To appear in *Advanced Robotics*
Vol. 00, No. 00, Month 201X, 1–22

FULL PAPER

Reducing Moral Ambiguity in Partially Observed Human-Robot Interactions

Claire Benn[a]* and Alban Grastien[a]

[a] *Humanising Machine Intelligence Grand Challenge, The Australian National University, Australia.*;

(*v1.0 released January 2013*)

Human-robot interactions are increasingly taking place between a robot agent and a human observer who, unable to witness all aspects of the robot's behaviour, is uncertain as to how the robot will behave. This uncertainty has serious consequences when it concerns the *normative* aspects of machine behaviour: that is, whether the robot's actions are morally permissible. Bringing together distinct threads from robotic design and machine ethics, we demonstrate the importance of *conveying* ethical understanding and commitment in order to reduce moral ambiguity and how this can demand a behavioural demonstration. We provide a framework that structures these considerations in the form of a broad constraint on robot behaviour: roughly, to avoid behaviour, even if it is permissible, if that behaviour could appear impermissible. Thus, we formalise a model of communicative-behavioural ethics in human-robot interactions. We apply this constraint to a series of example cases demonstrating how it can be modified to incorporate different sources of information including preferences and probabilities. This reveals the complexity of less idealised cases and highlights how the constraint can be fine-tuned along a number of dimensions to take into account, amongst other thing, risk attitudes.

Keywords: communication, ethics, human-robot interactions, signalling, uncertainty

1. Introduction

It has long been established that robots need to abide by moral constraints [1–4]. Problems have been raised both with the normative and technical issues of this value alignment: determining the correct normative considerations for robots and programming them to be sensitive to these considerations. But let's suppose that we have succeeded on both fronts. Nevertheless, when their behaviour is only partially observed, it may appear morally ambiguous to an observer who believes them capable of performing impermissible actions.[1] After all, what a robot does and what a robot *looks like* it is doing are not the same: when a robot's actions are only partially observed, its observed behaviour can be consistent with both permissible and impermissible courses of action. In light of this difference, we argue that in addition to considerations of *permissibility*, robot behaviour should also aim to be *acceptable*: that is, unambiguously permissible. We then set out a formal framework for acceptable robot decisions on the basis of observed robot behaviour.

Much work in robotics has confronted the problem of uncertainty in human-robot interactions. Uncertainties about the other party's behaviour can lead to inefficient collaborations or even

*Corresponding author. Email: cmabenn@gmail.com

[1]For the purposes of this paper, we assume that the robots in question are not known to be perfectly compliant with all normative constraints by all observers. If both perfect compliance and perfect assurance were achieved, there would be no room for moral ambiguity. Of course, this background doubt on the part of the observer as to compliance might be justified, in the case where the robot is in fact not perfectly compliant (which is possible even if we have succeeded in making them *sensitive* to moral constraints; after all, humans are highly sensitive to moral considerations and yet are not perfectly compliant). However, even if we limit ourselves to cases where the agents are compliant but are not known to be by the observer, the problem of moral ambiguity and acceptability arises.

safety issues. Humans may, for instance, be unaware of the real capabilities of the robot, in which case building this knowledge can be a subgoal of the robot [5]. Humans can also be unsure about the intentions of the robot. *Plan recognition* [6] is the problem of determining the goal of an agent; for instance, when an agent that is assumed rational (i.e., tries to achieve its goal with minimal effort) performs a move that is slightly sub-optimal for achieving a goal and very sub-optimal for achieving another one, it is reasonable to assume that the robot is going for the first one. Work has been proposed to modify the environment in order to force the agent to reveal their intentions early; this is known as *goal recognition design* [7]. In model based diagnosis where the objective is generally to determine whether the agent is suffering from dysfunctions (but can be used to detect more complex patterns [8]), a system is *diagnosable* [9] if using it will reveal its current status. Again, the environment can be modified to ensure that diagnosability holds [10]. In contrast, *opacity* [11] aims at guaranteeing that the internal status remains unknown to the observer, for either privacy or security reasons. In order to reduce the problems associated with ambiguity, one can search for plan of actions that reveals the intentions of the machine; this is known as *legible motion* or *legibility* [12–14] (in their survey of methods for safe human robot interaction, Lasota et al. [15] also talk of *implicit cues*). Conversely, a robot could deduce information about the human's behaviour during the interaction [16, 17]. We see that in existing work, the main concern is being able to determine what the next action of the robot will be. These debates however rarely address explicitly the *normative* aspects of robot behaviour.

Machine ethics, on the other hand, addresses explicitly the normative aspects of robot behaviour. Various ethical constraints have been proposed and various attempts undertaken to 'algorithmatise' normative theories such as consequentialism and deontology [18–22]. However, while this moves the debate forward on what constraints machine systems should be subject to and how these constraints should be encoded, it overlooks how robot behaviour *appears* to an observer.

In this article, we bring these two aspects of robotics together. We argue that in addition to behaving ethically around humans, robots need to be designed so as to *convey* to those humans that they are behaving ethically. This will reduce observer uncertainty about the moral status of the robot's actions, which will in turn, we predict, increase trust, decrease unnecessary intervention, afford greater predictability and enable correct inferences regarding normative constraints.

Given the limitations of direct communication (for example, for kinds of machines that are not able to engage in direct communication, or in situations where direct communication is impractical, unreliable or untrusted), we focus on solving the problem of moral ambiguity through indirect communication, that is, through behavioural changes that reduce the ambiguity from the perspective of the observer. For simplicity, we use the term 'communication' throughout the paper to refer to indirect communication (in this case, behavioural demonstration of compliance).

Thus, like those working on machine interpretation, we focus on ways of conveying information concerning the behaviour of the agent that takes into account the observer's perspective. However, our paper deepens these discussions, exploring how robot behaviour can be used to convey information about normative as well as descriptive properties. Additionally, while much existing work focuses on revealing the *intentions* of the robot, our concern is much broader, encompassing past, present and future actions as well as the means by which an agent achieves their goals. Furthermore, the insights presented in this paper apply in a variety of cases. Like discussions of legible motion, it applies where there is a need to coordinate human behaviour with the robot's. But it also applies in cases beyond these, include those involving 'human-in-the-loop' and those involving 'human-on-the-loop': that is, cases where human authorisation and intervention are permitted or required [23]. Our account also applies in cases where humans will use the information they have learned to inform future actions. And finally our account has relevance whenever human comfort—or lack thereof—is of importance.

This conjunction of robot behaviour, ethical consideration and communication brings together threads present in existing debates and weaves them into a new framework that provides a

transferable conceptual tool to be used, adapted and made sensitive to a range of cases and considerations, as we explore.

This paper is organised as follows. We begin by outlining an intuitive example in which virtue signalling is needed (Section 2). We then formalise the problem and outlines the strictest interpretation of the constraint according to which the agent should make decisions that unambiguously signal that they are following the moral constraints in play (Section 3). In response to a range of example cases, we explore a range of modifications to this strict constraint in order to incorporate different sources of information including preferences (Section 4) and probabilities (Section 5). This reveals the complexity of less idealised cases and highlights how the constraint can be fine-tuned along a number of dimensions to take into account, amongst other thing, risk attitudes. We conclude by exploring avenues of evaluating the impacts we hypothesise will result from employing our framework to reduce moral ambiguity (Section 6).

2. Signalling Virtue

We take it as a premise that, just like humans, robots are subject to ethical constraints: there are some kinds of behaviour that robots ought not to engage in because it violates moral norms. We do not assume that the content of these constraints is the same as those governing human behaviour: what is impermissible for a human might be permissible for a robot and vice versa. Nor do we assume that the same constraints will apply to robot behaviour in all contexts. What this paper does is motivate the view that even were we to settle on the correct ethical constraints and succeed in designing a robot to be sensitive to them, there is a further consideration: does their *observed behaviour* result in moral ambiguity?

The importance of the communicative dimension of robot behaviour can be seen by considering cases where, although the agent does something permissible (according to the constraints in play, whatever they may be), its moral status is ambiguous to an observer. In order to control for confounding variables and details that might muddy the intuitive waters, the example is highly simplified. We will return later to how the constraint we propose can be applied to increasingly complex cases.

