Towards a Framework Combining Machine Ethics and Machine Explainability*

Kevin Baum
Universität des Saarlandes
Saarland Informatics Campus
Saarbrücken, Germany
Department of Computer Science
and Department of Philosophy

Holger Hermanns
Universität des Saarlandes
Saarland Informatics Campus
Saarbrücken, Germany
Department of Computer Science

Timo Speith
Universität des Saarlandes
Saarland Informatics Campus
Saarbrücken, Germany
Department of Computer Science
and Department of Philosophy

We find ourselves surrounded by a rapidly increasing number of autonomous and semi-autonomous systems. Two grand challenges arise from this development: Machine Ethics and Machine Explainability. Machine Ethics, on the one hand, is concerned with behavioral constraints for systems, so that morally acceptable, restricted behavior results; Machine Explainability, on the other hand, enables systems to explain their actions and argue for their decisions in a way that human users can understand and justifiably trust them.

In this paper, we try to motivate and work towards a framework combining Machine Ethics and Machine Explainability. Starting from a toy example, we detect various desiderata of such a framework and argue why they should and how they could be incorporated in autonomous systems. Our main idea is to apply a framework of formal argumentation theory both, for decision-making under ethically motivated constraints and for the task of generating useful explanations based on these constraints given only limited knowledge of the world. The result of our deliberations can be described as a first version of an ethically motivated, principle-governed framework combining Machine Ethics and Machine Explainability.

1 Introduction

(Semi-)Autonomous systems are pervading the world we live in. These systems start to deeply affect our lives and we, in turn, become more and more dependent on their operations. Several important questions arise: How should machines be constrained, such that they act morally acceptably towards humans? And how should conflicts between such constraints and the traditional means-ends based decision-making be resolved? These questions concern *Machine Ethics* – the search for practically implementable behavioral constraints for systems, enabling them to exhibit morally acceptable behavior. Although some researchers believe that implemented Machine Ethics is a sufficient precondition for humans to reasonably develop trust in autonomous systems, we want to discuss why this is not true in cases of imperfect information of the systems. We instead see the need to supplement Machine Ethics with means to enable *justified trust* in autonomous systems. We argue that there is at least one feasible supplement for Machine Ethics doing the job: *Machine Explainability* – the ability of an autonomous system to explain its actions and to argue for them in a way comprehensible for humans. In this paper we try to demonstrate how these two fields, Machine Ethics and Machine Explainability, are intertwined, and we propose the nucleus of a formal framework combining Machine Ethics and Machine Explainability. This work thereby goes beyond the rough thoughts and ideas we presented in [8].

*This work is supported by the ERC Advanced Grant 695614 (POWVER) and by the Initiative for Excellence of the German federal and state governments through funding for the Saarbrücken Graduate School of Computer Science and the DFG MMCI Cluster of Excellence.

B. Finkbeiner and S. Kleinberg: Third International Workshop
on Formal Reasoning about Causation, Responsibility,
and Explanations in Science and Technology (CREST 2018)
EPTCS 286, 2019, pp. 34–49, doi:10.4204/EPTCS.286.4

Related Work. Machine Ethics is becoming a serious research field, as first systematic works have been published in the last years (cf. [3, 38], see also [16] for a short overview of techniques and challenges). As James H. Moor pointed out (cf. [34]), Machine Ethics can be understood as a rather broad term, ranging from ethically motivated restrictions on the behavior of complex and possibly autonomous systems to the implementation of full-fledged moral capacities, involving deep, philosophical concepts of autonomy and deliberation, as well as free will. Following the systematizing ideas from [38] regarding different degrees of moral artificial agents, we think that for now Machine Ethics should not aim directly for *true ethical* decision-making. The most pressing task of Machine Ethics is rather to find a way of describing and implementing *ethically constrained instrumental* decision-making.[1] This should allow for principle-based, unambiguous and formal guarantees that restrict the autonomous system's behavior in a way that makes the system *significantly morally better*, without necessarily implementing any moral theory and still allow it to do what it was designed for. Hence, the goal is an *overall morally acceptable and desirable* system that remains useful.

In contrast to Machine Ethics, Machine Explainability aims at equipping complex and autonomous systems with means to make their decisions *understandable* to different groups of addressees (cf. [1, 9, 25, 26, 31]) enabling a sufficient amount of transparency and perspicuity for these systems. Doing so becomes more and more urgent: For instance, the software doping cases that surfaced in the context of the VW diesel emissions scandals made obvious that the behavior of complex systems can be very hard – if not practically impossible – to comprehend even for experts (cf. [6, 7, 19]). These cases make clear that explainability plays a crucial role in regard to trust and whenever it comes to accountability questions where one needs to tell apart intentional misconduct from genuine malfunction. Especially in context of autonomous systems – which often promise positive, societal effects –, black box systems for which nobody is able to explain their decisions, predictions or behavior will plausibly lack trust in the long run. Also, many applications of computational intelligence systems – for instance as advisors of politicians and judges – presuppose more than naked numbers and probabilities, at least in context of liberal democracies. They need to be scrutinizable and their outputs must be justifiable at least in principle and upon request. Thus, even under the premise that the deployment of some systems is desirable from a moral point of view (thanks to their overall effects) and even if these systems would in fact behave as morally good as logically and conceptually possible (thanks to future advances in Machine Ethics): As long as people cannot justifiedly trust the systems and cannot understand the reasons for their decision, their implementations are threatened even where desirable, and they cannot be promoted with good conscience in many areas of potentially promising application. However, Machine Explainability still is a young field and especially formal frameworks supporting explanations are rare (cf. [12] for a simple one). With this paper, we take first steps towards a method to perform ethically constrained decision making – Machine Ethics – in a way that in itself grounds the very possibility of Machine Explainability.

2 Developing a General Framework

In the last decades and especially in recent years, researchers have made enormous progress in the development of autonomous systems. The knowledge and the tools to create artificial agents in the sense of autonomous problem solvers and good *instrumental* autonomous decisions-makers are broadly available. These agents are instrumental decision-makers insofar, as they decide *instrumentally rationally* (cf. [30]). The goal of Machine Ethics is to extend these methods such that the resulting agents not only solve their problems instrumentally well, but also in a morally acceptable way. That being said, it has to be admitted that, as of yet, there is not much foundational research pertaining to those topics available. For instance,

[1] Instrumental decision-making is means-ends oriented decision-making that tries to find the right means to achieve specific ends, where "right" here means as much as being most efficient or cost-effective. Philosophically speaking, it is a kind of *instrumental rationality*.

we lack the possibility to spell out problems in Machine Ethics in a formal way. We especially lack a formally precise, fruitful and unambiguous way to state moral constraints and principles. As one aspect of this paper we try to undertake steps to change that. Since we are still at the beginning of this admittedly ambitious research project, we impose some restrictions for now. For instance, while we do not need to assume a deterministic evolution of the environment, we resort to a probabilistic interpretation, for instance derived from past statistical evidence. Thus, our methods are made with aleatoric uncertainty in mind. We leave the question of how to handle cases involving *epistemic* uncertainty to future research.

2.1 The World of a Medical Care Robot

We start our discussion by describing a toy example that serves as an example context for motivating a formal framework and bringing it to life. The example is deliberately kept simple, but sufficiently complex and general to exemplify the challenges arising with respect to Machine Ethics.

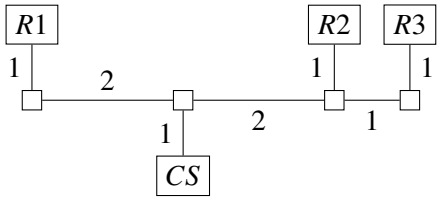

Figure 1: The medical care robot's realm

We consider a medical care robot working in a hospital. There are up to three patients the robot has to take care of. Each of these patients is in a separate room ($R1$, $R2$, $R3$), and the rooms are connected by several hallways. The spatial layout of the scenario is depicted in Fig. 1. The robot spends energy when traveling along a hallway. The energy costs depend on the distance traveled (distances are written next to the hallways). For one unit of distance, the robot needs one unit of energy. At some point the robot's battery will be depleted. To prevent this, there is a charging station (CS) where the robot can recharge its battery. Once the recharging process is started, it will not stop before the battery is full again. The robot 'knows' its current position and its energy level.

