initial situation s_0, a set of admissible plans Π, and ethical permissibility function $\mathcal{EP}$, the following conditions demonstrate some natural properties for selection functions.

Consistency Preservation (cp)
If $\Pi \neq \emptyset$, *then* $\mathcal{EP}(\Pi, s_0) \neq \emptyset$.

Iteration (it)
$\mathcal{EP}(\mathcal{EP}(\Pi, s_0), s_0) = \mathcal{EP}(\Pi, s_0)$.

Cut (c)
If $\mathcal{EP}(\Pi, s_0) \subseteq \Pi' \subseteq \Pi$, *then* $\mathcal{EP}(\Pi, s_0) \subseteq \mathcal{EP}(\Pi', s_0)$.

Aizerman (aiz)
If $\mathcal{EP}(\Pi, s_0) \subseteq \Pi' \subseteq \Pi$, *then* $\mathcal{EP}(\Pi', s_0) \subseteq \mathcal{EP}(\Pi, s_0)$.

Chernoff (ch)
If $\Pi' \subseteq \Pi$, *then* $\mathcal{EP}(\Pi, s_0) \cap \Pi' \subseteq \mathcal{EP}(\Pi', s_0)$.

Sen (s)
If $\mathcal{EP}(\Pi, s_0) \cap \mathcal{EP}(\Pi', s_0) \neq \emptyset$,
then $\mathcal{EP}(\Pi \cap \Pi', s_0) \subseteq \mathcal{EP}(\Pi, s_0) \cap \mathcal{EP}(\Pi', s_0)$.

Arrow (iia)
If $\Pi' \subseteq \Pi$ *and* $\mathcal{EP}(\Pi, s_0) \cap \Pi' \neq \emptyset$,
then $\mathcal{EP}(\Pi, s_0) \cap \Pi' = \mathcal{EP}(\Pi', s_0)$.

Here Π and Π' are choice sets; alternative plans from which we can select. $\mathcal{EP}$ is a *selection function* that selects preferred outcomes from the choice set(s). An interpretation of, and justification for, these properties can be found in [Lindström, 2022]. We note Lemma 4.8 [Lindström, 2022] that establishes the following relationships: (c) implies (it), (ch) implies (c), (cp) and (iia) together imply (ch) and (aiz), and (iia) is equivalent to (ch) and (s).

Consistency Preservation This property in choice theory requires that if there are available alternatives, at least one must be chosen. Intuitively, it shows that even among terrible alternatives, some are better than others. It may not be as compelling in the context of ethical permissibility since none of the available plans may be ethically permissible.

Iteration This property requires an ethically permissible plan to remain so, and so repeated application of the $\mathcal{EP}$ function does not change the result of the initial application.

Cut & Aizerman *Cut* and *Aizerman* are conditional converses of each other. Jointly they are interpreted such that removing ethically impermissible plans from the set of plans would not cancel the permissibility of any ethically permissible plan. Thus it is a special case of the *contextual blindness of ethical permissibility* that we alluded to earlier.

Chernoff This property might be taken to be a variation of *Aizerman*. It requires that if some plan π in Π that is judged ethically permissible (with respect to Π) is available in the subset $\Pi' \subseteq \Pi$, then it should be judged ethically permissible with respect to (the smaller set) Π' as well. In other words, if π is ethically permissible, removing any plan other than π should not change the permissibility of π.

Sen From a different perspective, *Sen* proposes that if there is at least one plan π that is judged ethically permissible with respect to both sets Π and Π', then no other plans should be considered ethically permissible with respect to $\Pi \cap \Pi'$. This means, if arbitrary plan π' is added to Π, the new set of ethically permissible plans should only be considered from $\mathcal{EP}(\Pi, s_0)$ and π', not some other plans.