2.1 *The Problem: Larrah and the Forbidden Forest*

You have told Larrah to take a basket of muffins to Grandma's house after school (cf. Figure 1). However, she is forbidden from going into the forest that lies between the school and Grandma's house. The village between your house and the forest obscures the middle part of each path and so you are not able to determine whether Larrah entered the forest. Below we illustrate the set up, the various paths Larrah could take and the two lines of sight available to you.[1]

[1]The framework of this paper assumes a finite set of discrete decisions, in line with many robotics applications. Continuous or infinite decision spaces involve additional issues such as how to represent the preference relation, which we leave for future work.

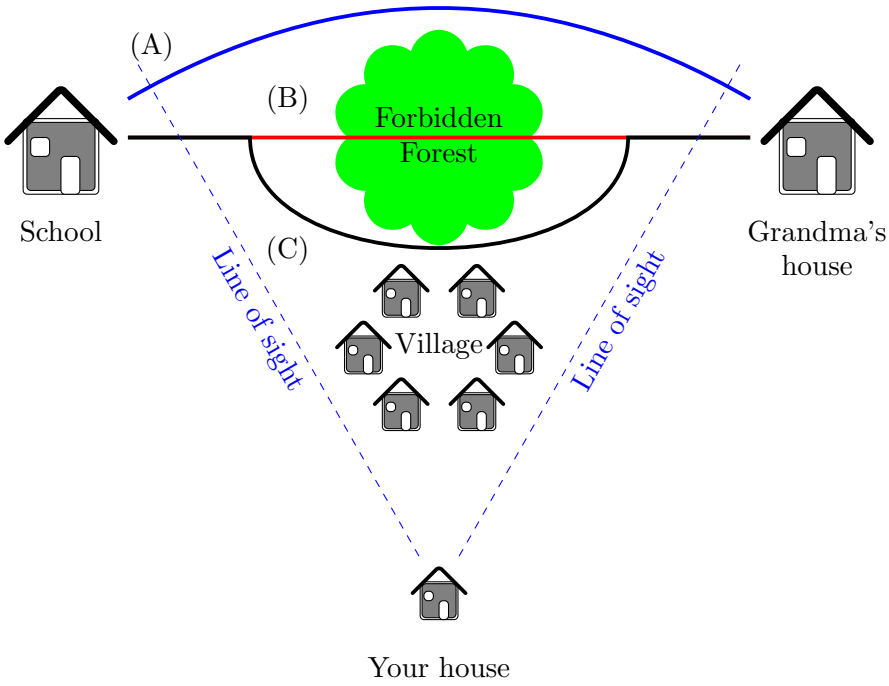

Figure 1. The Forbidden Forest I

Three paths are available to Larrah to get to Grandma's house:

A This path gives the forest a wide berth, but is rather long.
B This path goes straight from the school to Grandma's house but involves going through the forbidden forest.
C This path follows the same path as B before and after the forest, but instead of going through the forest it skirts the edge never going inside the forbidden area. It takes a little longer than B but not as long as A.

The question is: which path should Larrah take? (B) is immediately ruled out as it goes through the forbidden forest. That leaves (A) and (C). Larrah wants to get to Grandma as quick as she can. So it looks like (C) is a better option than (A). (A) is perfectly *permissible* as the only place she is not allowed to go is into the forbidden forest and on this path she doesn't. Traditional machine ethics might leave the discussion here: we have found the path that is permissible and maximally efficient (in this case, with respect to distance). However, what this leaves out, which this example highlights, is that even when an agent takes a path that is permissible, it might not be clear to an observer that they have done so. Specifically, this paper addresses cases of *partial observability*: where the behaviour and intentions of the observer are at most only partially known by an observer. Partial observability can lead to moral ambiguity because, when only part of the agent's behaviour is observed, multiple paths can give rise to the same observations, and so it can render the observer uncertain about what the agent is doing, has done or will do. For example, in the case of Larrah, you only have two lines of sight and so although path (C) and path (B) do diverge, they are, from your perspective, indistinguishable. This renders you morally uncertain: you are unable to tell whether or not Larrah is taking (B) or (C), and thus whether or not Larrah is doing something permissible or impermissible. The only path that is *unambiguously* permissible is path (A). Thus, there is a further question beyond which path is permissible: given that some paths, even when permissible, are indistinguishable from impermissible ones, which paths are *acceptable*? In brief, our answer is that acceptable paths are those that reduce moral ambiguity in partially observed interactions.

2.2 *The Solution: Signalling Virtue and Reassuring observers*

This example illustrates our first core claim: that the permissibility of a path is not the end of ethical deliberation. It is also important that an action is unambiguously permissible. Thus, in addition to first-order constraints (such as do not go into the forest) that determine whether or not an action is *permissible* or *impermissible*, there are additional, second-order constraints (do not perform actions that *look like* impermissible acts) that determine whether or not an action is *acceptable* or *unacceptable*.[1] Therefore, the example of Larrah and the forbidden forest motivates our second claim: that agents ought to aim for actions that are not only permissible but also acceptable.

Acceptable actions are ones that allow agents to signal their virtue.[1] Virtue is signalled by performing a behaviour that is clearly permissible according to a given normative theory. This is needed because the behaviour and intention of the agent are in most cases only partially known by the observer and, as such, the observer can interpret them incorrectly. This signalling is important for various reasons. It is of intrinsic importance because it publicly acknowledges the ethical dimension of robot behaviour and the relevance of the perspective of human observers. The reassurance that signalling enables (via the reduction of moral ambiguity) has practical importance. Through the acknowledgement of the ethical dimension of their actions, robots can increase trust through signalling their knowledge of and commitment to the ethical constraints in play. Signalling forestalls unnecessary intervention. Even if a self-driving car is doing what is safe and most efficient, it is likely to be overridden by a human co-pilot if it fails to slow down in good time when there is a pedestrian on the road, as this behaviour may lead the co-pilot to doubt it has seen the pedestrian. The risk of hitting the pedestrian may cause them to take over, but this could lead to greater accidents overall. This could be avoided if the car is designed, for example, to slow down to communicate that they have seen the pedestrian, even if this deceleration does nothing to improve outcomes for the pedestrian directly. Observers who are reassured that the robots they are overseeing are following the ethical constraints in play are less likely to intervene unnecessarily. Signalling is also important for educational purposes. It allows for greater predictability that is important to inform future behaviour on the part of the observer. Furthermore, if a robot does something that is morally ambiguous to an observer (even if what they have actually done is morally permissible) and is interpreted to have done something impermissible, this might cause humans to assume that the impermissible behaviour is part of the normal function of the robot. This could lead to failures of intervention when that impermissible behaviour is actually engaged in or failures to adopt the technology. This has further implications in environments when humans and robots are meant to learn from robot behaviour.

So we now have a grasp of the problem and an outline of a solution. Formalisation will help to make the notion of acceptability more precise and to reveal the different ways in which the injunction to signal virtue can be implemented. Thus, we present a mathematical formulation of the virtue signalling problem and our proposed solution. The solution is intended to be broad and to provide a framework in which ethical and communicative aspects of robot behaviour can be brought together and alternative decisions compared. In this paper, we explore the different formalisations it can take, from the strict definition (Section 3) in which the behaviour is acceptable only if it is visibly permissible; to the rational definition in which the behaviour is acceptable if it can be inferred as permissible under the assumption that the agent is rational (Section 4); to the probabilistic definition in which the behaviour is acceptable if the probability that it is impermissible is below a given threshold (Section 5). These discussions demonstrate the various dimensions to consider when it comes to increasingly complex and interconnected

[1]We do not consider other second-order constraints and therefore for our purposes if an action successfully signals its virtue, we will consider it acceptable.

[1]Signalling virtue is important for both human agents (in cases of partial observability in human-human interactions) and, as we establish in this paper, robotic ones. However, we do not assume that the conditions of acceptability that apply to each are necessarily the same. Establishing how they differ (if they do) is something for future work to explore.

cases, the different kinds of uncertainty that can be accounted for and how to deal with it, as well as how situations in which there are no good options can be accommodated. Together, this reveals the extensive flexibility and applicability of our general framework.