In our scenario, the robot listens to requests. At each point in time, each of the three patients may issue a request to the robot, asking for a task of a specific priority. Although each request has a priority when issued, this priority is deliberately *not transmitted* to the robot. Instead, the robot is assumed to know appropriate and justified probability distributions regarding the tasks associated with the requests. This is necessary, as otherwise the patients could get tempted to always issue tasks of the highest priority in order to get preferential treatment. We further assume that there is only a limited number of possible tasks which can be concealed by a request. Such tasks can range from simply fetching water to doing a reanimation. In the following we use the example to highlight a couple of central points.

2.2 Towards a General Framework

In this section, we work towards a *general* framework in which autonomous systems, including the above, can be described. We then extend it to be applicable to Machine Ethics in the next section.

World States and (Partial) Knowledge. We assume the autonomous system's world to be fully specifiable by assignments of a finite number of variables. Thus, a *state of the world* (short: ω) is represented by a tuple of variables $\omega := \langle \omega_1, \ldots, \omega_n \rangle$ with corresponding domains $D_1, \ldots, D_n$. We call the set of all possible world-states $\Omega \subseteq \bigtimes_{i=1}^{n} D_i$, and we let $|\omega|$ denote the number of elements in a tuple ω.

At each state, the system knows some, but not necessarily all, facts about the current world state. In fact, the set of known variables may vary from state to state. For instance, our robot may know

the brightness and temperature in some room if and only if it is in this room. Thus, a subset of a complete world state represents the variables the system has knowledge of. For the sake of readability, we conveniently assume the first k ($1 \leq k \leq n, n = |\omega|$) variables $\langle \omega_1, \ldots, \omega_k \rangle$ of a state of the world to be known variables at some given moment. We call the set of these (in general) partial world states $\Theta \subseteq \bigtimes_{i=1}^{k} D_i$. We do not rule out that the variables spanning Θ are dependent and, thus, the strict containment $\subsetneq$ to hold. Furthermore, we define the possible world states in the light of some knowledge $\theta \in \Theta$ as $\Omega_\theta := \{\omega \in \Omega \mid \omega_{1:k} = \theta\}$, where $\omega_{i:j}$ for $1 \leq i < j \leq n$ denotes the subtuple $\langle \omega_i, \ldots, \omega_j \rangle$ of some tuple $\omega = \langle \omega_1, \ldots, \omega_i, \ldots, \omega_j, \ldots, \omega_n \rangle$.

Even if the strict inequality $k < n$ holds (as it does in most cases), we assume the system to be not clueless with regards to the remaining variables $\langle \omega_{k+1}, \ldots, \omega_n \rangle$. The system, hence, has justified credence regarding these variables' assignments, representable as probability distributions over the variables' domains P_i: $D_i \to [0,1]$, $k < i \leq n$. The probability of an unknown variable having a specific value might very well be dependent on the values of known variables. So, for some $\omega_i \in D_i$ with $k < i \leq n$ and some assignments of the known variables $\theta \in \Theta$, we have $P_i(\omega_i) \neq P_i(\omega_i|\theta)$. Call the set of all these distributions Π. Finally, we can assign the overall credences of the system concerning a specific configuration $\omega \in \Omega$ as $P(\omega) = \prod_{i=1}^{n} P_i(\omega_i)$, where ω_i is the ith variable in ω. As the system normally has some knowledge $\theta \in \Theta$, we get for every $\omega \in \Omega_\theta$ the probability $P(\omega) = P(\omega|\theta) = \prod_{i=k+1}^{n} P_i(\omega_i|\theta)$.

Example 1 *For our robot, a world state would consist of variables encoding daytime, position, energy level, energy costs for traveling, the requests in its queue and the tasks associated with it as well as some more. Since by design our robot is quite omniscient, the knowledge subset of these world states would contain everything except for the tasks associated to the requests.*

Options and Actions. At each decision state – that is, in a state where the system given its context has to decide something, normally after performing an operation or triggered by some incoming event –, the system has to choose from a number of available operations. Call the operations the *options* and let $\Phi = \{\phi_1, \ldots, \phi_n\}$ be the set of them. The operations available to the system will normally depend on the current state of the world, but for the sake of simplicity we do not elaborate on this dynamic here any further. Instead, we assume them to be constant over Ω. An *action* is an instantiated (i.e., chosen and performed) option and thus, by assumption, the observable decision the system has made.

Example 2 *In case of our robot example, there are just two possible options: it can either serve the request (AnsReq) or recharge (Charge).*

Goal(s), Outcomes and Instrumental Decision-Making. We simply assume here that any system under consideration not only has at least one unambiguously defined goal, but also a method dec_{inst}^{Π} of deciding for the best means to achieve this goal (given the system's knowledge and a set of candidate options). Traditionally, the goal is to find an action that maximizes some kind of expected utility. This utility incorporates both, the uncertainty of the action's outcomes in the light of the world state's uncertainty (and possibly even some indeterminacy in the world's 'reaction' to some action) as well as rewards and penalties associated with the possible outcomes.

Here is, for instance, how such an instrumental decision can be made, if we model the issue as a Markov decision problem. We assume that there is a function that, given the current world state, assigns to some action and another world state (i.e., a candidate for an *outcome* of the action given the current world state and the action) the probability of that outcome. Formally this can be specified as $Outcome : \Omega \times \Phi \times \Omega \to [0,1]$. Using the probability distributions on unknown variables (given some knowledge state) from above, we can straight-forwardly derive a function $Outcome_\Pi : \Theta \times \Phi \times \Omega \to [0,1]$

$$Outcome_\Pi(\theta, \phi, \omega) = \sum_{\omega' \in \Omega_\theta} P(\omega'|\theta) \cdot Outcome(\omega', \phi, \omega)$$

that operates on partial world states (i.e., the system's knowledge at some point). Further, let us suppose a utility function $U : \Omega \to \mathbb{R}$, specifying rewards and penalties as incentives for or against a specific

behavior. This allows us to reformulate the need to achieve our goal as maximization of utility. We thus arrive at a well-posed Markov decision problem. Given some partial state of the world $\theta \in \Theta$ representing the system's knowledge and a set of currently available options Φ, following the standard approach to Markov decision problems, dec^{Π}_{inst} comes down to:

$$dec^{\Pi}_{inst}(\Phi, \theta) = \operatorname*{argmax}_{\phi \in \Phi} EU(\phi|\theta) =: Choice^{\Pi}(\Phi, \theta), \text{ where}$$

$$EU(\phi|\theta) := \sum_{\omega \in \Omega} Outcome_{\Pi}(\theta, \phi, \omega) \cdot U(\omega)$$

is the expected utility of an option $\phi \in \Phi$ given the system's knowledge θ. Note that $dec^{\Pi}_{inst}(\Phi, \omega) \subset \Phi$.

Up to this point, we described a rather general class of decision problems that, as we showed, can be solved by methods associated with Markov decision problems. In the current state of our framework, we can thus, by adjusting the utilities in distinct ways, support or effectively even enforce specific decisions.