Arrow This is a strong property and implies *Chernoff* and *Cut*. It says that if a subset $\Pi' \subseteq \Pi$ has at least one plan π that is judged ethically permissible with respect to the larger set Π, then **all and only** such plans should be judged ethically permissible wrt the smaller set Π'. *Arrow* works similarly to *Chernoff*, however it requires a stronger condition: if π is ethically permissible, removing any plan other than π should not change the permissibility of *all plans in the new set*. It is known as *Independence of Irrelevant Alternatives*.

5 Results and Discussion

In §3 we formally defined five ethical principles: *Utilitarianism, Hedonism, No New Harm, Remove Harm* and *Harm Avoidance*. Inspired by the discussion found in the ethics literature, in §4.1 we formalised a number of properties that one would expect ethical principles to satisfy: *Consequential Equivalence, Maximal Welfare, Aggregationism, Independence, Beneficence, Positive Responsibility, Negative Responsibility* and *Harmlessness*. Finally, in §4.2 we adapted to our framework of ethical permissibility desirable properties from social choice: *Consistency Preservation, Iteration, Cut, Aizerman, Chernoff, Sen* and *Arrow*. One would expect that most ethical principles would satisfy most of these properties, although it would be unrealistic to expect every ethical principle to satisfy each of these properties. Our principal formal result vindicates that expectation:

Theorem 1. *Assume a set of (admissible) plans* Π *and an initial situation* s_0*. Define the five ethical principles in §3, and their* 15 *potential properties as in §4. Then:*

1. **Utilitarianism** *satisfies Cons. Equivalence, Maximal Welfare, Aggregationism, Consistency Preservation, Iteration, Cut, Aizerman, Chernoff, Sen,* and *Arrow;*
2. **Hedonism** *satisfies Consequential Equivalence, Aggregationism, Independence,* and *Beneficence;*
3. **No New Harm** *satisfies Consequential Equivalence, Positive Responsibility, Harmlessness, Iteration, Cut, Chernoff, Sen,* and *Arrow;*
4. **Remove Harm** *satisfies Consequential Equivalence, Negative Responsibility, Harmlessness, Iteration, Cut, Chernoff, Sen* and *Arrow;* **and**
5. **Harm Avoidance** *satisfies Consequential Equivalence, Harmlessness, Iteration, Cut, Chernoff, Sen,* and *Arrow.*

We have provided proofs of these comprehensive results at bit.ly/3JUVrU9. We have also verified which properties are satisfied by each ethical principle. To our knowledge this is the first research of its kind. We highlight and discuss some of the more interesting observations. To facilitate this discussion we summarise the results of Theorem 1 in Table 1. We hasten to note that we do not so far have any representation result. So, instead of characteristic features of different ethical principles, we discuss their *salient* or *distinguishing* features based on Table 1. We discuss in §5.1 the properties inspired by the ethics literature, followed by a discussion of the properties inspired by social choice theory in §5.2.

We recall that of the five ethical principles, *Utilitarianism* and *Hedonism* are motivated by promotion of "good" while the other three are based on the value we attach to reduction in "harm". Hence we contrast the properties of "good" based principles (*Utilitarianism* with those of *Hedonism*), and analogously those of the "harm" based principles.

5.1 Normative Properties of Ethical Principles

We start with Utilitarianism. As noted in §3, according to JS Mill, the four properties, *Consequential Equivalence*, *Maximal Welfare*, *Impartiality* and *Aggregationism*, jointly characterise Utilitarianism. Nonetheless, since *Consequential Equivalence* implies *Impartiality*, we may say the distinguishing properties of Utilitarianism are: *Consequential Equivalence*, *Maximal Welfare* and *Aggregationism*. An inspection of Table 1 readily shows that our formalisation of this principle satisfies all those three normative properties. Now, while *Utilitarianism* recommends maximal welfare, *Hedonism* **only** requires some comparative benefit (while choosing among available plans/actions). Yet, because of the implicit universal quantifier, *maximal welfare* can be satisfied in the absence of any comparative benefit. Hence *Maximal Welfare*, and *Beneficence* are respectively distinguishing features of *Utilitarianism* and *Hedonism*. It is also understandable why *Independence* is guaranteed by *Hedonism* but not *Utilitarianism*: someone not the poorest in Timbuktu is not the poorest person in the world, however the richest person in Timbuktu is not necessarily the richest person in the world.