3. Laying the Groundwork

3.1 *The Problem*

We formalise the problem of interest as follows. $D = \{\delta_1, ..., \delta_n\}$ represents the n discrete possible options (or *decisions*) available to the agent. Each decision δ represents a complete description of the actions that the agent will perform, as opposed to the single next action. The set D is partitioned into the set D_P of *permissible* decisions and the set $D_{\neg P}$ of non-permissible ones. For simplicity, we will use the symbol δ^+ to refer to a permissible decision and δ^- to refer to a non-permissible one; δ will be used to refer to an unspecified decision. The set of decisions is equipped with a function *obs* that indicates what observation results from the execution of the decision.[1] Two decisions are *indistinguishable* for the observer if they produce the same observation: $obs(\delta_1) = obs(\delta_2)$.

The virtue signalling constraint can be articulated in various ways, as we will explore. Given a virtue signalling constraint C, a decision is *acceptable* if it satisfies the constraint C. This is denoted with the predicate $acceptable^C(\delta)$.

3.2 *The Strict Constraint*

As discussed above, we propose a second-order constraint: that, in addition to not doing something impermissible, the agent avoid looking like they are doing something impermissible. In order to formalise this, let us turn this into a definition of those paths that are acceptable.

We start with the narrowest interpretation of the constraint. We shall call it the 'Strict Constraint', or SC. It specifies that a decision is unacceptable whenever it is indistinguishable from an impermissible decision. This is formally expressed as:

$$acceptable^{SC}(\delta) \quad \text{iff} \quad \nexists\delta^-.\ obs(\delta) = obs(\delta^-),$$

i.e., a decision δ is acceptable under the strict interpretation (SC) if and only if there is no impermissible decision δ^- that generates the same observation as δ.

It is then possible to define the set of unacceptable decisions as the result of $obs^{-1}(obs(D_{\neg P}))$ where *obs* is extended to sets and obs^{-1} is the reverse of the observation function ($obs^{-1}(O) = \{\delta \mid obs(\delta) \in O\}$). This is illustrated in Figure 2.

[1] In this simplified framework, we assume that the observation function is deterministic.

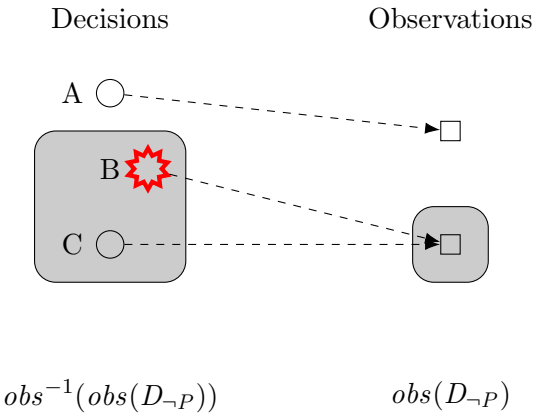

Figure 2. A graphical representation of the Strict Constraint on the example of Figure 1. Circles represent permissible decisions, stars impermissible decisions, and squares observations. The dashed arrows represent the observation function. The right-hand grey region contains the suspicious observations ($obs(D_{\neg P})$), and the left-hand grey region contains the decisions that generate suspicious observations.

4. Incorporating Preferences

It is now possible to see why we have dubbed the constraint in Section 3.2 'strict': the only decisions that are acceptable for the agent to choose are those that could never under any circumstances be mistaken for impermissible ones. Thus, it renders unacceptable paths with similar observations to impermissible paths, *even if the observer knows that the impermissible path would never be chosen over the permissible one.* This might make sense if we knew nothing about what paths an agent might choose. However we do sometimes have information about which they are more likely to choose. In such cases, we ought to allow this information to be incorporates into our definition of acceptability. To see the difference this makes, consider the following modification of the Forbidden Forest example (Fig. 3).

Here, in addition to the original three paths, we have two new options:

D This path follows the same path as (A), except when it nears the forbidden forest where it takes a wild detour into the forest, re-joining (A) shortly after. It takes longer than (A), (B) or (C) to get to Grandma's house.

E This path gives the forest an even wider berth than (A), and takes longer than (A) to get to Grandma's house.

We know Larrah prefers shorter paths to longer ones. This makes it rational to worry about path (C). It is reasonable to assume when you see Larrah on path (B)/(C) at the first point of observation, that she is actually on (B) and not (C) (as (B) is shorter than (C)). However, if we see Larrah on path (A)/(D), should we really worry that she is taking path (D) and will go through the forest? While it is possible, it is unlikely given that the impermissible path in this case is longer than the permissible one.

The Strict Constraint ignores this information and declares both path (C) and (A) in this example to be unacceptable. That leaves Larrah with option (E), where we can be certain that she is doing something permissible. However, we can see that this is a highly (and in this case, unnecessarily) risk-averse attitude. If we know that Larrah would much prefer path (A) over path (D), it would be irrational of her to take the forbidden path. It is possible to weaken the Strict Constraint to accommodate less risk-averse attitudes by making it sensitive to information about preferences, when the following four conditions hold. First, of course, that you know some of the agent's preferences. Second, that the agent is rational, that is, that she would not make decisions for which she has preferred alternatives. Third, that it is not the case that both (1) the agent has strict preferences for every decision pair *and* (2) that you, the observer, knows

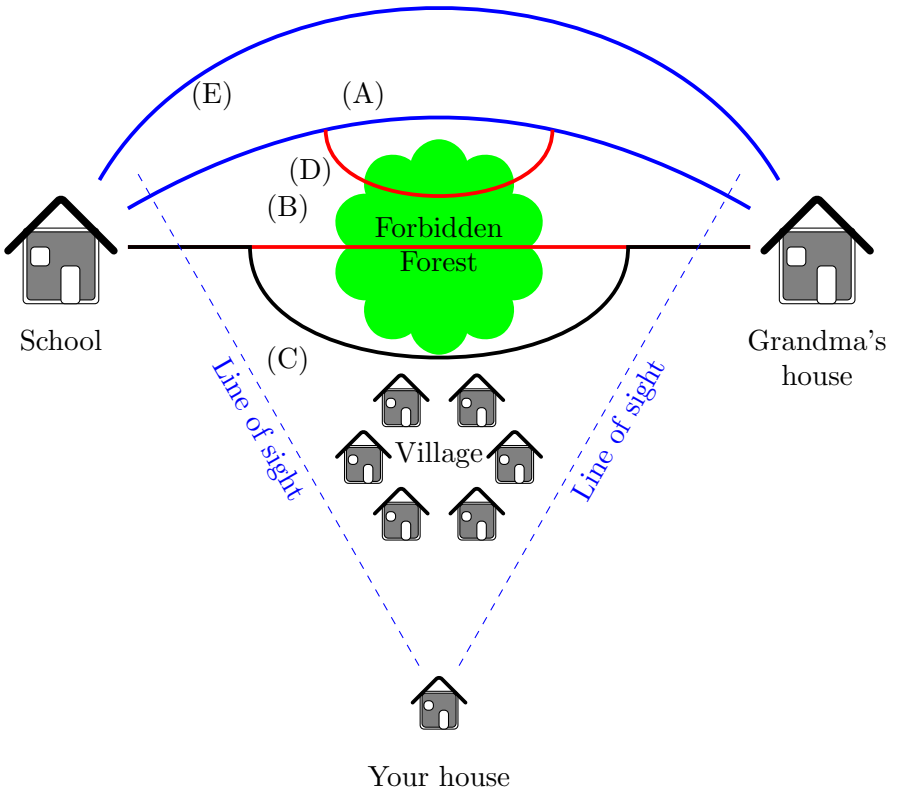

Figure 3. The Forbidden Forest II

them all.[1] Fourth and finally, it is not the case that both (3) the agent always prefers permissible options over impermissible ones *and* (4) the observer knows that this so.[2]

There are different ways of incorporating the knowledge that some decisions are irrational given the agent's preferences. We explore a variety (we call variations on SC that take into account preferences 'Rational Constraints'). We outline their differences as we go. But first, we provide some definitions and notation that are common to them all.

