Uncertain Machine Ethical Decisions Using Hypothetical Retrospection

Simon Kolker, Louise Dennis, Ramon Fraga Pereira, and Mengwei Xu

Department of Computer Science, University of Manchester, UK
{simon.kolker, louise.dennis, ramon.fragapereira, mengwei.xu}@manchester.ac.uk

Abstract. We propose the use of the hypothetical retrospection argumentation procedure, developed by Sven Hansson, to improve existing approaches to machine ethical reasoning by accounting for probability and uncertainty. Actions are represented with a branching set of potential outcomes, each with a state, utility, and either a numeric or poetic probability estimate. Actions are chosen based on comparisons between sets of arguments favouring actions from the perspective of their branches, even those branches that led to an undesirable outcome. This use of arguments allows a variety of philosophical theories for ethical reasoning to be used, even in flexible combination with each other.
We provide an implementation of the procedure, applying both Consequentialist and Deontological ethical theories, independently and concurrently, to a use case based on an autonomous library recommendation system. While more work is necessary, our approach provides a practical framework that meets the varied requirements of a machine ethics system: versatility under multiple theories, values and beliefs; and a resonance with humans that enables transparency and explainability.

1 Introduction

Autonomous machines are an increasingly prevalent feature of the modern world. From spam filters [24] and fraud detectors [3], to drivers [28], medical practitioners [39] and soldiers [36], machines are being developed to automate tasks. Any decision that affects real people has the potential for ethical impact. Therefore, machines are increasingly recognised as ethical agents. Moor [30] categorises machines with ethical impact as either *implicitly ethical* or *explicitly ethical*. Implicit ethical agents are built and situated by humans to have a neutral or positive effect, like an ATM machine, but they do not utilise concepts of right and wrong internally in their decision making. As autonomous systems make more decisions with more responsibility, the need grows for them to reason about ethics *explicitly*. Allen et al. identify two strategies for designing explicitly ethical systems [4]: *bottom-up* approaches train systems to make ethical decisions with learning techniques based on data from human decision making; *top-down* approaches encode principles and theories of moral behaviour (often drawn from philosophy) into rules for use in a selection algorithm, generally using techniques from the field

of symbolic Artificial Intelligence (AI). In this paper, we propose and implement a top-down, explicitly ethical approach.

When an action is taken in the real world, its exact results are typically uncertain. As such, a top-down machine ethical system needs a mechanism for handling uncertainty over outcomes. There are a variety of mechanisms for representing uncertainty available for AI, including Bayesian methods, Dempster-Shafer theory, fuzzy logics and others [32]. Nevertheless, it is currently unclear how these approaches might integrate with existing philosophy in machine ethics; there may be unanticipated philosophical implications.

We have opted instead, to operationalise and implement Sven Hansson's hypothetical retrospection procedure [22]. Originating in philosophy, the procedure was designed to guide ethical reasoning under uncertainty. It favours no specific ethical theory, but systematises the foresight argument pattern, extending an assessor's perspective to judge decisions by the circumstances in which they were made. Therefore, arguments can be grounded in a variety of different ethical theories. Over the past ten years, the field of machine ethics has implemented many theories from philosophy [37], but there has been no consensus over which is the most effective. Philosophy too has not agreed which ethical theory is morally correct, leaving implementers of machine ethics to choose from the perspective of stakeholder requirements and preferences—hence a mechanism for handling uncertainty that can be adapted to different ethical theories is desirable.

We outline how the procedure works via an example introduced in Hansson [22]. Suppose an agent is given the choice between an apple and flipping a coin. If the coin lands heads, they win a free holiday to Hawaii. If the coin lands tails, they get nothing. Selecting the coin is clearly a valid choice. How might this decision be justified? Under hypothetical retrospection, we list each possible outcome: choosing the apple; choosing to toss the coin and winning Hawaii; choosing to toss the coin and losing. Next, we *hypothetically retrospect* from each outcome's endpoint. Intuitively, the objective is to find outcomes which do not lead the agent to *regret* their action. First consider the coin's outcomes. After winning Hawaii, there cannot be any regret since Hawaii is the best outcome. Otherwise, after losing the agent has nothing, which is the worst outcome. However, there is no regret since the agent can justify that they had a good chance of winning Hawaii, which is far better than an apple. Now, consider choosing the apple. Here, there is regret since the agent missed a chance of a holiday, which is worth far more than the apple. We can see that selecting the coin does not lead to this regret. Therefore, the procedure advises we pick the coin, matching our intuition.

This paper operationalises the hypothetical retrospection procedure, and the foresight argument pattern it is based on. We implement the procedure and evaluate with moral theories from philosophy. We consider deontological theories, which specify a set of actions that are strictly forbidden [2], and a theory of Act-Consequentialism, which specifies that an action is only good if its consequences maximise good for the greatest number of people [31]. We illustrate our approach in action under the novel scenario of an autonomous library system.

We demonstrate the system's potential for explainability and versatility, while discussing issues and future work.

In Section 2, we will cover related work in the area and highlight this paper's contribution. In Section 3, we will cover background on symbolic argumentation and uncertainty in ethical philosophy. In Section 4 we will recap Hansson's description of hypothetical retrospection; in Section 5 we overview the implementation, including notation, the representation of probability and the argumentation model. Section 6 describes our test case of the autonomous library system, its formalism, and our results. Finally, in Section 7 we will identify the system's potential benefits and its shortfalls, left for future work.

2 Related Work

This is not the first attempt at building a top-down explicitly ethical machine. Tolmeijer et al. presents an exhaustive survey of implementations as of 2020, but finds the effect of uncertainty is rarely addressed [37]. Dennis et al. developed a framework suggesting how an autonomous system should act in unforeseen circumstances, where there are no positive outcomes. However, it does not address uncertainty between the likelihood of outcomes [18]. Probabilistic reasoning, such as Bayesian networks [35] and Markov models [17], has been applied to machine ethics, mostly with regards to maximising expected utility [15]. There are a number of criticisms of this approach which we will touch on in Section 3. Killough et al. goes further, architecting agents sensitive to utility risk and reward, with an ability to dynamically adjust risk-tolerance for the environment [26].