Coming to the ethical principles based on the reduction of harm, we first note that the principle *Harm Avoidance* (roughly: Avoid all avoidable harms!) combines the power of the other two related principles: *No New Harm* (Do not produce a harm if it can be avoided) and *Remove Harm* (Eliminate a pre-existing harm if possible) which is easily proved:

$$\mathcal{EP}_{aHarm} \subseteq \mathcal{EP}_{nnHarm} \text{ and } \mathcal{EP}_{aHarm} \subseteq \mathcal{EP}_{rHarm}.$$

As expected, the *No New Harm* principle satisfies *Positive Responsibility* and the *Remove Harm* principle satisfies *Negative Responsibility*. This is in line with some previous works such as [Lindner *et al.*, 2020; Dennis *et al.*, 2021]. However, interestingly, and somewhat counter-intuitively, the *Harm Avoidance* principle satisfies neither. This clash in intuition is re-

	Utilitarianism	Hedonism	No New Harm	Remove Harm	Harm Avoidance
Consequential Equivalence	✓	✓	✓	✓	✓
Maximal Welfare	✓	✗	✗	✗	✗
Aggregationism	✓	✓	✗	✗	✗
Independence	✗	✓	✗	✗	✗
Beneficence	✗	✓	✗	✗	✗
Positive Responsibility	✗	✗	✓	✗	✗
Negative Responsibility	✗	✗	✗	✓	✗
Harmlessness	✗	✗	✓	✓	✓
Consistency Preservation	✓	✗	✗	✗	✗
Iteration	✓	✓	✓	✓	✓
Cut	✓	✓	✓	✓	✓
Aizerman	✓	✓	✗	✗	✗
Chernoff	✓	✓	✓	✓	✓
Sen	✓	✓	✓	✓	✓
Arrow	✓	✓	✓	✓	✓

Table 1: Properties of different ethical principles: ✓ indicates satisfaction and ✗ indicates non-satisfaction

solved when we carefully look at the above displayed formulation: all plans that are permissible by the *Harm Avoidance* principle are also permitted by each of *No New Harm* principle and the *Remove Harm* principle but not the converse. There might be plans permitted by the *No New Harm* principle (respectively *Remove Harm* principle) that will not be permitted by the *Harm Avoidance* principle. For instance, the *No New Harm* principle will permit plans that do not introduce any new harm but retain some of the pre-existing harms. In light of this, it is not surprising that the *Positive (negative) Responsibility* property is not satisfied by the *Harm Avoidance* principle.

5.2 Social Choice

We now observe the principles inspired by social choice theory. Between the principles based on *promotion of "good"*, the only social choice property that distinguishes them is *Consistency Preservation*. The reason that *Utilitarianism* satisfies this property is that, even if the available plans are all terrible and worsen the pre-existing situation, at least one of them would be least bad "maximising the welfare". However, since that least-bad plan will not bring any positive benefit, it will not be permitted by *Hedonism*. Social choice theory often talks about aggregating people's preferences, and it is usually the case that we can design a system that consistently produces an outcome. Meanwhile, an ethical dilemma can be pictured as a scenario where all permissible plans violate a certain ethical principle. In this case, no plan will be appropriate as an outcome. We can see here that, even though in the previous section we laid out the similarity between social choice and ethical principles, it is clear that they both have their distinct features.