Example 3 *Instantiating these considerations to the case of our medical care robot it seems plausible to assign rewards to the fulfilling of tasks and penalties to running out of power. Then, an execution plan which serves the most (and best rewarded) requests over the longest time possible is what we are aiming for. By setting the rewards of rescuing a person (through reanimation) higher than the penalty of exhausting the battery, we would get the result of human lives being more important than robots operating.*

*But is this kind of 'tweakable' instrumental decision-making sufficient to get a satisfying way of making decisions in case of our robot? We believe not, because one can construct situations, in which maximizing the overall utility (for any assignment of utilities) plausibly is not what the robot ought to do. Assume, for instance, our robot is in room R*1 *and has to decide to either perform a reanimation there or to go back to the charging station. Let's assume further that the robot has enough power to reanimate, but then cannot make it back to the charging station afterwards. Assume now that with a high enough certainty, other high priority tasks – say even other reanimations – need to be performed later on. If our robot performs the reanimation now, it is not able to perform the other reanimations later. We can easily construct such a case in a way that makes the expected utility of charging higher than the expected utility of performing the current reanimation task.*

At least some ethicists would agree that the robot ought not to recharge now, nevertheless. It should give preference to rescuing the life at issue at the moment of decision. But even an ethicist that does not agree with this, would likely subscribe to the claim that a robot should not be constructed in the go-recharge-way, first and foremost because of the question of trust*: Imagine that in such cases the robot would be witnessed to turn around and leave toward its charging station. People would not develop trust in that robot – but it is important for people to trust in autonomous systems, as we already made plausible in the beginning. So, let us suppose that, overall, the robot ought not to weight lives that way.*

Instead, explicit and unambiguous constraints are needed that rule out some decisions and enforce others. Thus, Machine Ethics is a valid research field. This is the first central point we want to make.

3 A Framework of Machine Ethics

We think that we need substantially more than just instrumental decision-making as it is modeled up to now. In this section, we will give good indicators for this being the case and propose what we exactly need more.

Assume that Machine Ethics amounted to simply adjusting the utilities and disutilities in such a way that the induced behavior entirely adheres to a, say, consequentialist picture of morality,[2] we apparently could integrate this in an instrumental decision-making procedure as sketched above. Given a

[2]Consequentialist theories are normative theories – theories about the moral permissibility of actions – that solely focus on

full-fledged artificial system that is meant to qualify as a moral agent, and adopting such a picture of morality, adjusting the utilities, then, might very well be everything there is when it comes to implementing Machine Ethics, as such a system would take into account the effects its decisions make with regard to the question of trust as well.

However, neither does an autonomous system qualify as a full-fledged moral agent, nor is a consequentialist picture of morality common sense.[3]

Therefore, the decision-making ought to be guided and *restricted by explicit social and moral norms*. We, the people, want hard guarantees, forbidden actions and other desirable properties *a priori*. This is what is needed for achieving the goal of Machine Ethics. This motivates the essential building block of our approach to a framework of Machine Ethics: Moral Principles.

Moral Principles. Assume it has been decided that the decision-making process of the system ought to be constrained by a number of carefully chosen, ethically motivated and ethically justified *principles* $\Psi = \{\psi_1, \ldots, \psi_m\}$. Their concrete semantic interpretation will be discussed later. However, we already note that these principles are – in line with how principles are often understood within moral philosophy (cf. [13]) – meant to be *objective* in the sense that they are principles about which action ought (not) to be done with regards to certain states of the world. So, the question what ought to be done is determined by these principles irrespectively of the agent's information. We take this to be the most natural way to frame the core problem of Machine Ethics, because the behavioral restrictions one will need to implement will be of this kind, too. The restrictions themselves will be given by social or moral norms as well as by legislation, independent from the concrete design decisions and restrictions the system has.

In order to express a hierarchy of the principles, we define an order on Ψ in two steps. First, an equivalence relation $\approx_\Psi$ on Ψ is assumed which induces t equivalence classes $\Psi_1, \ldots, \Psi_t$, such that for $1 \leq i \leq t : \psi, \psi' \in \Psi_i : \psi \approx_\Psi \psi'$. For any arbitrary principle $\psi_i \in \Psi$, the class $[\psi_i]$ refers to the equivalence class of ψ_i. Second, we assume a strict total order $\succ_\Psi$ on these equivalence classes. This order is extended to the level of principles, such that for $\psi, \psi' \in \Psi : [\psi] \succ_\Psi [\psi'] \rightarrow \psi \succ_\Psi \psi'$, so as to define an overall (non-strict) weak order $\succ_\Psi \cup \approx_\Psi \;=:\; \succeq_\Psi \subset \Psi \times \Psi$, thus a total preorder on the principle set Ψ. We call $\mathfrak{P} := \langle \Psi, \succeq_\Psi \rangle$ a *principle structure*, giving us a hierarchy of moral principles.

Up to this point, we did not say anything about the principles' inner structure and about their content. We suggest to think of principles, in general, as functions $\psi : \Omega \rightarrow \mathcal{O}(\Phi)$ from a possible world state and the corresponding set of available options (which in context of this paper is assumed to be constantin Ω) to a subset of these options, the set of *permissible options* $Perm^{\psi}_{\Phi,\omega} \subseteq \Phi$. Either such a principle does shrink – or as we say 'filter' – the set of available options, then $Perm^{\omega}_{\psi} \neq \Phi$, or it does not, and thus $Perm^{\psi}_{\Phi,\omega} = \Phi$. We say that ψ has grip if and only if $Perm^{\psi}_{\Phi,\omega} \neq \Phi$. The set of worlds $\omega \in \Omega$ in which ψ has grip, namely $\{\omega \in \Omega \mid Perm^{\psi}_{\Phi,\omega} \neq \Phi\} \in \mathcal{O}(\Omega)$, can be understood as predicate c_ψ and we write $\omega \models c_\psi$ to express that $\omega \in c_\psi$. Correspondingly, in this paper we represent principles as conditionals[4] of the form $(c_\psi \rightarrow o_\psi)$ where each o_ψ is an *option structure* $\langle \Phi, \succeq_\Phi \rangle$, defined in complete analogy to the principle structures just introduced, but over the action space Φ. Given that representation of principles, we owe the reader how to determine $Perm^{\psi}_{\Phi,\omega}$ from these option structures.

Given that some $\omega \models c$ (for some $\omega \in \Omega$ and some principle $(c \rightarrow o) \in \Psi$), this induces a (non-strict weak) *permissibility order* $\succeq^{\psi}_{\Phi,\omega}$, a total preorder on the option set Φ. We refer to the topmost

the actions' outcomes. The consequentialist picture is driven by the idea of maximizing (moral) value and that what has value (and disvalue) are states of affairs. Hence, what makes an action right (or wrong) is what the action changes (or promise to change) in the world. The classical source of consequentialism can be found in [10] and a systematic discussion of it in [11].

[3]The non-consequentialist competitors in the realm of normative ethics are the families of deontological theories (cf. [28, 36]) and virtue theories (cf. [4, 5, 20, 37]). Philippa Foot prominently emphasized the tension between consequentialism and common sense (in [20]). For a recent consequentialist approach to avoid such clashes, see [35].

[4]We want to emphasize that we do not mean to imply that principles in fact (whatever that means) need to adhere to such a structure. We just say that for the purposes at hand principles might be *modeled* as conditionals. We leave more sophisticated models for future research.

class $[\phi]$ of this order (in the sense that $\forall\hat{\phi} \in [\phi] : \hat{\phi} \succeq^{\psi}_{\Phi,\omega} \phi'$ for all $\phi' \in \Phi$) as $Perm^{\psi}(\Phi,\omega) \subseteq \Phi$, the set of *permissible options* relative to principle ψ given some world state ω. The intention behind this construction is that the action to perform according to principle ψ in state ω needs to be picked from $Perm^{\psi}(\Phi,\omega)$, the set highest in the permissibility order associated with that principle. Naturally, if multiple principles apply for a given state of the world (i.e., given some world state, more than one antecedent is fulfilled), the one highest up in the principle structure $\langle\Psi,\succeq_{\Psi}\rangle$ is deemed decisive. But this does not exclude the probability that different principles of the same (topmost) equivalence class apply.