This paper is interested in a framework that incorporates a variety of philosophical ethical theories and allows for the combination of multiple theories, such as Deontology [2], Contractualism [7] and Virtue Ethics [23]. Different philosophical theories can advise on different courses of action, not only in tricky dilemma situations but sometimes even in situations where the moral choice seems intuitively obvious. There has been some work within machine ethics on comparing and combining different theories. For instance, Sholla et al. weights different principles and then uses fuzzy logic to decide between their recommendations [34]. Ecoffet and Lehman [20] use a voting procedure in which different ethical theories vote on recommendations but struggle with the difficulty of comparing consequentialist theories that return a score for actions with deontological theories that tend to return a judgement that the action is either permissible or impermissible. Our framework enables a flexible approach in which the construction of an argument can treat all ethical theories equally, or allow one to have precedence over another. The HERA project [27] is of interest here – while it does not combine ethical theories it provides a single framework in which many theories can be formalised and operationalised, allowing their recommendations to be compared. Cointe et. al [16] do something similar with an Answer-Set Programming approach though focused, in this case, on enabling the agent to make moral judgements about others. These systems could, potentially, be integrated

into our argumentation framework to supply judgements on the rightness of an action and its consequences from the perspective a particular moral theory.

Atkinson and Bench-Capon have developed a framework for ethical argumentation [8]. Like our work, assessments of action's outcomes are modelled as arguments. However, Atkinson and Bench-Capon's work remains concerned with epistemic conflicts between arguments (i.e. disputes between the truth of argument's circumstances) and annotates attacks and defends within the argumentation framework with values, aligning it with the philosophical theory of Virtue Ethics. Our work pivots away, focused purely on the ethical conflicts between arguments. We can assume epistemic truth because arguments are based only on potential, purely hypothetical, versions of events, each created from a single, shared set of information. This allows us to address moral conflict directly. It also lets us build uncertainty into the argumentation mechanism, instead of delegating it to a detail of argument attacks.

3 Background

The effect of uncertainty on machine ethics has been relatively unexplored largely due to the lack of research on how uncertainty impacts ethics in general. As Altham explains, there seems to be a gap in moral theory for uncertain situations [5]. He postulates this could be due to a belief among philosophers that no special principles are required; moral philosophy decides the virtues and it is up to decision theory to decide how they should be maximised under uncertainty.

Hansson shows that consequentialist utilitarian theories are straightforward in this regard [22]. The notion of expected utility uses probabilities as weights to discount the value of improbable outcomes. Hansson critiques this adaptation for the same reason as actual Utilitarianism: its assumption that outcomes can be appraised in terms of a single number (or at least done so both easily and accurately) often produces unintuitive outcomes. In the Apple-Coin scenario from Section 1, although it is evident that a trip to Hawaii holds more value than an apple, the extent of the difference in value remains uncertain. Adding more apples, such as 100, 1000, or 1001, does not necessarily make the deal any more appealing. In other words, apples and holidays are not proportionally comparable. There is no method of assigning relative utilities to all possible states. Brundage briefly surveys other critiques against consequentialist theories in the following points. First, they fail to account for personal social commitments, i.e. to friends and family. Second, they do not consider individual differences and rights, tending to favour the majority over any minority. Lastly, they place excessive demands on individuals to contribute to others [13].

Traditional deontological ethical systems [2] are made of principles which should never be violated. Hansson points out that any form of probabilistic absolutism, where an action is not permitted if there is any chance of a rule violation, would be too restrictive. Therefore, an approach involving probability thresholds is often suggested to solve this problem. Here, an action is only forbidden when the probability that it violates a law exceeds some limit. The exact

value of this limit is open for debate. It is tempting to suggest the limit should have some relation to the action's potential benefits, but this could soon reduce to some elaborate form of Utilitarianism, adamantly against the essence of the original theory.

Noticeably, most humans do not consciously rely on one philosophical, moral theory to make their decisions [12]. Nor do we think it is our place to choose a single theory to apply to machine ethics. As such, one of Hansson's key contributions is providing an argumentation procedure that can frame multiple, possibly conflicting theories rationally. To model this, we look to the study of abstract argumentation. Dung creates a framework of logically generated, non-monotonic arguments [19]. They can discredit each other with attacks, modelled as a binary relation between the arguments. Dung goes on to specify properties of a well-founded framework; he gives procedures for believing arguments based on their membership to framework extensions. This paper will take only take the simple structure of Dung's framework. We leave it to Hansson's philosophy to define attacks and select arguments.

4 Hypothetical Retrospection

In this section, we overview Hansson's description of the hypothetical retrospection procedure, as given in [22]. Hypothetical retrospection is an argumentation framework that systematises perspective-extending argument patterns.

Much of moral philosophy can be interpreted as an attempt to extend a decision maker's perspective. In promoting empathy, we invoke a perspective extending argument pattern to consider other's perceptions of our actions. For cases of uncertainty, it is instead helpful to extend our perspective with future perceptions of our actions. This means viewing, or hypothetically retrospecting, on a choice from the endpoint of its major foreseeable outcomes. As a result, the potential outcomes, or the *branches of future development*, can be referred to in a valid argument about what to do in the present.