We make two brief notes about the principles based on harm reduction. (1) All three principles fail to satisfy *Consistency Preservation*. This is not surprising since it is possible that the pool of available plans may not contain any that neither introduces new avoidable harms nor eliminates pre-existing eliminable harms. This suggests that, to better harness the power of social choice theory in the context of the principles based on harm reduction, some more intuitive properties should be formulated and examined. (2) The other interesting observation is that all three principles fail *Aizerman* but satisfy *Cut*, although these two properties are taken to be two sides of the same coin [Lindström, 2022]. Nevertheless, our formalisation of the *avoid-harm permission* approaches satisfy *Cut* but fail *Aizerman*. Though interesting, it is not really surprising since, unlike in social choice theory where the "best" alternatives are chosen, here **all** the plans that "reduce" (in the relevant sense) harm are being permitted. A larger pool of plans might have some harm-reducing members that are not in a subset of it explaining why *Aizerman* would fail.

6 Conclusions

This paper takes an essential step in the direction of formalising machine ethics by utilising situation calculus for reasoning about action and change. We introduce two important aspects: 1) a formalisation of several important consequentialist ethical principles from the philosophical literature; and, 2) a formalisation of several important properties that we claim should be considered when analysing and characterising approaches to machine ethics. Moreover, we prove properties satisfaction in our formalisation of ethical principles, aiding in distinguishing consequentialist ethics. This research is a crucial step for theoretical analysis and practical implementation of machine ethics, allowing future studies to validate their formalisations against our properties and providing precise characterisations of ethical principles.

Our work opens many important opportunities for future research. This includes various deontological and hybrid approaches, where prioritisation of principles is critical. Accordingly, it is also essential that, given several important distinguishing properties, future work can show that these principles produce a consistent result regarding the property satisfaction when we need to include uncertainty.

References

[Abel *et al.*, 2016] David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decision making. In *Workshops at the Thirtieth AAAI conference on Artificial Intelligence*, pages 54–61, 2016.

[Anderson and Anderson, 2011] Michael Anderson and Susan Leigh Anderson. *Machine Ethics*. Cambridge University Press, 2011.

[Arrow, 1951] Kenneth Joseph Arrow. Social choice and individual values. 1951.

[Asaro, 2019] Peter M Asaro. AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine*, 38(2):40–53, 2019.

[Baum, 2020] Seth D Baum. Social choice ethics in artificial intelligence. *AI & Society*, 35(1):165–176, 2020.

[Beckers, 2021] Sander Beckers. The counterfactual ness definition of causation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6210–6217, 2021.

[Bentzen *et al.*, 2018] Martin Mose Bentzen, Felix Lindner, Louise Dennis, and Michael Fisher. Moral permissibility of actions in smart home systems. In *Federated Logic Conference 2018*, 2018.

[Berreby *et al.*, 2017] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. A declarative modular framework for representing and applying ethical principles. In *16th Conference on AAMAS*, Sao Paulo, Brazil, May 2017.

[Bochman, 2018] Alexander Bochman. Actual causality in a logical setting. In *IJCAI*, pages 1730–1736, 2018.

[Bremner *et al.*, 2019] Paul Bremner, Louise A Dennis, Michael Fisher, and Alan F Winfield. On proactive, transparent, and verifiable ethical reasoning for robots. *Proceedings of the IEEE*, 107(3):541–561, 2019.

[Claßen and Delgrande, 2020] Jens Claßen and James Delgrande. Dyadic obligations over complex actions as deontic constraints in the situation calculus. In *Proceedings of the International Conference on KR*, volume 17, pages 253–263, 2020.

[Dennis *et al.*, 2021] Louise A Dennis, Martin Mose Bentzen, Felix Lindner, and Michael Fisher. Verifiable machine ethics in changing contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11470–11478, 2021.

[Driver, 2011] Julia Driver. *Consequentialism*. Routledge, 2011.

[Driver, 2013] Julia Driver. *Ethics: The Fundamentals*. John Wiley & Sons, 2013.

[Ghallab *et al.*, 2004] Malik Ghallab, Dana Nau, and Paolo Traverso. *Automated Planning: theory and practice*. Elsevier, 2004.

[Gotterbarn, 2001] Donald Gotterbarn. Informatics and professional responsibility. *Science and Engineering Ethics*, 7(2):221–230, 2001.