Given an arbitrary set of principles $\hat{\Psi} \subseteq \Psi$ of principles and an arbitrary world state $\omega \in \Omega$, we refer to the subset of principles which apply in this world state $\{(c \to o) \in \hat{\Psi} \mid \omega \models c\}$ as $\hat{\Psi}^{\omega}$ and call $\hat{\Psi}^{\omega}_{\max}$ the set of the maximally ranked principles in $\hat{\Psi}^{\omega}$ according to $\succeq_{\Psi}$. It seems right to identify the set of the overall permissible options $Perm^{\hat{\Psi}^{\omega}_{\max}}_{\Phi,\omega}$ with the intersection of the permissible options according to the principles $\psi \in \hat{\Psi}^{\omega}_{\max}$, as this will result in the set of options permissible according to all these maximally ranked principles. Thus

$$Perm^{\hat{\Psi}^{\omega}_{\max}}_{\Phi,\omega} := \bigcap_{\psi \in Perm^{\hat{\Psi}^{\omega}_{\max}}_{\Phi,\omega}} Perm^{\psi}_{\Phi,\omega}.$$

In general, then, we can identify the set of overall permissible options in light of some principle structure $\mathfrak{P}$, some world state $\omega \in \Omega$ and some corresponding options space Φ as $Perm^{\mathfrak{P}}(\phi,\omega) := Perm^{\Psi^{\omega}_{\max}}_{\Phi,\omega}$.

The question arises, whether this intersection will be guaranteed to be non-empty. The answer obviously depends on the properties we require for principles in the same equivalence class. If we require our system both, to only perform permissible options in the above defined sense and to never stop operating (and thus fulfill liveness), then we should require principles in the equivalence classes to be such that this intersection is never empty.

This finding, however, echoes that sometimes there seemingly are principles which would result in what philosophers call *true moral dilemmas*: situations where no option is permissible at all. Not all moral theories allow for such dilemmas to exist (for instance, consequentialist theories normally do not allow them). While we will not rule out that an unsatisfiable (relative to some $\omega \in \Omega$) subset of principles of the same equivalence class may be a subset of a valid set of principles for the purposes of Machine Ethics, we will disregard the question what to do in such a situation.[5] We leave an axiomatization of this requirement for future research.

All this – the principle structure and the method of finding the permissibility relation on the action – coalesce into a function we call *deontic filter*. Here, "deontic" indicates that something is about what *ought* to be the case according to some standard or norm, such as a social or moral norm. The deontic filter thus is

$$dec^{\mathfrak{P}}_{filter}(\Phi,\omega) = Perm^{\mathfrak{P}}(\Phi,\omega).$$

Example 4 *We now apply this part of our framework to our robot example. As mentioned before,* $\Phi :=$ *{AnsReq, Charge}. Below is one plausible way the robot's deontic filter could look like. Since we want to use the example just to give a vivid picture of how things would look, we spare applying the whole formalism here, but for a little more detail, see [8].*

$$dec_{filter}(\omega,\Phi) = \begin{cases} \{AnsReq\} & \text{, if the priority of the task associated to the request is high and the current energy level would suffice to serve it;} \\ \{Charge\} & \text{, if the priority of the task associated to the request is low and the current energy level would not suffice to serve it and then to go back to the charging station;} \\ \{AnsReq, Charge\} & \text{, otherwise.} \end{cases}$$

[5] Another, more permissive, way to define the set of permissible options would be to set $Perm^{\mathfrak{P}}(\Phi,\omega)$ as the union of permissible sets regarding the principles in the highest ranked, non-empty equivalence class that have grip in ω.

In order to determine which options are permissible in the light of these principles, the knowledge of the current world state is presupposed. For instance, our robot needs to know the associated task to the current request. But there were good reasons for not equipping the robot with that knowledge.

Notably, deontic filtering presupposes perfect knowledge about the current state of the world. In full generality, this might still not be enough, since the evaluation of dec_{filter} might even presuppose knowledge about de facto outcomes. So, what to make out of this result? First, as we shall discuss in the sequel, given the necessary (but practically impossible) kind of full world knowledge (perfect information), the task of Machine Ethics becomes quite simple. Second, in the practically much more interesting case of incomplete knowledge (imperfect information), we must be prepared for morally suboptimal behavior (as we have already argued in [8]), but also we cannot straightforwardly implement the notations introduced in this section.

Before we turn to the question of how to incorporate uncertainty in the context of deontic filtering (in Section 4), we first turn to the idealized case, presenting an easy, sequential way to incorporate deontic filtering into the overall decision process in case of perfect knowledge.

3.1 An Idealized Overall Decision Pipeline

We believe that the ingredients characterized above are all we need to specify a well-posed problem of Machine Ethics. But how do we then solve such a problem operationally? There are two possibilities. Either, one interlocks the two decision modules $dec^{\mathfrak{P}}_{filter}$ and dec^{Π}_{inst} into one. Or, one applies $dec^{\mathfrak{P}}_{filter}$ before dec^{Π}_{inst}. More specifically, dec^{Π}_{inst} must be applied to the options *surviving* the deontic filtering. Given that the deciding system has all it needs to evaluate $dec^{\mathfrak{P}}_{filter}$, we believe the latter is more natural and easier to achieve. One just has to concatenate the $dec^{\mathfrak{P}}_{filter}$ and dec^{Π}_{inst} into one larger decision procedure *dec*, such that for each decision $Perm^{\mathfrak{P}}(\Phi,\omega) \subset \Phi$ becomes the foundation of $Choice^{\Pi}(\Phi,\theta)$, rather than the full set of options Φ. In such a sequential picture, if one wants to ensure *liveness* of the system, then true moral dilemmas must be ruled out as aforementioned.

Such an overall decision pipeline would consist of the following steps: *deontic filtering*, *instrumental decision-making* (as already introduced) and, finally and trivially, *picking*: just picking a random element out of the options 'surviving' the first two steps. The whole decision pipeline *dec* is visualized as flow diagram (in Fig. 2).

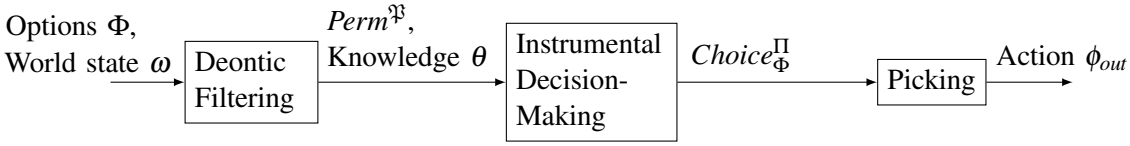

Figure 2: The sequential decision pipeline *dec*

However, for realistic applications one needs a $dec^{\mathfrak{P},\Pi}_{filter}$ version of the deontic filter that operates with the de facto knowledge of the system, as dec^{Π}_{inst} does. As pointed out above, the principles encoded in the deontic filter are – for good reasons – objective. They only can be applied to determine the set of permissible options if one has perfect knowledge. That is, one needs full information on the complete state of the world. In other words, what we have is $dec^{\mathfrak{P}}_{filter}(\Phi,\omega)$ and what we need is $dec^{\mathfrak{P},\Pi}_{filter}(\Phi,\theta)$. The framework, as it stands up to this point, thus can be understood only as a partially idealized version of what we finally long for. We turn to our proposal to tackle this issue in the next section.

4 Incoporating Uncertainty & Enabling Machine Explainability

4.1 Arguments as Enablers

How do we incorporate uncertainty in the deontic filter function? We want to present an approach that does not only solve that problem, but also enables Machine Explainability: We envision explanations as byproduct of an argument-based decision-making process.