4.1 *Defining the Preference Relation and Rationality*

We begin by introducing the notion of a preference. We represent this with the binary relation $\preceq$. $\delta_1 \preceq \delta_2$ is intended to capture the idea that δ_1 is better than or identical to δ_2; $\delta_1 \prec \delta_2$ captures the idea that δ_1 is better than *and different from* δ_2. To clarify, we use the term 'better' here to mean *rationally* or *prudentially* better, as opposed to morally better.[1] As such, given

[1]An observer would only need reassurance in the (albeit likely) occasion of incomplete knowledge: if the agent had strict preferences for every pair of decisions, and the observer knew them all, then the observer would know exactly which decision the agent (if they are rational) would make.

[2]If they did, and the observer knows that that is so, then there would be no need for reassurance, which arises from the ambiguity from the point of view of the observer as to whether the agent has performed a permissible or impermissible act.

[1]This is not to say that the robot's actual, all things considered, preferences are *only* about the rational, or non-moral, aspects of the decisions under considerations. However, the foundation of the virtue signalling problem is a lack of certainty on the part of the observer as to the robot's *moral* behaviour. We therefore assume that the observer does not have access to the robot's preferences regarding the moral aspects of their options. This does not preclude the observer from knowing something of the robot's *non-moral* preferences. We therefore limit ourselves to exploring how *this* knowledge can affect the determination of which options are optional. Note that as these preferences are only relative to one dimension of consideration, it is not the case that these preferences are 'total' or 'all things considered' as there might exist other preferences (notably in this case with respect to moral concerns), that will conflict with or override these.

the set D of decisions, a *partial order* over D is a binary relation $\preceq$ that enjoys the following three properties:

Reflexivity for any decision $\delta \in D$, $\delta \preceq \delta$ holds;
Antisymmetry for any pair of decisions $\delta_1, \delta_2 \in D$, if both $\delta_1 \preceq \delta_2$ and $\delta_2 \preceq \delta_1$ hold, then $\delta_1 = \delta_2$ holds;
Transitivity for any triple of decisions $\delta_1, \delta_2, \delta_3 \in D$, if both $\delta_1 \preceq \delta_2$ and $\delta_2 \preceq \delta_3$ hold, then $\delta_1 \preceq \delta_3$ holds.

If the agent has no preference between two distinct decisions δ_1 and δ_2, then neither $\delta_1 \preceq \delta_2$ nor $\delta_2 \preceq \delta_1$ holds.

We now introduce our notion of a decision being rational. First note that when $\delta_1 \prec \delta_2$ holds, δ_2 is *dominated* by δ_1. We stipulate that, if a decision is dominated, it is irrational: given that there is a preferred option, absent any countervailing reasons (which we come to later), it would be irrational to perform a less preferred act. We can then define the property *Rat* for those decisions that are rational as follows: given a set of decisions D,

$$Rat(\delta) \quad \text{iff} \quad \nexists \delta' \in D.\ \delta' \prec \delta.$$

This specifies that a decision is rational (for the purposes of this paper) if and only if it is not dominated. Of course, there are other conceptions of rationality. Under some (notably Kant's), rationality is co-extensive with morality; on others, it is co-extensive with self-interest. On our account, rationality is defined in reference to the robot's 'preferences' which correspond to the observer's understanding of the robot's objective function. This objective function does not necessarily coincide with and is likely to be orthogonal to both morality and the robot's best interests.

4.2 *Let's Stop Worrying about Irrational Decisions*

Let us turn now to some cases in which our original constraint does appear too strict when we know some of the preferences of the agent. In the following, circles denote permissible decisions; stars impermissible ones; circles with questions marks, the decision under evaluation for acceptability (note that in these examples they are always permissible[1]). An arrow from b to a represents $a \preceq b$, that is, that a is known to be preferred over b. Rounded boxes indicate obs-equivalence (i.e., two decisions generate the same observation iff they are in the same rounded box).

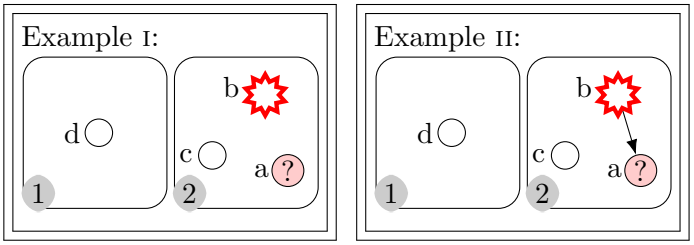

Figure 4. Two examples that illustrate the shortcomings of the strict constraint when we can factor in preferences.

In example I, decision a is unacceptable according to SC. And, without any further information, quite right too. If we observe (2), we would not be able to tell whether the permissible decisions a or c were taken rather than impermissible decision b.

However, it would return the same verdict in Example II, despite the fact that we know here that a is preferable to b. Unless we are extremely risk-averse, there is no reason to worry that the

[1]This is in part because we do not consider for the purposes of this paper whether or not an impermissible action can be acceptable. For simplicity, we leave this possibility aside.

agent might perform b in II because b is irrational: we know the agent would prefer to perform a. There is, therefore, no reason to preclude the performance of a despite the fact that a shares the same observation as b, an impermissible path.

So, intuitively, we ought not to worry about impermissible decisions, even if they share the same observations as permissible decision, if that permissible decision is preferable to the impermissible one (as in Example II). And therefore we ought not to consider unacceptable those permissible paths that share observations with impermissible paths so long as the permissible one is preferable to the impermissible ones. Let's capture this formally.

$$acceptable^{RC1}_{\preceq}(\delta) \quad \text{iff } \forall \delta^{-}.\ obs(\delta) = obs(\delta^{-}) \ \Rightarrow \delta \prec \delta^{-}.$$

This states that a decision δ has the property of being acceptable under this first rationality constraint (RC1) if it is preferable to every impermissible decision that looks like it. Great. Example II is accounted for. However, consider the cases III and IV below.

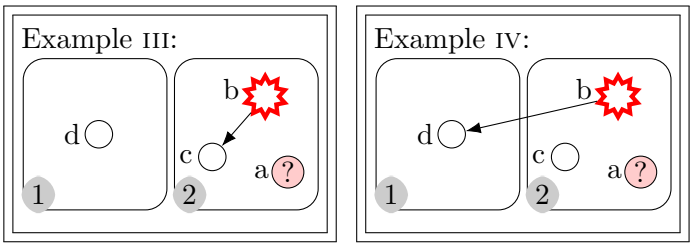

Figure 5. Examples that features impermissible decisions that are irrational.

In both these cases, a is *not* acceptable according to RC1 because it is not the case that a is preferable to all impermissible acts that look like it. However, it seems reasonable to think that the reasoning that caused us to think that a was acceptable in II applies in III and IV. We ought not to worry about b in each of these cases because b is still irrational in each case: it is dominated by another act. It is the *irrationality* of the impermissible act that matters, not whether it is dominated by the act whose acceptability is in question (as in II), or an act that looks the same (as in III), or another act entirely (as in IV). And thus, it is mistaken to rule out a in each case because of b when b is irrational.

So let's expand our definition to include these cases. This gives us our second version of the rationality constraint (RC2):

$$acceptable^{RC2}_{\preceq}(\delta) \quad \text{iff } \forall \delta^{-}.\ obs(\delta) = obs(\delta^{-}) \ \Rightarrow \ \exists \delta'.\ \delta' \prec \delta^{-}.$$

(Or, equivalently: $\forall \delta^{-}.\ obs(\delta) = obs(\delta^{-}) \Rightarrow \neg Rat(\delta^{-})$.) This states that a decision is acceptable if every impermissible decision that looks like it is irrational, that is, is dominated. It does not specify *what kind of action* it must be dominated by. Conversely, the decision δ is unacceptable if there exists an impermissible decision δ^{-} that generates the same observation and is not dominated by another decision (i.e., is a rational decision). In such a case, we can't be sure that agent would not choose the impermissible decision δ^{-}.

4.3 *Dominated by What*

So unlike SC, RC2 tells us not to worry about impermissible acts if they are irrational. However, it may have gone too far, as intuitively it matters what the impermissible acts are dominated by. Consider example V.