Hansson lays principles for determining each action's branches as a search problem. Theoretically, the effects of decisions may be infinitely complex and far-reaching. The major search principle, therefore, is to find the most probable future developments which are the most difficult to defend morally. This will increase the chance of considering unethical scenarios. Searched branches should be described to an endpoint sufficiently far to capture all morally relevant information. All intermediate information has to be captured too: rule violations occurring before the point of retrospection still need to be considered. Additionally, and for the sake of comparison, branches should be described with the same type of information where possible[1]. Hansson sees no reason not to create alternate branches based on the uncertainty of the decision maker's own future

[1] The way in which consequences are discussed here may seem to exclude non-consequentialist theories. Hansson emphasizes that this is not the case. In his approach, consequences are broadly defined and their *information* includes agency, virtue intentions, and any other information necessary for moral appraisal.

choices, considering human's inability to control their future actions. Whether an autonomous system has uncertainty over its own future actions depends on the nature of the agent and its application architecture.

For this paper, we focus on the assessment of actions, assuming their potential branches have already been found. In future work, a planning algorithm could be adapted to the requirements above. For instance, the Probabilistic Planning Domain Definition Language (PPDDL) [38] is able to formalise different stochastic planning settings, e.g., Markov Decision Process (MDP) [21], Stochastic Shortest Path problems (SSP) [11], and Fully Observable Non-Deterministic planning [14]. This has recently been superceded by the Relational Dyanamic Influence Diagram Language (RDDL) [33] which has been adopted by the International Probabilistic Planning Competition (IPPC)[2] and is thus the target input language for many planning implemenations.

Using their potential branches, actions are morally judged. Judgements are purely comparative. Hansson stresses we are not to assess actions in isolation; assessments are to be composed of comparisons. This is because decisions are not made in isolation. Given a choice between actions A and B, choosing A is choosing A-instead-of-B. Building action assessments from comparisons ensures all morally relevant information is taken into account.

Actions are compared by hypothetically retrospecting from the endpoint of each action's potential branches of future development. We search for an action which never leads an agent to morally regret its choice in retrospect. Hansson argues against the term *regret* since it is considered a psychological reaction; humans often feel regret for actions they did not commit, or that they could not have known were wrong. By regret, therefore, we mean that the decision making was logically flawed under retrospection. As a result, we shall use the term *negative retrospection* to reflect this more technical definition. By hypothetically retrospecting between actions' branches, we search for an action which does not lead to negative retrospection, or has full acceptability among its branches. If no such action exists, one should be selected that maximises acceptability in its most probable branches.

5 Implementation

5.1 Formalism

We formalise an ethical decision problem as follows.

Definition 1 (Ethical Decision Problem). *An Ethical Decision Problem is a tuple of* $\langle A, B, S, U, F, I, m\rangle$ *where* A *stands for a set of available actions,* B *the set of all possible branches of future development,* S *the set of Boolean state variables,* U *an ordered set of* utility classes*,* F *a set of forbidden state assignments,* I *the initial assignment of boolean values to the variables in* S*, representing the initial state, and* $m : A \to \mathcal{P}(B)$ *(where* $\mathcal{P}$ *is the powerset function) is a mapping of actions to possible branches of development.*

[2] https://ataitler.github.io/IPPC2023/

The mapping m, associates every action, $a \in A$, with its potential branches of future development. Each branch, $b \in m(a)$ is an ordered sequence of *events* that could occur after action a.

Definition 2 (Event). *An event is a tuple of $\langle s, \phi, p \rangle$ where $s \in S$, ϕ is the new Boolean value of s, and p is the probability that the event occurs.*

An event therefore represents the change in value of one state variable in S. A branch is a sequence of events that can occur after the action is taken.

As an example, we formalise the Coin-Apple scenario. There are three state variables: s_1 represents whether or not we have an apple, s_2 whether or not we have gambled, and s_3 is whether or not we won a trip to Hawaii. In the initial state I, all these variables are false. There are two available actions in the set A. Action a_1 represents choosing the apple. It maps to one branch $b_1 \in m(a_1)$, containing one event, $\langle s_1, True, 1 \rangle$—if we choose to have an apple, we gain an apple; we have not gambled or won a holiday to Hawaii. Action a_2 represents flipping the coin. It maps to two branches, $b_2, b_3 \in m(a_2)$. The branch b_2 contains one event, $\langle s_2, True, 1 \rangle$—we gambled, but we have no apple and no holiday to Hawaii. The branch b_3 is the sequence of events $\langle s_2, True, 1 \rangle$ then $\langle s_3, True, 0.5 \rangle$—first we gambled, then we won a holiday to Hawaii. The formalism is represented graphically in Figure 1.

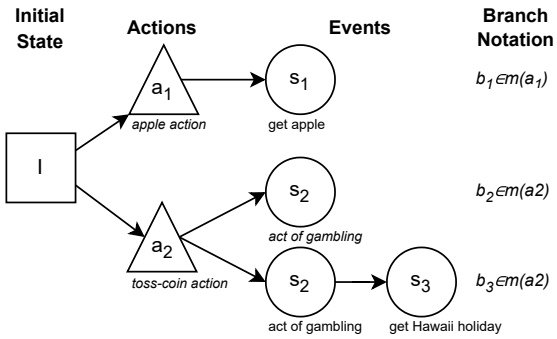

Fig. 1. Diagram for Coin-Apple scenario. Event nodes represent assignment to a state variable. Actions map to a set of branches, represented by rows of event nodes.

To capture the issue described in Section 1, where the outcomes of some events have an immeasurably greater/lower utility than others, we have introduced utility classes. U is an order over such classes.

Definition 3 (Utility Class). *A utility class is an unordered set of individual utility assignments represented as tuples of $\langle s_k, \phi, v \rangle$, where s_k denotes a state variable in S and $v \in \mathbb{R}$ represents the variable's utility when assigned Boolean value ϕ.*

The utility classes in U are ordered by importance in descending order. Where $i < j$, all the positive utilities in u_i are considered greater than any utility in u_j; all the negative utilities in u_i are considered less than any utility in u_j. To reiterate, the absolute utilities in lower indexed classes are immeasurably greater.

Continuing the Coin-Apple example, we define two utility classes in U, with one assignment each. The first class contains utility assignment, $\langle s_3, True, 1\rangle$ representing a utility of 1 for the getting the Hawaii holiday. The second class has utilities immeasurably lower. It contains one assignment, $\langle s_1, True, 1\rangle$ representing a utility of 1 for getting the apple.