We can easily think of the possible cases as ground for argumentation: If this or that case was true, it would give me, thanks to this or that principle, a reason for the one action rather than for the other. The probability of the cases together with the importance of the principle in question determines the strength of the resulting reason. We think that arguments can be understood as encoded reasons. And, as already proposed by Benjamin Franklin (cf. [21]), a decision-making process (in the sense of everyday understanding of the term) can be naturally interpreted as the weighing of reasons in order to determine the right action or decision. In other words, decision-making can be thought of as an internal argumentation with pro and contra arguments for (or against) the decision or action. Furthermore, the kind of reasoning involved in our everyday decision-making seems to be non-monotonic – further information or evidence may require the systems to retract from its decision – and arguments are *the* tool for non-monotonic reasoning as pointed out by Dung (cf. [18]).[6]

But can arguments be the right kinds of explanations? After all, there are many kinds of explanations: scientific explanations in the form of deductive-nomological models (cf. [23]), causal explanations that relate causes with their effects, psychological explanations and many more. What we are looking for primarily are explanations that are, in terms of Davidson ([15]), *rationalizations*. These rationalizations are meant to make available to us the reasons of why the explained system decided and/or acted the way it did. Interestingly, our approach can also be seen as some kind of inductive-statistical explanation as proposed by C.G. Hempel (cf. [24]). Given probabilities of a particular world state, we are able to derive the relative probability of each option being chosen. With this we are able to derive the robot's behavior both before and after it is witnessed. This is an important property of the kind of explanations our approach provides: a person can – independently from the robot – arrive at the same result.

If the system comes to its decisions based on an internal, argumentative process, the decision-making can be made transparent and rationalized in exactly the way explainability longs for. And if the argumentation-based decision-making models idealize deliberation using traditional human concepts, the obtained explanations can be expected to be *comprehensible* explanations (to put it into the terms of [9]: we have *graspable* explanations). We, thus, suggest to resort to an *argumentation-based* approach, since we believe that such an approach can not only solve the problem of finding principle-based decisions under uncertainty, but also allows to generate explanations for the resulting decisions as a byproduct.

So, provided argument-based reasoning is an appropriate approach to decision-making in the context of Machine Ethics, and arguments are the right kind of structure to encode explanations, adopting a framework of formal argumentation theory is the obvious choice for modeling and implementing these issues.[7] Machine Explainability, then, is a *byproduct* of artificial moral decision making, since the explanations are (or are extractable from) the argumentation graphs that represent what led to a decision.

[6] A sophisticated framework in the context of decision-making and explanations can be found in [2]

[7] What if our robot's decision-making component is a black box, for instance, because dec^{Π}_{inst} results from some kind of machine-learning approach? Is achieving explainability hopeless then? We think that this is not necessarily so: in principle, the argumentation graphs could be derived in hindsight and maybe even "from the outside" (e.g. by some process as sketched in [9]). This might come with the problem of our justifications being *post hoc* rationalizations and, thus, not reflecting the *true* reasons or reasoning (i.e. one needs to guarantee what [9] calls *accuracy*). We leave solving this problem for future research. An interesting starting point, however, could be attempts of falsification starting from inductive-statistical prognoses (cf. [29]).

4.2 Generating an Argumentation Graph

We propose a *three step approach* for generating the argumentation graph and introduce this approach in this section. But before doing so we remark two issues: First, we shall leave out for now the question of *how* to generate explanations from the argumentation graph. It will become obvious, we believe, that the (generic) arguments we describe in this section are a solid ground for spelling out explanations. Second, we suggest a procedure that is – contrary to the above sketched approach to a framework with *idealized* deontic filtering – *not* a sequential application of dec_{filter} and dec^{Π}_{inst}. Instead we propose a decision procedure that interlocks the two decision modules. We elaborate on that point towards the end of this section.

We return to the three steps and the overall graph generation. They are (in their corresponding order): *Case Distinction*, *Reason Aggregation* and *Final Action Determination*. Building upon the ingredients from the above framework – especially $\Omega, \Phi, \Pi, EU, \mathfrak{P}$ –, these three steps result in a *bare argumentation graph* $\Gamma \coloneqq \langle V, E \rangle$ with $V \coloneqq V_1 \cup V_2 \cup V_3, E \coloneqq E_{1,2} \cup E_{2,3}$ and a couple of weight-functions

$$relevance^{\mathfrak{P}} : V_1 \to \mathbb{R}^+, force_{\text{protanto}} : E_{1,2} \to \mathbb{R}^+ \text{ and } force_{\text{overall}} : E_{2,3} \to \mathbb{R}^+.$$

The resulting structure $\mathfrak{G} : \langle \Gamma, relevance^{\mathfrak{P}}, force_{\text{protanto}}, force_{\text{overall}} \rangle$ then is the *complete argumentation graph*. We postpone motivating the three weight functions to the point when they are defined, but shortly introduce the elements of Γ now. V_i are the vertices (which are or contain arguments) generated in step i, E is the set of edges representing the *influences* of arguments from earlier to later steps – thus $E_{1,2} \subset V_1 \times V_2$ and $E_{2,3} \coloneqq V_2 \times V_3$.

Prior to going through the generation process step by step, we note two things: First, the whole process has to be performed for each decision that is to be made and, thus, $\mathfrak{G}$ is a function of the current knowledge θ and the set of available options Φ. Second, the process results not only in a graph, but also in a decision for a particular option. The graph is meant to be stored for the purpose of being accessible later in order to explain (or enable to explain) the corresponding decision.

Step 1: Case Distinction. As we basically aim to model something like a human agent's inner deliberative process of pondering on what she ought to do, we suggest that the systems need to take into account all possible cases $\omega \in \Omega$ that respect the systems world-knowledge $\theta \in \Theta$. Recall that $\Omega_\theta \subset \Omega$ is the set that does so. In general, where each world state is defined by n variables and θ contains, by construction, the first k variables of each world state ω, we need to take into account $|\Omega_\theta| \leq \prod_{i=k+1}^{n} |D_i|$ different world states. In order to deal with the existing uncertainty, the occurrence probabilities of the cases $\omega \in \Omega_\theta$ must be taken into account. Additionally, various cases will plausibly need to be considered not only once, but with regards to every principle $\psi \in \Psi$ that applies to that case. We thus need an argument Arg^{ω}_{ψ} for every $\psi \in \Psi^{\omega}$ of every $\omega \in \Omega_\theta$. In sum, this makes $\prod_{\omega \in \Omega_\theta} |\Psi^{\omega}|$ arguments for V_1. In this sense, not only the case's probability has to be considered, but also the *relevance* of these applying principles, which correlates with $\succ_{\Psi}$.

First, to the form and content of the arguments. Each of these arguments consists of three premises, logically linked by two concatenated modus ponens applications. The first premise, P_ω, states plainly that $\omega \in \Omega_\theta$ is the case; the second premise, P_ψ, consist just of ψ i.e., of some $(c \to o)$; the third and final premise, P_{Perm_o}, determines the set of permissible options according to o. By construction of the argument, obviously $\omega \models c$, since $\psi \in \Psi^{\omega}$. Table 1 shows the general form and generic content of this first layer's arguments that we now can define explicitly:

$$V_1 \coloneqq \{Arg^{\omega}_{\psi} \mid \omega \in \Omega_\theta \wedge \psi \in \Psi^{\omega}\}.$$

The arguments in V_1 only result in sets $Perm^{\psi}(\Phi, \omega)$, but performing the full deontic filtering is supposed to give us $Perm^{\mathfrak{P}}(\Phi, \omega)$. According to our idealized framework above, $Perm^{\mathfrak{P}}(\Phi, \omega)$ is the intersection of all sets $Perm^{\psi_i}(\Phi, \omega)$ of the principles $\psi_i \in \Psi_{\max}$ (i.e., the maximally ranked equivalence class with a

principle that applies for ω). But this does not hold under uncertainty, because the *de facto obtaining* ω is unknown. For this reason we have done the case distinction in the first place. We suggest to incorporate the uncertainty in a quantitative aggregation method, respecting the *probability* $P(\omega|\theta)$ of the case (the ω) under consideration (given the system's knowledge θ) and the importance or relevance $relevance^{\mathfrak{P}}$ of the corresponding applicable principle $\psi \in \Psi^{\omega}$, correlating with $\succeq_{\Psi}$ given in $\mathfrak{P}$. But how should the relevance relate to $\succeq_{\Psi}$ exactly?