[Grisez, 1978] Germain Grisez. Against consequentialism. *American Journal of Jurisprudence*, 23, 1978.

[Hegde *et al.*, 2020] Aditya Hegde, Vibhav Agarwal, and Shrisha Rao. Ethics, prosperity, and society: Moral evaluation using virtue ethics and utilitarianism. In *29th International Joint Conference on Artificial Intelligence (IJCAI 2020). doi*, volume 10, 2020.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.

[Lindner *et al.*, 2020] Felix Lindner, Robert Mattmüller, and Bernhard Nebel. Evaluation of the moral permissibility of action plans. *Artificial Intelligence*, 287:103350, 2020.

[Lindström, 2022] Sten Lindström. A semantic approach to nonmonotonic reasoning. *Theoria*, 88(3):494–528, 2022.

[List, 2022] Christian List. Social Choice Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2022 edition, 2022.

[Loreggia *et al.*, 2018] Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K Brent Venable. Preferences and ethical principles in decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 222–222, 2018.

[McCarthy, 1958] John McCarthy. Programs with common-sense. In *Proceedings of the Symposium on Mechnization of Thought Processes*, volume 1, pages 77–84. Her Majesty's Stationery Office, 1958.

[Mill, 1863] John Stuart Mill. Utilitarianism. *London: Parker, Son, and Bourn*, 1863.

[Mill, 1978] John Stuart Mill. On liberty, ed. *Elizabeth Rapaport, Indianapolis: Hackett*, 1978.

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

[Moulin, 1985] Hervé Moulin. Choice functions over a finite set: a summary. *Social Choice and Welfare*, 2(2):147–160, 1985.

[Noothigattu *et al.*, 2018] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence, 32(1)*, 2018.

[Raphael, 1991] David Daiches Raphael. *British Moralists, 1650-1800: Hume*, volume 2. Hackett Publishing, 1991.

[Reiter, 2001] Raymond Reiter. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT press, 2001.

[Rodriguez-Soto *et al.*, 2021] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. Multi-objective reinforcement learning for designing ethical environments. In *Proceedings of the 30th IJCAI*, pages 545–551, 2021.

[Rossi and Mattei, 2019] Francesca Rossi and Nicholas Mattei. Building ethically bounded AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9785–9789, 2019.

[Sarmiento *et al.*, 2022] Camilo Sarmiento, Gauvain Bourgne, Daniele Cavalli, Katsumi Inoue, and Jean-Gabriel Ganascia. Action languages based actual causality in ethical decision making contexts. *CoRR*, abs/2205.02919, 2022.

[Scarre, 2020] Geoffrey Scarre. *Utilitarianism*. Routledge, 2020.

[Smart and Williams, 1973] John Jamieson Carswell Smart and Bernard Williams. *Utilitarianism: For and Against*. Cambridge University Press, 1973.

[Speith, 2022] Timo Speith. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, 2022.

[Takeshita *et al.*, 2023] Masashi Takeshita, Rzepka Rafal, and Kenji Araki. Towards theory-based moral AI: Moral AI with aggregating models based on normative ethical theory. *arXiv preprint arXiv:2306.11432*, 2023.

[Thornton *et al.*, 2016] Sarah M Thornton, Selina Pan, Stephen M Erlien, and J Christian Gerdes. Incorporating ethical considerations into automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 18(6):1429–1439, 2016.

[Tolmeijer *et al.*, 2020] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38, 2020.

[Winfield *et al.*, 2014] Alan FT Winfield, Christian Blum, and Wenguo Liu. Towards an ethical robot: Internal models, consequences and ethical action selection. In *Conference Towards Autonomous Robotic Systems*, pages 85–96. Springer, 2014.

[Zhan *et al.*, 2022] Xiao Zhan, Stefan Sarkadi, Natalia Criado, and Jose Such. A model for governing information sharing in smart assistants. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 845–855, 2022.