To incorporate a wider range of theories, F defines the states forbidden by deontological theories. This is not the same as defining a negative utility in U. Negative utilities can be outweighed by a greater positive utility. In certain decision making environments, forbidden states can not be outweighed. They could represent, for instance, that someone was deceived, that a law (e.g., trespass) was broken, and so on – any action or outcome that can not be justified.

Definition 4 (Forbidden State). *A Forbidden State is a tuple $\langle s, \phi\rangle$ where $s \in S$ denotes a state variable forbidden from being assigned the Boolean value ϕ.*

For the Coin-Apple example, F could contain a forbidden state, $\langle s_2, True\rangle$ representing a rule against gambling.

This formalism assumes that the high-level rules have been translated into domain-level rules, applicable to the variables in the state variables, S.

Given an ethical decision problem defined as a tuple of $\langle A, B, S, U, F, I, m\rangle$, a solution is a permissible action from the set of actions A, formally defined as follows.

Definition 5 (Permissible Action). *Given an ethical decision problem, defined as a tuple of $\langle A, B, S, U, F, I, m\rangle$, a permissible action is an action, $a \in A$, such that for all potential branches of future development $b \in m(a)$, there is acceptability over their events in state space S. If no such actions exist, action a is permissible if it maximises the cumulative probability of its acceptable branches.*

The definition of whether a branch is acceptable depends upon the ethical theory or theories under consideration (see Section 5.3).

5.2 Probability Representation

An additional feature of this formalisation is support for estimative probabilities as well as exact probabilities. In many scenarios, while a person may have an intuition that some events are more probable than others, their exact probabilities are unknown. Even so, a decision may need to be reached. This is most common when interacting with humans and complex systems.

Kent found that intelligence reports tend to use *poetic* words to describe probabilities, like *probable* or *unlikely* [25]. The issue is that people have different interpretations of their meaning. Kent defined a relation for poetic words to mathematical probability ranges, as given in Table 1 from [25]. Our implementation supports using these estimates instead of exact numerical probabilities.

100% Certainty			
The General Area of Possibility	93%	Give or take 6%	Almost Certain
	75%	Give or take 12%	Probable
	50%	Give or take 10%	Chances about even
	30%	Give or take 10%	Probably not
	7%	Give or take 5%	Almost certainly not
0% Impossibility			

Table 1. Mathematical to poetic relation from Kent's estimative probability [25].

5.3 Argumentation Model

Given each action, and its potential branches of future development, a system can hypothetically retrospect from each endpoint and search for an action with complete acceptability. We have implemented this with a simple argumentation network, partially based on the work of Atkinson et al. [9]. We implemented their process by generating arguments for each branch, representing positive retrospection on its action.

As in Atkinson et al. [9], arguments are generated logically from an *argument scheme*. For an action $a \in A$, selected in initial state I, resulting in the branch $b \in m(a)$ with probability p, the following argument is generated:

"From the initial state I, it was acceptable to perform action a, resulting in consequences b with probability p."

For notation, this argument will be written as $Argument(b)$. This can be viewed as a default starting argument that any action is acceptable. The retrospective argument below is generated for branch b_3 in our running example: namely, winning the Hawaii holiday after tossing the coin.

"From the initial state I, where $s_1 = s_2 = s_3 = False$, it was acceptable to perform the action a_2, resulting in consequences with $s_2 = s_3 = True$ with probability 0.5."

To find whether a retrospective argument is correct, we search for attacks from other action's arguments. An incoming attack implies negative retrospection for not choosing the attacking argument's action. Attacks are generated by posing critical questions on the arguments' claims [9]. For the branches $b_1 \in m(a_1)$, $b_2 \in m(a_2)$ and any generic moral principle, the following critical questions are asked for $Argument(b_1)$ to attack $Argument(b_2)$.

CQ1 *Does b_2 violate a moral principal that b_1 does not?*
CQ2 *Does a_2 hold a greater probability of breaking the moral principle than a_1?*

$Argument(b_1)$ only attacks $Argument(b_2)$ if both of these questions are answered positively. They represent negative retrospection for missing the chance to avoid violating a principle. The critical questions are asked both ways between

all arguments supporting different actions, for every moral principle under consideration. The time and space complexity of answering the questions will differ for different theories. The critical questions for embedding Utilitarianism and a generic deontological Do-No-Harm principle are given as follows:

- Utilitarian CQ1: *Does* b_2 *bring greater utility value than* b_1?
- Utilitarian CQ2: *Did* a_2 *expect greater utility value than* a_1?
- Do-No-Harm CQ1: *Does* b_2 *cause harm where* b_1 *does not*?
- Do-No-Harm CQ2: *Did* a_2 *expect a greater probability of causing harm than* a_1?

After searching for attacks on all branches, an action should be selected with complete acceptability. If no such action exists, an action should be selected with maximal acceptability, i.e. summing the probability of each non-attacked argument and selecting an action with a maximal sum.

6 Autonomous Library Test Case

To illustrate our implementation in action, we present an uncertain ethical decision problem and discuss our results under five sets of ethical considerations.

Suppose a student logs onto their University's autonomous library to revise for a test the next morning. All the other students started revision a month ago. As the student constructs various search terms for a recommendation, the system recognises that all other students have taken out the same book, implying it is very useful.

Should the autonomous library system use this data to recommend the book, allowing the student to revise quicker on the night before the test? While the student would prefer this, if the others find out, they may feel unfairly treated; they get no more credit than someone who simply waits for them to find the reference and then use it.

The potential outcomes and probabilities are depicted in Figure 2. If the agent recommends, there is a 0.6 chance the student will use the book. If the student has the book, there is a 0.7 chance they will pass; without the book, there is a 0.7 chance they will fail. If the book is recommended, there is a 0.05 chance the other students will find out.