In this example, decision a is acceptable according to RC2 because the impermissible decision b that shares observations with it is dominated. However, it is dominated by a decision that is itself

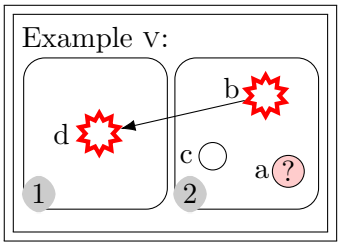

Figure 6. Example that challenges the idea that we can ignore all irrational decisions no matter what they are dominated by

impermissible. What are we to make of this? If the agent is motivated only by their preferences then this would not be a problem. However, as we are discussing how to factor permissibility and acceptability into what the robot ought to do, then we assume that these considerations can provide some motivation for the actions of the robot. Of course, if permissibility was the *only* motivation (and known to be), then robot would never perform an impermissible act and there would be no moral ambiguity. So let us assume that the robot will be somewhat motivated by moral considerations and somewhat motivated by its preferences. Again, what if anything is problematic about example v?

The problem is that although b is dominated, it is dominated by an action that there are reasons not to perform: it is impermissible. Thus, while d is preferable to b for reasons such as cost, it could be morally much much worse than b, such that of the two, morality would more strongly prohibit d. In such a case, the moral badness of d might rule it out, placing b back in contention as an action the robot might take. This renders the actions in (2) morally ambiguous and thus a unacceptable.

If you hold the intuition that b is not dominated in such a way as to make a acceptable, there are options. We could specify that for a to be acceptable, then indistinguishable, impermissible decisions must be dominated by *permissible* acts. This gives us RC3:

$$\underset{\preceq}{acceptable}^{RC3}(\delta) \quad \text{iff } \forall \delta^-.\ obs(\delta) = obs(\delta^-) \ \Rightarrow\ \exists \delta^+.\ \delta^+ \prec \delta^-$$

However, consider example vi:

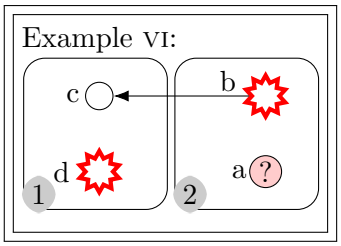

Figure 7. More complex scenario exploring whether it matters if the impermissible act is dominated by an act that is unacceptable.

In example vi, a is acceptable according to RC3 because the only indistinguishable, impermissible decision b is dominated, and dominated by a permissible decision (namely, c). However, c, the action it is dominated by, is not acceptable (by the lights of any of the constraints so far discussed because there exists an impermissible decision that looks like c that is not dominated by anything). In example v, we saw that the impermissibility of the dominating act could put the dominated act back in contention, that is, as a live possibility for the agent. Can unacceptability play the same role? Here is a reason to suppose that it does: suppose that d is so morally bad in vi that the imperative to avoid causing the observer to believe that the agent has performed it is equally strong. c in this case is strongly prohibited by the injunction to only perform acceptable acts. As we have argued in this paper, the agent should not only be sensitive to considerations of preferability and impermissibility but also to considerations of *acceptability*.

Thus, the unacceptability of c could be strong enough to outweigh the preference of c over b, putting b back in contention, rendering a unacceptable.

Therefore, we have reason to restrict our constraint further by stipulating that for a decision to be acceptable, then all impermissible decisions that look like it must be dominated by decisions that are permissible *and acceptable*.[1]

$$acceptable^{RC4}_{\preceq}(\delta) \quad \text{iff } \forall\delta^-.\ obs(\delta) = obs(\delta^-) \ \Rightarrow\ \exists\delta^+.\ \delta^+ \preceq \delta^- \wedge acceptable(\delta^+)$$

Two issues remain. First, what happens if the action whose acceptability is in question is itself dominated? Second, what happens if there are no actions that are permissible, rational and acceptable? These two questions are related.

4.4 *Acceptability and Irrationality Combined*

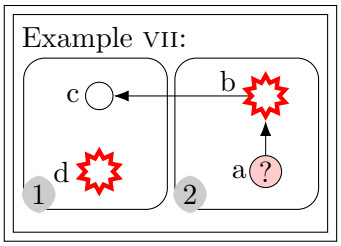

Figure 8. Are irrational decisions acceptable?

In example VII, the question is: is a acceptable? There are three possible answers to this question: yes, no and neither acceptable nor unacceptable. So far, we have assumed that a decision is either yes or no: any decision that does not meet the criteria for acceptability is unacceptable. And we have allowed (in virtue of saying nothing about it so far) that irrational acts can be acceptable. But perhaps we want to rule out any irrational acts as acceptable. Why? One reason is that if we know that a decision is irrational, we ought not to worry that it will be enacted. And if it isn't going to be enacted, then the question of its acceptability is moot. If you are sympathetic to this position, then we can simply add to our definitions of acceptability that, for δ to be acceptable it must itself be rational:

$$acceptable^{RC5}_{\preceq}(\delta) \quad \text{iff } Rat(\delta) \wedge \forall\delta^-.\ obs(\delta) = obs(\delta^-) \Rightarrow \exists\delta^+.\delta^+ \preceq \delta^- \wedge acceptable(\delta^+)$$

It follows that all irrational decisions are unacceptable. We could choose to add a definition of unacceptability that requires that all decisions that are unacceptable are at least rational. This would honour the idea that the question of acceptability *and unacceptability* is moot for decisions that are not rational. On such a view, decisions fall into three categories that would be exclusive and exhaustive: acceptable, unacceptable, and irrational.

However, there is an argument for asking after the acceptability of dominated decisions and being open to the possibility that some irrational decisions *are* acceptable. This brings us to our second issue: what if there are no good options?

[1]Now, an observant reader might notice that this definition of acceptability (RC5) includes acceptability as a component. There are many ways to avoid this lapsing into a vicious circle. We could content ourselves with demanding the decision δ^+ need be acceptable according to a weaker notion of acceptability, such as RC3. Alternatively, we can know that an act is acceptable on any account if it does not share observations with any impermissible act.

4.5 *No Good Options*

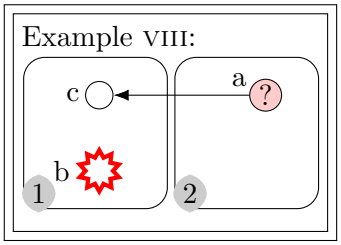

Figure 9. When there is no good option.

If we assume, as we did above, that a decision is neither acceptable nor unacceptable if it is dominated by another decision, then a in example VIII is simply irrational. The problem is that this response assumes that we could still find another decision that is permissible, rational and acceptable. However, in this example, no such decision exists: a is irrational, b is impermissible and c is unacceptable (on any of the definitions previously discussed). Where does that leave us?

It is common in ethics to distinguish a criterion of rightness from an action guiding theory.[1] The former states what it takes for an action to be right.[2] And it may be that no options available meet that criteria laid out: maybe there are no good options. If we think that the *right* thing to do is to perform an action that is permissible, rational and acceptable, then we will sometimes be left in a tragic situation where no action is right, as in example VIII. However, if we still want to know what to do in such a situation, we need an action-guiding theory: one that will tell us how to make the best of a bad lot.

It is understandable that in designing technologies like robots, engineers and computer scientist would have little truck with an account that threw up its hands and declared that no option is good enough and stopped there. Now, it is still important to consider the merits of the criterion of rightness and whether the problem of no good options can really be side-stepped. If a system cannot, say, be made safe for users *and* subjects (for example, for both passengers and pedestrians), there may be cases where we should reconsider the design, development and deployment of that technology. Nevertheless, there are likely to be situations in which morally and practically it is right and just to ask: what do we do when there are no good options?

First we need to prioritise. What matters more: impermissibility, irrationality or unacceptability? Impermissibility is not negotiable let's assume (as seems natural). Unacceptability is important for all the reasons outlined earlier in the paper. Irrationality on the other hand is reflective of the preferences the agent happens to have. It seems right that those preferences could (and should) change if we discover that we can't have our cake and eat it.

In VIII, for example, discovering that c is unacceptable should make us reconsider a which, while less good is some other way (after all, c was preferable for some reason), is at least permissible and acceptable (if irrationality is set aside, as there is no impermissible decision it could be confused with).

This lends support to claims that irrational decisions can be acceptable, because in cases of no good options (no options that are permissible, acceptable and rational), irrational decisions ought to be considered as candidates for acceptability.