Obviously, the relevance should reflect the priority ranking $\succ_{\Psi}$ over the equivalence classes $\Psi_1, \ldots, \Psi_t$ induced by $\approx_{\Psi}$. Hence, $relevance^{\mathfrak{P}}$ should be *monotone* relative to $\succeq_{\Psi}$: For all $\psi, \psi' \in [\ \psi\ ]$ it holds that $relevance^{\mathfrak{P}}(\psi) = relevance^{\mathfrak{P}}(\psi')$ and for all ψ, ψ' with $[\ \psi\ ] \succ_{\Psi} [\ \psi'\ ]$ it holds that $relevance^{\mathfrak{P}}(\psi) > relevance^{\mathfrak{P}}(\psi')$.

However, that leaves a lot of further properties of $relevance^{\mathfrak{P}}$ unspecified. Depending on the specific properties of $relevance^{\mathfrak{P}}$ different design decisions of the reasoning process could be made. For instance, assume we want to allow that a sufficiently large number of lower ranked principles that are fulfilled (with some probability smaller than 1) can *outweigh* a few higher ranked principles that are fulfilled (with the same or a lower probability). That is, we want $relevance^{\mathfrak{P}}$ to be *archimedean*.

In other contexts than health care, reasoning that allows for, e.g., defeater reasons, – roughly: reasons that silence other reasons, canceling out their strengths – may be needed. Also, one could want to follow some kind of precautionary principle sometimes. Then a tiny chance of a principle to apply might be already enough for the system to be morally required to decide in line with this principle, no matter how improbable the case is (cf. [22] for a discussion of advantages and disadvantages of doing so). If in such cases one still wants to work with weights, the weights of specific principles might need to be infinite, so that the underlying principle is enforced. Basically, this would allow then again for *true moral dilemmas* as mentioned above in context of perfect knowledge scenarios.

So, if we would disallow any such weighing between principle fulfillments of principles in different equivalence classes, then we would want $relevance^{\mathfrak{P}}$ to map ψ into sets *closed* under scalar multiplication (and define an order over these sets in accordance with $\succ_{\Psi}$).

Independent of how we design $relevance^{\mathfrak{P}}$ concretely, it is later assigned to each argument.

Example 5 *Turning one last time to our example scenario (the remaining steps are completely generic), this step is instantiated by constructing arguments for any possible task that might be concealed by a request (since there are no other unknowns in this case). Given the original* dec_{filter} *method from section 4, the robot knows what it is permitted to do in each possible case under consideration. Together with its probability estimates for each such case and in the light of the order of the principles, it can compute and assign all relevant aspects of this step.*

Step 2: Reason Aggregation. In the second step, we aggregate the results from the first step. Thus, we make an argument for all the actions that 'survived' the first step. Let $Perm_{V_1}(\Phi) := \bigcup_{Arg^{\omega}_{\psi} \in V_1} Perm^{\psi}(\Phi, \omega)$ be the set of all options $\phi \in \Phi$ for which it holds that they are permissible according to at least one applicable principle $\psi \in \Psi$. Vice versa, let $Support(\phi)$ be the set of arguments from V_1 supporting (in the sense of permitting) some option $\phi \in \Phi$. Hence, $Support(\phi) := \{Arg^{\omega}_{\psi} \in V_1 | \phi \in Perm^{\psi}(\Phi, \omega)\}$. For each of these options, we then need to consider what speaks in favor of it. So we start by defining:

$$V_2 := \{Arg_{\phi} \mid \phi \in Perm_{V_1}(\Phi)\}$$

We will call the output of an argument $Arg^{\omega}_{\psi} \in V_1$ that must be taken into account into these arguments $Arg_{\phi} \in V_2$, *pro tanto reasons for option* ϕ. Consequentially

$$E_{1,2} := \{\langle Arg^{\omega}_{\psi}, Arg_{\phi} \rangle \mid \phi \in Perm^{\psi}(\Phi, \omega)\}.$$

We use $force_{\text{protanto}}$ as function for encoding the strength of the pro tanto reason for an option ϕ given some argument Arg^{ω}_{ψ} It is a function of the case's probability $P(\omega|\theta)$ and the involved principle's

relevances $relevance^{\mathfrak{P}}(Arg^{\omega}_{\psi})$. As both quantities are weights, it seems right to aggregate them by multiplying. We leave the discussion of other kinds of aggregations, like maxing out, in special contexts for future research. We thus get:

$$force_{\text{protanto}}(\langle Arg^{\omega}_{\psi}, Arg_{\psi}\rangle) = P(\omega) \cdot relevance^{\mathfrak{P}}(Arg^{\omega}_{\psi})$$

Since there is no difference in strength between any two options $\phi, \phi' \in Perm^{\Psi}(\Phi, \omega)$ – that is, between two options supported by the same argument $Arg^{\omega}_{\psi} \in V_1$ –, we will, as a shorthand, just write "$force_{\text{protanto}}(Arg_{\psi})$" in order to refer to that strength.

Now we turn to the generic form and content of the arguments in V_2. Fundamentally different from the arguments in V_1, the form of the arguments in V_2 is dynamic: The number of premises in Arg_{ϕ} depends on the number of incoming edges, each representing a reason supporting ϕ. In other words, every $Arg_{\phi} \in V_2$ contains one premise for any $Arg \in Support(\phi)$, bringing the contributed strength with it into the argument. Additionally, one further premise is added, determining the aggregation of all these incoming reasons' strengths. So the aggregation is handled within the arguments in V_2. The most intuitive candidate for aggregation is simple summation of the weights. However, it may be even more controversial, whether a simple summation is the best way to aggregate reasons, than it is how to incorporate principle relevance with case probability. Answering this question (comprehensively) is clearly outside this paper's scope and, again, we leave this highly interdisciplinary question that should make use of the rich literature on that topic, to future research (cf. [27, 32, 33]). Table 2 shows a generic version of the arguments in V_2.

Step 3: Final Action Determination. The last step is rather simple, but involves a couple of design decisions nevertheless. The remaining task, after all, is to decide for one of the options given the previous results. For this, we propose to combine the *normative, moral force* and the *normative, instrumental force* of all the remaining options. That is, we see the remaining problem as a *multi-objective optimization problem* where we aim at maximizing the *moral reason responsiveness* of the system on the one hand and the *instrumental means-end optimality* represented by *EU* maximization on the other. Before we elaborate on that point, first let us define the third layer of the graph.

First, V_3 and $E_{2,3}$. One only needs one final argument, thus $V_3 \coloneqq \{Arg_{dec}\}$. Since all arguments of the second level contribute to the final argument, we have $E_{2,3} \coloneqq V_2 \times V_3$.

The final argument Arg_{dec} consists of a varying number of premises, one for each $\phi \in Perm_{V_1}(\Phi)$, each importing the strength of the overall reason supporting ϕ. Additionally, one or more premises are included which reflect the design-decision one needs to make as mentioned above.

For the importing premises, we define the second strength function $force_{\text{overall}}$ (i.e., the weights for the edges from V_2 to V_3). This strength represents the aggregative normative strength of the supporting reasons in favor of each option $\phi \in Perm_{V_1}(\Phi)$. These are exactly the options $\phi \in \Phi$ for which it is possible in light of the system's current knowledge $\theta \in \Theta$ that they are permissible according to some principle $\psi = c \to o \in \Psi$. Here "possible" means that $P(\omega|\theta) > 0$ for some $\omega \in \Omega_{\theta}$ such that $\omega \models c$. One could, in principle, filter out options with only tiny overall forces, for instance, in order to make the argumentation graphs computationally feasible. However, for now, we remain with the general structure as we do not see sufficient reason for such thresholding on theoretical grounds.