6.1 Problem Model

For the Autonomous library scenario, we define an ethical decision problem $\langle A, B, S, U, F, I, m\rangle$, where A is a set of two possible actions. In detail, the action a_1 represents the action to *recommend* the book and action a_2 to *ignore* the book. Each action's potential branches of future development depend on how far into the future the branches' endpoints are set, and how their information is described.

To capture all the morally relevant information, the *recommend* action a_1 leads to eight branches, while the *ignore* action a_2 leads to two

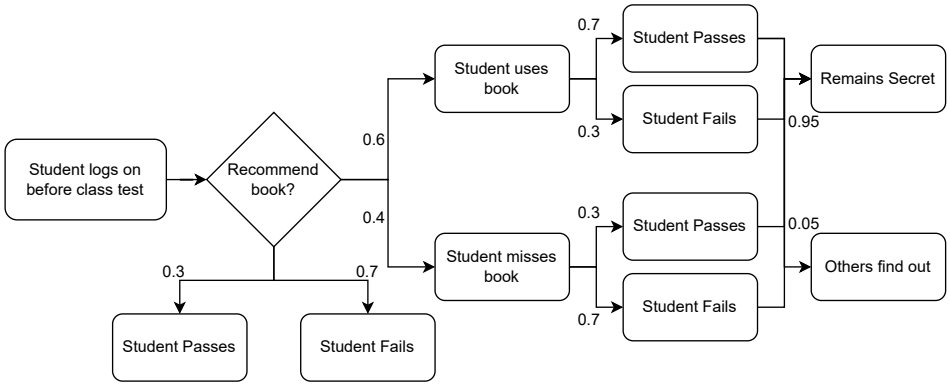

Fig. 2. Tree describing possible events in the autonomous library problem.

branches. These branches change state across four variables in S, namely $s_1 =$ *dataProtectionViolation*, $s_2 =$ *readsBook*, $s_3 =$ *passesTest*, $s_4 =$ *othersFindOut*. The *othersFindOut* variable refers to the class finding out their data was used to help the student. The *dataProtectionViolation* variable refers to the act of using data for a non-consented purpose. Finally, each state variable is *False* in initial state I.

Based on the tree in Figure 2, we develop Table 2, where each action's branches of future development are represented with their numeric probability, given their action is selected.

$b_1 \in m(a_1)$	$\langle s_1, True, 1\rangle$, $\langle s_2, True, 0.6\rangle$, $\langle s_3, True, 0.7\rangle$, $\langle s_3, False, 0.95\rangle$	0.399
$b_2 \in m(a_1)$	$\langle s_1, True, 1\rangle$, $\langle s_2, True, 0.6\rangle$, $\langle s_3, False, 0.3\rangle$, $\langle s_4, False, 0.95\rangle$	0.171
$b_3 \in m(a_1)$	$\langle s_1, True, 1\rangle$, $\langle s_2, False, 0.4\rangle$, $\langle s_3, True, 0.3\rangle$, $\langle s_4, False, 0.95\rangle$	0.114
$b_4 \in m(a_1)$	$\langle s_1, True, 1\rangle$, $\langle s_2, False, 0.4\rangle$, $\langle s_3, False, 0.7\rangle$, $\langle s_4, False, 0.95\rangle$	0.266
$b_5 \in m(a_1)$	$\langle s_1, True, 1\rangle$, $\langle s_2, True, 0.6\rangle$, $\langle s_3, True, 0.7\rangle$, $\langle s_4, True, 0.05\rangle$	0.021
$b_6 \in m(a_1)$	$\langle s_1, True, 1\rangle$, $\langle s_2, True, 0.6\rangle$, $\langle s_3, False, 0.3\rangle$, $\langle s_4, True, 0.05\rangle$	0.009
$b_7 \in m(a_1)$	$\langle s_1, True, 1\rangle$, $\langle s_2, False, 0.4\rangle$, $\langle s_3, True, 0.3\rangle$, $\langle s_4, True, 0.05\rangle$	0.006
$b_8 \in m(a_1)$	$\langle s_1, True, 1\rangle$, $\langle s_2, False, 0.4\rangle$, $\langle s_3, False, 0.7\rangle$, $\langle s_4, True, 0.05\rangle$	0.014
$b_9 \in m(a_2)$	$\langle s_3, True, 0.3\rangle$	0.3
$b_{10} \in m(a_2)$	$\langle s_3, False, 0.7\rangle$	0.7

Table 2. Each action's alternate branches of future development for the autonomous library problem. S contains state variables: $s_1 =$ *dataProtectionViolation*, $s_2 =$ *readsBook*, $s_3 =$ *passesTest*, $s_4 =$ *othersFindOut*. Events setting variables to $False$ are not required, but included for readability.

An argument is generated from each branch representing positive retrospection. Using the argument scheme from Section 5, $Argument(b_1)$ is as follows:

"From the initial state, I, where $s_1 = s_2 = s_3 = s_4 = False$, it was acceptable to perform the action, a_1, resulting in consequences with $s_1 = s_2 = s_3 = True$, with probability 0.399."

In informal terms, the argument claims it was acceptable to recommend the book, resulting in a data protection violation, the student reading the book, the student passing, and nobody finding out about the data breach, all of which had 0.399 probability.

6.2 Consequentialism with One Assignment

We begin testing our implementation only considering the ethical theory of Act-Consequentialism. We set U to have one utility class, containing one utility assignment, $\langle passesTest, 1, True \rangle$. The only value is the student passing the test. Intuitively, the action which maximises the probability of passing should be chosen and hypothetical retrospection agrees. The argumentation graph given in Figure 3 shows the retrospection.