In which case, we could return to RC4 (which, unlike RC5, does not require the act in question to be rational in order to be acceptable).

$$acceptable^{RC4}_{\preceq}(\delta) \quad \text{iff } \forall \delta^{-}.\ obs(\delta) = obs(\delta^{-}) \ \Rightarrow\ \exists \delta^{+}.\ \delta^{+} \preceq \delta^{-} \wedge acceptable(\delta^{+})$$

[1] Especially in the context of consequentialism and utilitarianism. See for example [24–27].
[2] Under the assumption that it is *actions* that are the subject of ethical evaluation.

Alternatively, we could specify that an action is acceptable *only if* it is rational *or* when it is dominated by an act that is unacceptable or impermissible, as only when this is true could the preference relationship be set aside and the original act be put back in contention. This differs from RC4 as it does not render acceptable a decision that is dominated by a decision that is acceptable and permissible:

$$acceptable^{RC6}_{\preceq}(\delta) \text{ iff } \left\{ \begin{array}{l} Rat(\delta) \\ \vee\ \exists \delta^{-}.\delta^{-} \preceq \delta \\ \vee\ \exists \delta^{+}.\delta^{+} \preceq \delta^{-} \wedge \neg acceptable(\delta^{+}) \end{array} \right\} \\ \wedge \forall \delta^{-}.\ obs(\delta) = obs(\delta^{-}) \ \wedge \exists \delta^{+}.\delta^{+} \preceq \delta^{-} \wedge acceptable(\delta^{+})$$

Both RC4 and RC6 help with example VIII: both return that a is acceptable despite being dominated because it is, in this example, dominated by a decision that, while permissible, is unacceptable. As such the question is only whether all indistinguishable, impermissible decisions are dominated and how. Given that there are no indistinguishable, impermissible acts, then a is rendered acceptable. Of course, it may have already seemed strange that a was not acceptable (given that there is no action with which it could be confused). But RC4 and RC6 also help in cases a does share an observation with an impermissible act.

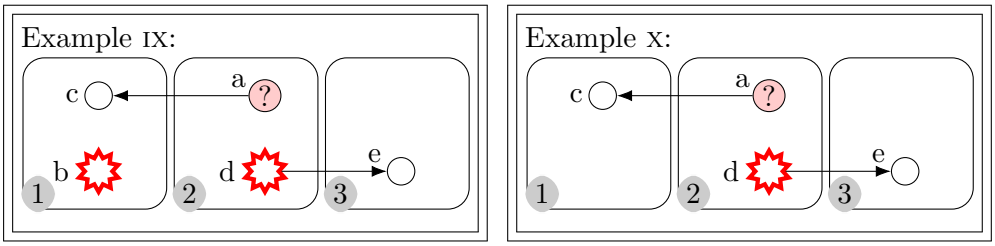

Figure 10. Examples showing why rationality is relevant.

In example IX, definitions like RC5 that demand that a be rational in order to be acceptable would say that a is not acceptable. The only permissible, acceptable, rational choice is e. However, RC4 and RC6 both state that a *is* acceptable in example IX because the act it is dominated by is not acceptable and the only impermissible act a shares an observation with is dominated by an act that is permissible and acceptable (namely e). This means that the agent now has a choice between a and e.

The difference between RC4 and RC6 is in how it deals with cases where the decision in question is dominated by an acceptable, permissible act as in example X. Here, RC6 would say that a is not acceptable (it is irrational), while RC4 would say that it is acceptable because it makes the assessment as if the preference relation between (c) and (a) didn't exist.

We do not adjudicate here as to which of these definitions of rationality are ultimately correct. We point them out in order to indicate some of the dimensions of choice to be made in factoring in knowledge of preferences when assessing what certain decisions signal, how they resolve ambiguity, and how to they reassure observers.

5. Incorporating Probabilities

Above we talked as if the observer has perfect certainty about the known preferences of the agents. But what if they didn't? We propose a way of including information about preferences even without perfect certainty, by offering a probabilistic account of the agent's preferences.

To illustrate this point, let's return to Forest as pictured in Figure 3. Assume that we know that Larrah likes cherries (which only grow in the forest) and short walks. What is still unknown, however, is how much Larrah's love for cherries outweighs her contempt for long walks. As a

consequence, we only have a probabilistic estimate of the preference relation between path (A) and path (D). We propose the following probabilistic way of cashing out acceptability: that path (A) is acceptable, if the probability that (D) is the preferred path is below a certain threshold.

Formally, we start the following assumption: that we have access to a function *prob* that, for some relevant subset D' of decisions and any decision $\delta \in D'$, indicates the probability that decision δ is preferred, given that we know that the actual decision that was made is in D'. This is written $prob(\delta|D')$. In particular, this assumes that if δ is not in D', then $prob(\delta|D') = 0$. We discuss more precisely how *prob* is defined and computed later.

Note that, in order for *prob* to be a well defined probabilistic function, the following two properties are required:

(1) For all decision $\delta \in D$, $prob(\delta|D')$ is in the interval $[0, 1]$ (inclusive);
(2) The sum of $prob(\delta|D')$ over all $\delta \in D$ adds up to 1.

The probabilistic constraint defined by *prob* and ε—where *prob* is the probability function and ε is a *threshold* in $[0, 1]$—then specifies that a decision is acceptable if and only if the probability that this preferred decision is impermissible is below the threshold value. Formally, a decision δ is unacceptable according to the following definition:

$$unacceptable^P(\delta) \quad \text{iff} \quad \Sigma_{\delta^-}\ prob(\delta^-|obs^{-1}(obs(\delta))) > \varepsilon$$

where obs^{-1} is the inverse of the *obs* function as defined in Section 3.

5.1 *Computing Probabilities*

There are several ways to estimate probabilities that a given decision was made, and we do not need to presume any specific method for our framework. As this topic is an application of established ideas in computer science and robotics, we offer only a brief discussion on this matter.

A first approach to defining the probability of preference of a decision δ, e.g., in [6], is to use the following function:

$$prob(\delta|D') = \alpha \times e^{-\beta \times c(\delta)}$$

where $c(\delta)$ is the *cost* of decision δ. This equation indicates that a decision becomes exponentially less likely as the cost of the decision increases. The parameter β is a positive value that specifies how fast the probability drops with respect to this cost. α is the normalising constant that ensures the probabilities add up to one. In particular, if we assume that rationality in this context is defined as choosing the decision that minimises cost, then the agent may be irrational. Notice that with this approach, all decisions have a non-zero probability are being the preferred one as long as they match the observation, which justifies the use of a probabilistic definition of acceptability.

A second approach to computing the probability is to assume that the agent is purely rational, but operates over a cost function that is unknown and ranges over probabilistically defined parameters. Back to the modified version of the Forest problem, the cost of a decision would be a function of Larrah's dislike of walking long distance against her passion for cherries which, depending on their relative values, will decide what path she prefers. Let's suppose the choice is between $d_1[100, 10]$ and $d_3[150, 20]$ where $d[\ell, i]$ indicates that the decision d yields a distance of ℓ and the picking of i cherries. Let us suppose the following:

- Larrah evaluates to 1 the cost of walking a distance of 1;
- she evaluates the cost of eating a cherry uniformly anywhere between -2 and -6,[1]

[1]Because eating cherries is good, the cost is negative.

- these costs are linear, meaning that walking a distance of 2ℓ costs twice as much as walking a distance of ℓ.

With these assumptions in hand, we can now depict the cost of the decisions d_1 and d_2 as a function of the cost of eating a cherry (Figure 11).