This time, since the aggregation of the strengths was deliberately part of the arguments in V_2, we only need to identify the strength of the edges with the strength in the conclusions of these arguments – which we denote by $Arg_{\phi}.\text{ConStr}$[8]:

$$force_{\text{overall}}\left(\langle Arg_{\phi}, Arg_{dec}\rangle\right) \coloneqq Arg_{\phi}.\text{ConStr} = \sum_{Arg \in Support(\phi)} force_{\text{protanto}}(Arg)$$

[8] We use this (a little bit bulky) way of stating our idea in order to emphasize that the decision of how to aggregate pro tanto reasons for options is contained in the arguments in V_2 and is *not* part of the argumentation graph generation. Whatever one plugs into the arguments has to come out as weights of the exiting edges.

Argument Arg^{ω}_{ψ}	
(P_{ω})	ω
(P_{ψ})	if c then o
(P_{Perm_o})	if o, then $Perm^{\psi}(\Phi,\omega)$.
(C_i)	Thus: $Perm^{\psi}(\Phi,\omega)$.

Table 1: Case Distinction Arguments. The argument exemplifies the general form and generic content of the first level arguments (V_1). Note that by construction $\omega \models c$.

Argument Arg_{dec}	
$(P_{\phi_{i_1}})$	There is an overall reason supporting ϕ_{i_1} with strength $force_{\text{overall}}(\phi_{i_1})$.
$\vdots$	$\vdots$
$(P_{\phi_{i_w}})$	There is an overall reason supporting ϕ_{i_w} with strength $force_{\text{overall}}(\phi_{i_w})$.
(P_{max})	Perform one randomly picked option ϕ_{out} of those in $\text{argmax}_{\phi \in Perm_{V_1}(\Phi)} force_{\text{overall}}(\phi) + EU(\phi\|\theta)$.
(C_{final})	Thus: Perform ϕ_{out}.

Table 3: The Final Argument. We set $w := |Perm_{V_1}(\Phi)|$.

Argument Arg_{ϕ_i}	
(P_{i_1})	There is a reason r_1 with strength $force_{\text{protanto}}(r_1) := force_{\text{protanto}}(Arg_{i_1})$ for A_1
$\vdots$	$\vdots$
(P_{i_v})	There is a reason r_v with strength $force_{\text{protanto}}(r_v) := force_{\text{protanto}}(Arg_{i_v})$ for A_1
$(P_{\sum})$	For any number of reasons u: If there are some reasons $r_1,\ldots,r_u$ supporting the same option ϕ with strengths $force_{\text{protanto}}(r_1),\ldots,force_{\text{protanto}}(r_u)$, then there is an overall reason supporting ϕ with strength $\sum_{i=1}^{u} force_{\text{protanto}}(r_i)$.
(C_{ϕ_i})	Thus: There is an overall reason supporting ϕ_i with strength $\sum_{Arg \in Support(\phi_i)} force_{\text{protanto}}(Arg)$.

Table 2: Reason Aggregation Arguments: The argument exemplifies the generic form and content of the second level arguments (V_2). We set $v := |Support(\phi_i)|$.

Again, we will use a shorthand, this time "$force_{\text{overall}}(\phi)$", to refer to the strength supporting a specific option ϕ. Table 3 shows the generic argument Arg_{dec}, including the final aggregation we suggest.

Let us return to our suggestion of the simultaneous, interlocked decision method. We start by defending our decision. We believe our approach to be superior to sequential approaches in context of the here discussed quantitative, uncertainty incorporating deontic filter method for two reasons: First, if we used a sequential approach, we could run into cases like the following. Imagine two options ϕ_i and ϕ_j with $force_{\text{overall}}(\phi_i) = force_{\text{overall}}(\phi_j) + \varepsilon$ for a negligible $\varepsilon \in \mathbb{R}^+$. Presuppose that this difference rules out $force_{\text{overall}}(\phi_j)$ as impermissible. Now, imagine further that $EU(\phi_i) \ll EU(\phi_j)$. It seems odd that such a small difference in the supportive reasons should be decisive against an option which otherwise is much more suitable for the ends of the system. There might be filter functions operating on the reason's strengths overcoming this problem, but they will have to be more complicated and meticulously designed then a simple threshold filtering. Second, as our naming of the strength functions already indicates, we think of strengths of reasons as some kind of (normative) forces. The principles induce what could be called *moral (or maybe societal) normative force*, while the instrumental design decisions encode sources of what could be called *instrumental normative force*. Normative forces, in our eyes, are what pulls and pushes an agent into the direction of some option or set of options and should be combined the same way as forces are combined traditionally, namely by *summation*. This justifies our decision to maximize the sum and not, for instance, the product of the two objective functions. This results in preferring some option ϕ_i over some option ϕ_j where $force_{\text{overall}}(\phi_i) = 18, EU(\phi_i|\theta) = 3$ and

$force_{\text{overall}}(\phi_j) = 10, EU(\phi_j|\theta) = 10$ respectively (and vice versa, for interchanged $force_{\text{overall}}$ and EU values) in contrast to what would be the case if we decided for multiplication instead of summation. That is, we do not punish differences between the two objectives systematically. As we mentioned, we have not finally decided whether the suggestion we make here is the right one. Maybe there are good reasons for modeling this aggregation as linear combination of these forces with weights as functions of the distances of the different kinds of forces induced by some metric. For now, we stick to the simple summation and leave further deliberation, again, to future research that will need heavy involvement of Philosophy and the debate around the question of how to weigh reasons.

There is one potential practical disadvantage of our non-sequential approach. Our approach makes it rather impossible to get strong guarantees on the system's behavior. For instance, it will not be verifiable for a medical care robot (that decides using our method) that it will, whenever there is the smallest hope, attempt to rescue a life even if running out of power afterwards. There might be cases where the corresponding case is too improbable, such that the relevance of the corresponding, applicable principle is outweighed by some much more probable case in combination with a less relevant, applicable principle.

Still, softer properties are possible and are even necessarily given by construction. Using our approach, the system would be designed such that it chooses an option that, thanks to the construction up to now, is permissible according to at least one principle applicable to at least one possible world state. Additionally, it will always choose an option that maximizes the sum of both, the combined strength of the overall reason supporting the option and the expected utility of the options given the current knowledge. In other words, it will always act upon best reasons and will be able to offer an explanation for its behavior. Maybe, our system is not *verifiably* 'ethical', but it is such that its behavior always is *justifiable*.

We believe this result to be the right result for many but not all contexts involving autonomous systems. For very vital or dangerous situations, so in contexts that need hard regulations – like autonomous trading systems or even lethal autonomous weapon systems – we, the people, demand harder guarantees. Deontic filtering then should be able to absolutely override instrumental considerations at cost of losing liveness. All this seems true to us. But now that we have defined the whole argumentation graph and finished the sketch of our framework combining Machine Ethics and Machine Explainability, we are confident that we made the corresponding 'adjusting screws' evident. Our framework can thus be adapted to meet also these requirements. We believe that in this area of tension – desired, verifiable properties on the one side and different possible design decisions on the other side – new promising future research can be identified.

5 Conclusion

In this paper, we introduced a formal and general framework combining Machine Ethics and Machine Explainability. We did so in two (major) steps: first, we motivated and introduced a framework of Machine Ethics. Second, we constructed an instantiation of this framework enabling Machine Explainability.

In our discussion, a couple of details were left for future work. While we characterized the form and content of the arguments, we omitted a *formal* characterization of their contents. Obviously, for most of the practically interesting cases, they consist of first order, modal, temporal or deontic logic formulas. This being the case, the graph needs to be supplemented by expressive means to draw conclusions: it needs a logical system, a calculus with inferences rules. Another aspect still to be explored is the space and time complexity of our approach. Additionally, we postponed a couple of optimization questions. For instance, there might be significantly fewer world states needing consideration if some variables that constitute ω are dependent. Also we ignored that some variables might have very large or even uncountably infinite domains, such that considering all possible cases would be practically infeasible or even impossible. Here heuristics are needed to restrict the number of options to the most probable or important ones.