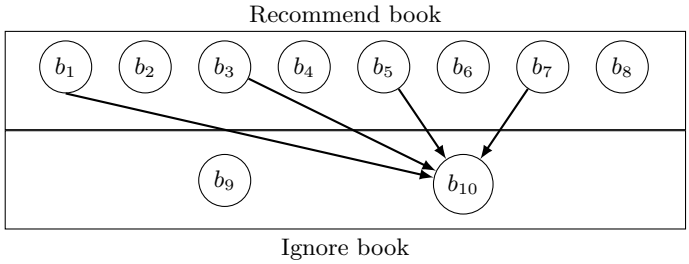

Fig. 3. Graph of retrospection between hypothetical branches of development with only the utility of the student passing in consideration.

Every branch has acceptability, except $b_{10} \in m(a_2)$ where the student fails after the system chooses *ignore*, with 0 utility and 0.3 probability (*'probably not'* in Kent's words). This branch has a lower utility than the four *recommend* branches where the student passes: $b_1, b_3, b_5, b_7 \in a_1$. This causes $Argument(b_{10})$ to answer Critical Question 1 positively when attacked by these branches. Since *recommend* has a greater utility expectation, or a greater probability of the student passing, $Argument(b_{10})$ cannot defend itself in Critical Question 2. There is no reason to select *ignore*; from the perspective of b_{10}'s endpoint there is negative retrospection. There are no other attacks. Therefore by hypothetical retrospection action a_1, *recommend*, should be selected.

6.3 Utilitarian with Two Equal Assignments

Now we consider two utility assignments of the same class: $\langle passesTest, 1, True \rangle$ and $\langle othersFindOut, -1, True \rangle$. This invokes the risk of others finding out their

data was used. However, having the students find out is judged to be as bad as the student passing is good. Retrospection is shown in Figure 4.

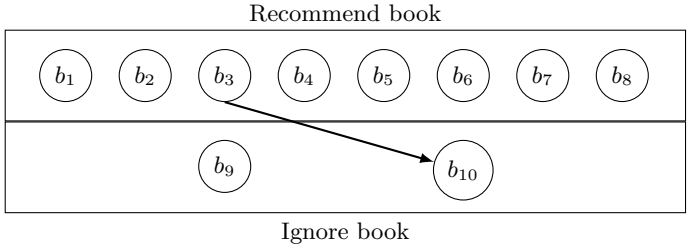

Fig. 4. Graph of retrospection between hypothetical branches of development with the cost of others finding out data was compromised equaling the utility of the student passing.

Again, the only branch with negative retrospection is $b_{10} \in m(a_2)$, when the system selects *ignore* and the student fails. This time, only two of *recommend*'s branches have greater utility, $b_1, b_3 \in m(a_1)$. Action *recommend* has a greater utility expectation, so *ignore* cannot be defended. Therefore, *recommend* is chosen again.

6.4 Consequentialism with Two Unequal Assignments

Judging that others finding out has -1 utility is subjective. It could be lowered such that *recommend*'s expected utility is lower than *ignore*'s. This is true for the following distribution: $\langle passesTest, 1, True\rangle$ and $\langle othersFindOut, -5, True\rangle$ in one utility class in U. Retrospection is shown in Figure 5. Finally, attacks fire the other way. When the system chooses *recommend* and other students find out, as in $b_5, b_6, b_7, b_8 \in m(a_1)$, there is utility lower than either of *ignore*'s branches. There is no defence since *ignore* has a greater utility expectation. *Recommend* can still lead to high utility branches with b_1 and b_3, but unlike before, b_{10} defends by citing *ignore*'s higher utility expectation. Thus, *ignore* is selected with full acceptability.

6.5 Consequentialism with Two Incomparable Assignments

Deciding the utilities is difficult without further detail. If the student's grades are poor, the utility of $passesTest$ should be raised. If the students could press legal action, the utility of $othersFindOut$ should be lowered; if students were pleased their data helped a fellow student, $othersFindOut$ should be raised.

Ideally we would continue developing branches until enough morally relevant information is described. This is not always computationally viable and it ignores the second half of the problem. The issue is that, even with all the relevant

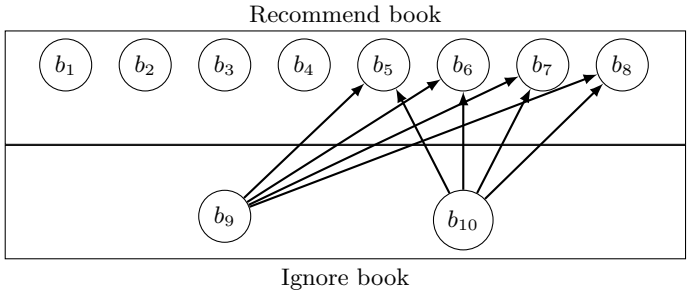

Fig. 5. Graph of retrospection between hypothetical branches of development with the cost of others finding data was compromised far out outweighing the utility of the student passing.

information, assigning exact utilities is subjective. We confront this with the notion of utility classes. If we suppose $othersFindOut$ has negative utility, but immeasurably so compared with $passesTest$, we can form two utility classes. The first class has assignment $\langle othersFindOut, -1, True\rangle$ and the second class has $\langle passesTest, 1, True\rangle$. This leads to the same retrospection as Figure 5; the cost of others knowing outweighs the benefits of passing.

6.6 Deontology

Finally we consider a deontological theory objecting to the act of misusing the others' data. For the demonstration, it does not matter which ethical theory is being used or which rule is violated. For instance, collecting and using data without consent is considered a violation of the UK's Data Protection Act, requiring "personal data to only be used for specified, explicit purposes." [1] and so this could violate a deontological rule against breaking the law. It could also be a violation of Deontology's Doctrine of Double Effect. This has four conditions [29]: 1. that the action in itself from its very object be good or at least indifferent; 2. that the good effect and not the evil effect be intended; 3. that the good effect be not produced by means of the evil effect; 4. that there be a proportionately grave reason for permitting the evil effect. If we consider non-consensual use of students' data as bad and helping a student to pass the exam to be good, then the fact that the bad effect is required in order to bring about the good effect breaks the third condition above, and, therefore, is not permissible.