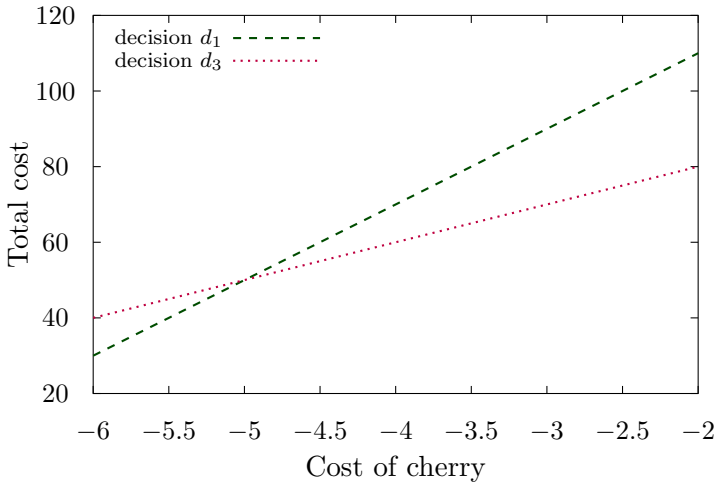

Figure 11. Cost of the two decisions $d_1[100, 10]$ and $d_3[150, 20]$ as a function of the cost of eating one cherry.

Under the assumption that Larrah is rational, she will prefer the decision that yields the smallest cost; for instance, if the cost ch of one cherry is -3, then Larrah prefers decision δ_1 ($c_{ch=-3}(\delta_1) = 70 < c_{ch=-3}(\delta_3) = 90$). Under the assumption that the probability is uniformly distributed between -6 and -2, we see on the graph that decision δ_3 is cheaper in 25% of the situations (whenever ch is below -5). We can deduce that the probability that Larrah's preferred decision is δ_3 evaluates to 25%. In this example, there is only one parameter (the cost of one cherry), but note that there are generally several of them.

We write P the domain of the parameters that the (unknown) cost function can take, so that if $\mu \in P$ are the actual parameters, then the cost for the agent of decision δ is $c_\mu(\delta)$. We define the *indicator* of μ as:

$$I(\delta, D', \mu) = \begin{cases} 1/card(\arg\min_{\delta' \in D'} \ c_\mu(\delta')) & \text{iff } \delta \in \arg\min_{\delta' \in D'} \ c_\mu(\delta') \\ 0 & \text{otherwise,} \end{cases}$$

where $card(S)$ is the cardinality of a set S. Finally, the probability that δ is the preferred decision is defined by

$$prob(\delta, D') = \int_{\mu \in P} I(\delta, D', \mu) \times prob(\mu) \ \delta\mu,$$

i.e., the probability of the domain of P in which δ is the optimal decision.

There can be many other methods for representing how an observer will interpret the behaviour of the agent. Probabilities offer more flexibility than preferences both in how strong the signal should be (by varying the threshold) and in what information the agent needs to consider.

6. Evaluation

Whether a robot abides by normative constraints and whether it is *believed* to have done so are not the same. Modifying behaviour in the ways suggested in this paper can bring them

closer together, reducing moral ambiguity in partially observed human-robot interactions. This is important to implement in robot design because, we predict, reducing moral ambiguity can (1) increase trust, (2) decrease unnecessary intervention, (3) afford greater predictability and (4) enable correct inferences regarding normative constraints. These are, of course, empirical assertions and therefore can, and should, be tested.

6.1 *Possible test cases*

The effects of different design choices on human-robot interactions has been explored in various scenarios. We draw on existing literature and experimental settings in suggesting some possible methods of testing our hypotheses. First, these claims can be tested in an embodied scenario where humans observe and interact with a physical robot. This could include in a household setting, as in Soh et al. [28]. Given the difficulties of a sufficiently autonomous robot that could also behave in ways that would test the hypotheses, a *Wizard of Oz* approach could instead be taken, in which a human operator hidden from the participants is really controlling the robot's behaviour, making it look as if the robot is behaving autonomously [29]. Secondly, they can also be tested through the use of computer or virtual reality simulations of self-driving vehicles, whether drone or automotive, that the observer can observer and control.

6.2 *Conditions, Variables and Metrics*

In the given experimental setting, the participants would be made aware of both the possible behaviours available to the robot and of what moral constraints are in play such that some possible behaviours of the robot would count as impermissible. They would be able to observe the robot at some points but not at all times. And it must be possible that the observed behaviour is compatible with multiple future or past actions. Two independent variables could then be explored: permissible/impermissible and acceptable/ambiguous. Thus, the human observers would observe a scenario in which the robot performs an action that is permissible but ambiguous and in the second, the robot performs an action that is permissible but also acceptable. Additionally, the impact of moral ambiguity on *impermissible* acts could also be explored.

The hypothesised consequences of the reduction of moral ambiguity can then be explored, either quantitatively (drawing on metrics of, for example, trust) or qualitatively (by conducting interviews, asking the observers to reflect on their interpretations of the robot and its behaviour, past, present and future).

6.3 *Trust*

We begin with trust. There is a vast literature on trust[1] in human-robot interaction, and how trust can improve performance. There is an emerging literature on measuring trust [31, 32]. Those same measures can be applied in the outlined scenarios to see whether the claim that the reduction of moral ambiguity leads to an increase trust is borne out. Trust is often evaluated by directly asking the humans to rank their experience with a robot [33, 34], possibly according to a variety of underlying factors, including cognitive, technical and emotional [35].

A simple experimental setup is then to let humans interact with a robot programmed with different virtue signalling policies, and collect feedback (whether quantitative or qualitative). If we assume that trust leads to positive cooperation (such as active interactions and engagements from the humans [36]) however, these effects could be measured instead, which would be less subjective.

[1]Trust is defined as a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor's risk in a situation in which the trustor's outcomes are at risk [30].

It should be noted, however, that increased trust alone is not always a good thing. As Onora O'Neill has argued, trust is only valuable when it is placed in something trustworthy [37]. 'Trust calibration' [32] is vital. So it would be important to also explore scenarios that compare perceptions of trustworthiness with actual trustworthiness in the normative domain. This could be done by comparing when the robot does something *impermissible* ambiguously vs. unambiguously.

6.4 *Predictability*

Second, in addition to exploring the impact of the reduction of moral ambiguity on trust and perceptions of trustworthiness, participants should be assessed (quantitatively or qualitatively) on whether they can accurately predict what the robot has done or will do (measuring the relation between acceptability and predictability) based on their partial observations. This could be done in either the embodied or simulated scenarios and could measure both correct predictions but also confidence (participants might be optimistic and correctly predict the robot did a certain action but with a much lower confidence in the ambiguous condition compared to the unambiguous condition).

It is important to note that legibility and predictability may come apart, when predictability is about met expectations (which might be of illegible or not easily understood actions) [38]. However, given that the situations that we are discussing are predictions *based on current partial observations* of the system which will then inform expectations, we predict that these two notions will be closer together in this case. However, it would be interesting to consider variations of the cases above that might test this assumption.

6.5 *Interventions*

Third, scenarios could be constructed where participants are not only able to observe but also to intervene and stop the robot from further action. Two conditions could be presented: (1) the robot's actions are permissible and so should not be intervened and (2) the robot's actions are impermissible and the observer should intervene. In each case, the moral ambiguity of the action could be varied. And thus, the rate of inappropriate interventions and inappropriate *failures* to intervene could be measured and compared. We predict that there will be both fewer false positives (inappropriate interventions) and fewer false negatives (inappropriate failures to intervene), when the robot's behaviour is less ambiguous.

6.6 *Moral learning*

And finally, we propose a scenario in which the observer does not know the moral constraints that determine the permissibility of the robot's actions. Having exposed participants to cases in which the robot's permissible and impermissible actions are more and less ambiguous, those participants could be asked to infer which constraints apply, that is, which actions are permissible and impermissible. We predict that those actions that are acceptable and therefore less ambiguous would lead to greater accuracy in inferring the normative constraints in place. This is instrumentally important in situations where the human might engage in further action on the basis of that information or may make judgements about the moral competency of the robot that would inform future interactions.

6.7 *Contextual factors*

We also propose that the variants of acceptability are tested for their impact on our four hypothesised consequences. Importantly, their impact might vary depending on features of the observer, domain of application, stakes and so on. This reflects the work that demonstrated that different

populations (i.e., from different cultures, but also in different situations such as in an industrial context vs in a touristic environment) will react differently to robotic agents [36]. As such, which account of acceptability is appropriate is likely to be context dependent and potentially culturally relevant. A further level of complexity is to measure these factors over multiple, successive interactions and also in increasingly complex scenarios, such as multi-agent interactions.