Additionally, there are at least four interesting and pressing interdisciplinary research questions left open. First there is the question of how to model the content and ordering of principles in a more sophisticated way and how to quantify these orderings – and if one even needs to do so. After all, in light of our results, one could be inclined to switch to a framework genuinely relying on reasons if one finds a way to decouple reasons from principles ([17] might offer a useful approach here). Second, the principle order might be context dependent. Basically, this would mean to become a particularist instead of an generalist, believing that there are *no* general principles at all governing what ought to be done and that, rather, what is a normative reasons varies from context to context in an unsystematic way (cf. [13] and [14]). Third, there is a need to make decisions regarding the question of how to aggregate and weigh reasons, where the answer, as indicated before, might well be dependent on the context of application. Fourth, we have not discussed at all how to extract useful explanations of the right kind from the generated argumentation graphs. Finally, we postponed also the question of how to handle cases involving *epistemic* uncertainty (i.e., pure non-determinism) to future research as well. There is more than enough left to work on in Machine Ethics and Machine Explainability.

References

[1] Jose M. Alonso & Gracian Trivino (2017): *An Essay on Self-explanatory Computational Intelligence: A Linguistic Model of Data Processing Systems*. In: *Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)*, doi:10.18653/v1/W17-3704.

[2] Leila Amgoud & Henri Prade (2009): *Using arguments for making and explaining decisions*. *Artificial Intelligence* 173(3), pp. 413–436, doi:10.1016/j.artint.2008.11.006.

[3] Michael Anderson & Susan Leigh Anderson (2011): *Machine Ethics*. Cambridge University Press, doi:10.1017/CBO9780511978036.

[4] Gertrude E. M. Anscombe (1958): *Modern Moral Philosophy*. *Philosophy* 33(124), pp. 1–19, doi:10.1017/S0031819100037943.

[5] Aristotle: *The Nicomachean Ethics*.

[6] Gilles Barthe, Pedro R. D'Argenio, Bernd Finkbeiner & Holger Hermanns (2016): *Facets of software doping*. In: *International Symposium on Leveraging Applications of Formal Methods*, Springer, pp. 601–608, doi:10.1007/978-3-319-47169-3_46.

[7] Kevin Baum (2016): *What the Hack Is Wrong with Software Doping?* In: *International Symposium on Leveraging Applications of Formal Methods*, Springer, pp. 633–647, doi:10.1007/978-3-319-47169-3_49.

[8] Kevin Baum, Holger Hermanns & Timo Speith (2018): *From Machine Ethics To Machine Explainability and Back*. Available at http://isaim2018.cs.virginia.edu/papers/ISAIM2018_Ethics_Baum_etal.pdf.

[9] Kevin Baum, Maximilian A. Köhl & Eva Schmidt (2017): *Two Challenges for CI Trustworthiness and How to Address Them*. In: *Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)*, doi:10.18653/v1/W17-3701.

[10] Jeremy Bentham (1789): *An Introduction to the Principles of Morals and Legislation*. doi:10.1093/oseo/instance.00077240.

[11] Krister Bykvist (2009): *Utilitarianism: A guide for the perplexed*. Bloomsbury Publishing.

[12] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng & Max Kramer (2017): *Moral Decision Making Frameworks for Artificial Intelligence*.

[13] Jonathan Dancy (2004): *Ethics Without Principles*. Oxford: Claredon Press, doi:10.1093/0199270023.001.0001.

[14] Jonathan Dancy (2017): *Moral Particularism*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, winter 2017 edition, Metaphysics Research Lab, Stanford University.

[15] Donald Davidson (1963): *Actions, Reasons, and Causes*. The Journal of Philosophy 60(23), pp. 685–700, doi:10.2307/2023177.

[16] Louise Dennis & Michael Fisher (2018): *Practical Challenges in Explicit Ethical Machine Reasoning*. arXiv preprint arXiv:1801.01422.

[17] Franz Dietrich & Christian List (2017): *What Matters and How It Matters: A Choice-Theoretic Representation of Moral Theories*. Philosophical Review 126(4), pp. 421–479, doi:10.1215/00318108-4173412.

[18] Phan Minh Dung (1995): *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*. Artificial intelligence 77(2), pp. 321–357, doi:10.1016/0004-3702(94)00041-X.

[19] Pedro R. D'Argenio, Gilles Barthe, Sebastian Biewer, Bernd Finkbeiner & Holger Hermanns (2017): *Is Your Software on Dope?* In: European Symposium on Programming, Springer, pp. 83–110, doi:10.1007/978-3-662-54434-1_4.

[20] Philippa Foot (1967): *The problem of abortion and the doctrine of double effect*. doi:10.1093/0199252866.003.0002.

[21] Benjamin Franklin (1887): *Letter to J. B. Priestley, 1772*, p. 522. Putnam, New York.

[22] Stephen M. Gardiner (2006): *A core precautionary principle*. Journal of Political Philosophy 14(1), pp. 33–60, doi:10.1111/j.1467-9760.2006.00237.x.

[23] Carl G. Hempel (1965): *Deductive-Nomological Explanation*. Aspects of Scientific Explanation, pp. 335–376.

[24] Carl G. Hempel (1965): *Inductive-Statistical Explanation*. Aspects of Scientific Explanation, pp. 381–393.

[25] Monika Hengstler, Ellen Enkel & Selina Duelli (2016): *Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices*. Technological Forecasting and Social Change 105, pp. 105–120, doi:10.1016/j.techfore.2015.12.014.

[26] Helmut Horacek (2017): *Requirements for Conceptual Representations of Explanations and How Reasoning Systems Can Serve Them*. In: Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017), doi:10.18653/v1/W17-3703.

[27] John Horty (2007): *Reasons as Defaults*. Philosophers' Imprint 7, pp. 1–28, doi:10.1093/acprof:oso/9780199744077.001.0001.

[28] Immanuel Kant (1785): *Groundwork for the Metaphysics of Morals*.

[29] Popper Karl (1963): *Conjectures and Refutations: The Growth of Scientific Knowledge*. doi:10.1063/1.3050617.

[30] Niko Kolodny & John Brunero (2016): *Instrumental Rationality*. In Edward N. Zalta, editor: The Stanford Encyclopedia of Philosophy, winter 2016 edition, Metaphysics Research Lab, Stanford University.

[31] Pat Langley, Ben Meadows, Mohan Sridharan & Dongkyu Choi (2017): *Explainable Agency for Intelligent Autonomous Systems*.

[32] Errol Lord & Barry Maguire (2016): *Weighing Reasons*. Oxford University Press USA, doi:10.1093/acprof:oso/9780199315192.001.0001.

[33] Susanne Mantel (2017): *Worldly Reasons: An Ontological Inquiry Into Motivating Considerations and Normative Reasons*. Pacific Philosophical Quarterly, doi:10.1111/papq.12094.

[34] James H. Moor (2006): *The nature, importance, and difficulty of machine ethics*. IEEE intelligent systems 21(4), pp. 18–21, doi:10.1109/MIS.2006.80.

[35] Douglas W. Portmore (2011): *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford University Press USA, doi:10.1093/acprof:oso/9780199794539.003.0007.

[36] William D. Ross (1930): *The Right and the Good*. Oxford University Press, doi:10.1093/0199252653.001.0001.

[37] John J. C. Smart & Bernard Williams (1973): *Utilitarianism: For and Against*. Cambridge University Press, doi:10.1017/CBO9780511840852.

[38] Wendell Wallach & Colin Allen (2008): *Moral machines: Teaching robots right from wrong*. Oxford University Press, doi:10.1093/acprof:oso/9780195374049.001.0001.