We build on our first retrospection, shown in Figure 3, which had one utility assignment $\langle passesTest, 1, True\rangle$, and a permissible action, *recommend*. Now we add a forbidden state $\langle dataProtectionViolation, True\rangle$ to the set of forbidden states F. Retrospection is shown in Figure 6. We observe that every argument from *ignore* attacks every argument from *recommend* because all branches in *recommend* contain the forbidden event, and none in *ignore* do.

Under Utilitarianism, the action *recommend* is still the better choice, with the same attacks on $Argument(b_{10} \in m(a_2))$ as before. This conflict represents

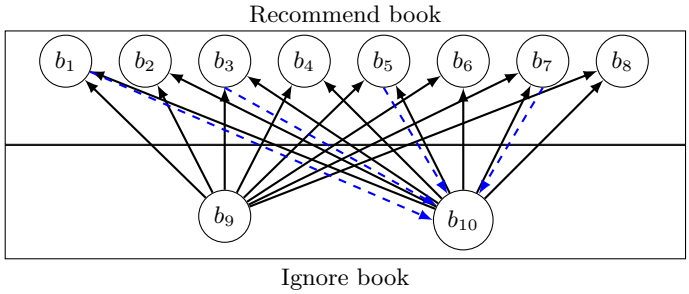

Fig. 6. Graph of retrospection between hypothetical branches of development with one consequentialist assignment and one deontological law. Consequentialist attacks are dashed blue; deontological attacks are solid black.

a moral dilemma, where no choice is normatively inferior to another [22]. The aim is to maximise acceptability amongst the most probable branches. Since all arguments from *recommend* are attacked, there is 0 acceptability; one argument from *ignore* is attacked with 0.7 probability, or *probable* in Kent's words. Therefore, *ignore* is selected with 0.3 acceptability.

7 Discussion

Our goal here is to extend beyond the typical approach to machine ethics, which is the assessment of a single action from the perspective of a single ethical theory often without any account of probability or uncertainty. We have formalised Hansson's theory of hypothetical retrospection. This formalisaiton systematises moral assessments as comparisons, between sequences of consequences in a way that allows richer judgements to be formed beyond just the evaluation of utilities. Our moral assessments are comparisons between retrospective justifications of hypothetical consequences. One may might ask how this differs from directly analysing the properties of consequences? For machines, it gives a procedure for selecting actions and providing justifications. For humans, it offers realism that sharpens our intuitions and allows us to make clearer judgements [22]. It also enables us, in the future, to build on existing work for evaluating actions from the perspective of individual ethical theories and combining those judgements into arguments. Essentially our proposal extends, rather than replaces, existing mechanisms for evaluating actions against a single ethical theory.

The retrospective procedure provided by the two critical questions resembles real life discussion: a claim against an argument and a chance to refute. It's common that, if someone took action a_2 and a moral principle was broken, they may be accused of wrongdoing on the basis that they could have taken a_1. The retrospective argumentation would follow the steps modelled by the two critical questions:

1. You should have chosen a_1 because it didn't break this moral principle.

2. No, because there is a greater probability of breaking some other principle with a_1. If I was given the decision again, I would make the same choice

Admittedly, real life discussion may not be so civil, but if all facts were agreed upon, this dialogue is the logical nature of discussion. Resemblance to real life has clear utility for agent transparency and explainability, both of which are important for ethical AI [10] and for stakeholder buy-in to the use of AI systems.

Our implementation could be adapted into a module to be placed on top of an existing autonomous system, possibly similar to Arkin's governor architecture [6]. So long as appropriate domain information is supplied, our implementation can be used in principle. The implementation is also theory-neutral, in that it allows multiple principles and theories to be considered at once, more analogous to human decision-making. Implementational work remains to be done, not least the integration into a planning system to generate branches, but also evaluation against a wider range of ethical theories (e.g., Virtue Ethics) to see how easily they can provide information to the critical questions. We also wish to further develop the evaluation of action's consequences along branches, not just at the branches end – for instance, if someone is made unhappy as a consequence of some action, but then we compensate them so they are happy by the end of the exploration of the branch, can we ignore the fact that we caused them (albeit temporary) unhappiness?

Our current implementation has a fairly simple approach to the integration of ethical theories. Some theories are directly incompatible, potentially leading to situations where a "worst of both worlds" option is chosen. Additionally, the use of utility classes needs careful handling. When utilities are of a greater class, they are prioritised, no matter how remote their probabilities. Extending the Coin-Apple scenario, suppose an agent is offered a free apple every day – as opposed to some number of apples all at once, or suppose the chance of winning the holiday in Hawaii is extremely low, or both. The justification for sacrificing a lifetime supply of apples in favour of a small chance at a holiday in Hawaii is considerably weaker than sacrificing some number of apples delivered now for a 50/50 chance. Expected utility clearly has a part to play, even if the calculation of such utilities is non-trivial. The difficultly in estimating utilities, and the fact that utilities may depend upon unknown factors such as a person's financial situation, mean there is uncertainty in the evaluation of state utilities which our framework currently does not address.

There will be some computational complexity in searching and representing actions' potential branches of future development. In Section 4, we note Hansson's principles for optimising search to find the most problematic branches but it remains to be seen if this can be implemented as a practical strategy for keeping the planning problem tractable for common problems.

Nevertheless we believe the hypothetical retrospection framework provides a practical approach to handling many of the issues in machine ethics – particularly the handling of uncertainty and the lack of any real agreement on the best moral theory to adopt in applications.

Acknowledgements

We would like to thank the University of Manchester for funding and EPSRC, under project Computational Agent Responsibility (EP/W01081X/1).

Open Data Statement

 The tools/examples shown in this paper and instructions on reproducibility are openly available on GitHub at
https://github.com/sameysimon/HypotheticalRetrospectionMachine

References

1. Data protection. Ministry of Justice, https://www.gov.uk/data-protection
2. Alexander, L., Moore, M.: Deontological Ethics. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Winter 2021 edn. (2021)
3. Alhaddad, M.M.: Artificial intelligence in banking industry: A review on fraud detection, credit management, and document processing. ResearchBerg Review of Science and Technology **2**(3), 25–46 (2018)
4. Allen, C., Smit, I., Wallach, W.: Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and information technology **7**(3), 149–155 (2005)
5. Altham, J.E.: Ethics of risk. In: Proceedings of the Aristotelian Society. vol. 84, pp. 15–29. JSTOR (1983)
6. Arkin, R.C.: Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction. pp. 121–128 (2008)
7. Ashford, E., Mulgan, T.: Contractualism. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Summer 2018 edn. (2018)
8. Atkinson, K., Bench-Capon, T.: States, goals and values: Revisiting practical reasoning. Argument and Computation **7**(2-3), 135–154 (2016). https://doi.org/10.3233/aac-160011
9. Atkinson, K., Bench-Capon, T., McBurney, P.: Justifying practical reasoning. In: Proceedings of the fourth international workshop on computational models of natural argument (CMNA 2004). pp. 87–90 (2004)
10. Balasubramaniam, N., Kauppinen, M., Hiekkanen, K., Kujala, S.: Transparency and explainability of ai systems: ethical guidelines in practice. In: Requirements Engineering: Foundation for Software Quality: 28th International Working Conference (REFSQ). pp. 3–18. Springer (2022)
11. Bertsekas, D.P., Tsitsiklis, J.N.: An analysis of stochastic shortest path problems. Mathematics of Operations Research **16**(3), 580–595 (1991)
12. Białek, M., Neys, W.D.: Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. Judgment and Decision Making **12**(2), 148–167 (2017)
13. Brundage, M.: Limitations and risks of machine ethics. Journal of Experimental & Theoretical Artificial Intelligence **26**(3), 355–372 (2014)

14. Cimatti, A., Pistore, M., Roveri, M., Traverso, P.: Weak, strong, and strong cyclic planning via symbolic model checking. Artificial Intelligence **147**(1-2), 35–84 (2003)
15. Cloos, C.: The utilibot project: An autonomous mobile robot based on utilitarianism. AAAI Fall Symposium - Technical Report (01 2005)
16. Cointe, N., Bonnet, G., Boissier, O.: Ethical judgment of agents' behaviors in multi-agent systems. In: Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems. p. 1106–1114. AAMAS '16, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2016)
17. Davis, M.H.: Markov models and optimization. Routledge (2018)
18. Dennis, L., Fisher, M., Slavkovik, M., Webster, M.: Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems **77**, 1–14 (2016)
19. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence **77**(2), 321–357 (1995)
20. Ecoffet, A., Lehman, J.: Reinforcement learning under moral uncertainty. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 2926–2936. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/ecoffet21a.html
21. Hansen, E.A., Zilberstein, S.: Lao*: A heuristic search algorithm that finds solutions with loops. Artificial Intelligence **129**(1-2), 35–62 (2001)
22. Hansson, S.: The ethics of risk: Ethical analysis in an uncertain world. Springer (2013)
23. Hursthouse, R., Pettigrove, G.: Virtue Ethics. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Winter 2022 edn. (2022)
24. Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., Alazab, M.: A comprehensive survey for intelligent spam email detection. IEEE Access **7**, 168261–168295 (2019). https://doi.org/10.1109/ACCESS.2019.2954791
25. Kent, S.: Words of estimative probability. Studies in intelligence **8**(4), 49–65 (1964)
26. Killough, R., Bauters, K., McAreavey, K., Liu, W., Hong, J.: Risk-aware planning in bdi agents. In: International Conference on Agents and Artificial Intelligence. vol. 2, pp. 322–329. SciTePress (2016)
27. Lindner, F., Bentzen, M.M., Nebel, B.: The hera approach to morally competent robots. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6991–6997. IEEE (2017)
28. Ma, Y., Wang, Z., Yang, H., Yang, L.: Artificial intelligence applications in the development of autonomous vehicles: a survey. IEEE/CAA Journal of Automatica Sinica **7**(2), 315–329 (2020). https://doi.org/10.1109/JAS.2020.1003021
29. Mangan, J.T.: An historical analysis of the principle of double effect. Theological Studies **10**(1), 41–61 (1949)
30. Moor, J.H.: The Nature, Importance, and Difficulty of Machine Ethics, p. 13–20. Cambridge University Press (2011)
31. Mosdell, M.: Act-Consequentialism. Encyclopedia of Global Justice, Springer Netherlands pp. 2–2 (2011)
32. Saffiotti, A.: An ai view of the treatment of uncertainty. The Knowledge Engineering Review **2**(2), 75–97 (1987)
33. Sanner, S., et al.: Relational dynamic influence diagram language (rddl): Language description. Unpublished ms. Australian National University **32**, 27 (2010)

34. Sholla, S., Mir, R.N., Chishti, M.A.: A fuzzy logic-based method for incorporating ethics in the internet of things. International Journal of Ambient Computing and Intelligence (IJACI) **12**(3), 98–122 (2021)
35. Stephenson, T.A.: An introduction to bayesian network theory and usage. Tech. rep., Idiap (2000)
36. Szabadföldi, I.: Artificial intelligence in military application–opportunities and challenges. Land Forces Academy Review **26**(2), 157–165 (2021)
37. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in machine ethics: A survey. ACM Computing Surveys (CSUR) **53**(6), 1–38 (2020)
38. Younes, H.L.S., Littman, M.L.: Ppddl1.0: An extension to pddl for expressing planning domains with probabilistic effects. In: Technical Report –Carnegie Mellon University (2004)
39. Yu, K.H., Beam, A.L., Kohane, I.S.: Artificial intelligence in healthcare. Nature biomedical engineering **2**(10), 719–731 (2018)