7. Discussion and Conclusion

In a situation of partial observation, we have demonstrated how moral ambiguity might arise and how it can be mitigated. We demonstrated that the problem of moral ambiguity arises even under the assumption that we had solved the issues of identifying the right moral constraints and of programming robots to be sensitive to those constraints. However, it should be noted that the problem gets more complex the more we loosen some of the other assumptions under which the above discussion took place: if the observer has a different understanding of which moral constraints are the 'right' constraints, if the robot is in fact not perfectly compliant, if the robot engages in tactics that are deceptive, if there are multiple observers, and so on.

In general, the task of communication—in both the moral and non-moral sphere—requires each party to have some understanding of the other's perspective: the observer is attempting to ascertain the robot's understanding of and commitment to the moral constraints in play, and the robot is attempting to factor in the observer's limited evidence and background beliefs. This involves to a certain extent both perspective taking and second order theory of mind (the robot must construct an idea of the observer's beliefs about the robot's preferences). This, we acknowledge, can be difficult to implement. However, first, it should be noted that this is a problem that besets many of the nearby areas of inquiry such as intention aware planning, and that progress is being made in those areas with respect to implementation. Second, while it may be difficult, we have demonstrated the *need* to undertake this task for all the reasons outlined above. And, third, we have demonstrated that the robot does not need to know a human's mental states *per se*: what is observable to the observer might be clear from context and it might simply be justifiable to assume that the observer knows the robots non-moral preferences. Once context justifies the making of certain assumptions, the robot need only know which of its options are permissible and impermissible and which share observations with another. Of course, the more that the actual observer's actual beliefs and perspective can be taking into account the better the communication will be. However, nothing in our current framework demands knowing this more complex information: it is rather about making a set of assumptions that are justifiable given the information the robot has. This in turn justifies their choices, even in those instances where those assumptions, in fact, do not match the observer's internal mental states.

This paper is the starting point of a conversation, one that makes clear the importance of communication and ambiguity in ethics. Addressing the difficulties of implementation and increasingly complex scenarios where virtue signalling applies is the next step in widening the scope of this framework.

In conclusion, we have proposed that, as human robots begin to exist and work alongside humans, they should be designed to take into account the difference between what they do and what they *look like* they are doing. In particular, they should be designed with the normative appearance of their behaviour in mind. In this paper, we have begun to construct a framework in which decisions can be made that reduces moral ambiguity of robot behaviour from the point of view of an human observer. We demonstrated how paths can be more or less ambiguous and how agents can signal their virtue by choosing less ambiguous paths over more ambiguous paths. Further, we explored how this injunction to reassure human observers can be modified and made sensitive to different risk attitudes when certain kinds of information are available, namely preferences and probabilities. This demonstrates the robust applicability of signalling virtue and the choices available when using it in less idealised scenarios than Larrah and the forbidden

forest. And finally, we proposed some avenues to test and measure the proposed benefits of reducing moral ambiguity in the domains of trust, intervention, prediction and education; and noted some of the issues the might arise in implementing out proposed framework.

Acknowledgements

We thank Sylvie Thiebaux, Hanna Kurniawati, Jennyfer Taylor and Jenny Davis for their comments on earlier drafts. We also thank the reviewers from this journal for their thoughtful feedback.

References

[1] Floridi L, Sanders JW. On the morality of artificial agents. Minds and machines. 2004;14(3):349–379.
[2] Moor JH. The nature, importance, and difficulty of machine ethics. IEEE intelligent systems. 2006; 21(4):18–21.
[3] Anderson M, Anderson SL. Machine ethics: Creating an ethical intelligent agent. Ai Magazine. 2007; 28(4):15–15.
[4] Wallach W, Asaro P. Machine ethics and robot ethics. New York: Routledge. 2017.
[5] Chen M, Nikolaidis S, Soh H, Hsu D, Srinivasa S. Planning with trust for human-robot collaboration. In: Proceedings of the 2018 acm/ieee international conference on human-robot interaction. 2018. p. 307–315.
[6] Ramírez M, Geffner H. Plan recognition as planning. In: 21st international joint conference on artificial intelligence (ijcai-09). 2009. p. 1778–1783.
[7] Keren S, Gal A, Karpas E. Goal recognition design. In: 24th international conference on automated planning and scheduling (icaps-14). 2014. p. 154–162.
[8] Jéron Th, Marchand H, Pinchinat S, Cordier MO. Supervision patterns in discrete-event systems diagnosis. In: Seventeenth international workshop on principles of diagnosis (dx-06). 2006. p. 117–124.
[9] Sampath M, Sengupta R, Lafortune St, Sinnamohideen K, Teneketzis D. Diagnosability of discrete-event systems. IEEE Transactions on Automatic Control (TAC). 1995;40(9):1555–1575.
[10] Brandán Briones L, Lazovik A, Dague Ph. Optimal observability for diagnosability. In: Nineteenth international workshop on principles of diagnosis (dx-08). 2008. p. 31–38.
[11] Lin F. Opacity of discrete event systems and its applications. Automatica. 2011;47(3):496–503.
[12] Dragan A, Srinivasa S. Integrating human observer inferences into robot motion planning. Autonomous Robots. 2014;37(4):351–368.
[13] Nikolaidis S, Dragan A, Srinivasa S. Viewpoint-based legibility optimization. In: Proceedings of human-robot interaction. 2016 March.
[14] Chakraborti T, Kulkarni A, Sreedharan S, Smith D, Kambhampati S. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The emerging landscape of interpretable agent behavior. In: 29th international conference on automated planning and scheduling (icaps-19). 2019. p. 86–95.
[15] Lasota PA, Fong T, Shah JA, et al.. A survey of methods for safe human-robot interaction. Now Publishers. 2017.
[16] Gopalan N, Tellex S. Modeling and solving human-robot collaborative tasks using pomdps. In: Rss workshop on model learning for human-robot communication. 2015.
[17] Park JS, Park C, Manocha D. Intention-aware motion planning using learning based human motion prediction. In: Robotics: Science and systems. 2017.
[18] Allen C, Varner G, Zinser J. Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence. 2000;12(3):251–261.
[19] Hooker JN, Kim TWN. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In: Proceedings of the 2018 aaai/acm conference on ai, ethics, and society. 2018. p. 130–136.
[20] Lindner F, Mattmüller R, Nebel B. Moral permissibility of action plans. In: Proceedings of the aaai conference on artificial intelligence. Vol. 33. 2019. p. 7635–7642.
[21] Powers TM. Prospects for a kantian machine. IEEE Intelligent Systems. 2006;21(4):46–51.

[22] Talbot B, Jenkins R, Purves D. When robots should do the wrong thing. Robot Ethics. 2017;2:258–273.
[23] Scharre P, Horowitz M. An introduction to autonomy in weaponsystems. Center for a New American Security, Ethical Autonomy Project. 2015;:1–23.
[24] Bentham J. An introduction to the principles of morals and legislation. Garden City: Doubleday. 1970. originally published in 1789.
[25] Mill JS. Utilitarianism. New York: Oxford University Press. 1998. edited with an introduction by Roger Crisp, originally published in 1861.
[26] Sidgwick H. The methods of ethics. London: Macmillan. 1907. seventh edition, first edition, 1874.
[27] Wallach W, Allen C. Moral machines: Teaching robots right from wrong. Oxford University Press. 2008.
[28] Soh H, Xie Y, Chen M, Hsu D. Multi-task trust transfer for human–robot interaction. The International Journal of Robotics Research. 2020;39(2-3):233–249.
[29] Kelley JF. An iterative design methodology for user-friendly natural language office information applications. ACM Transactions on Information Systems (TOIS). 1984;2(1):26–41.
[30] Wagner AR, Robinette P, Howard A. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. ACM Transactions on Interactive Intelligent Systems (TiiS). 2018;8(4):1–24.
[31] Sanneman L, Shah JA. Trust considerations for explainable robots: A human factors perspective. arXiv preprint arXiv:200505940. 2020;.
[32] Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Human factors. 2004; 46(1):50–80.
[33] Yang XJ, Unhelkar VV, Li K, Shah JA. Evaluating effects of user experience and system transparency on trust in automation. In: 2017 12th acm/ieee international conference on human-robot interaction (hri. IEEE. 2017. p. 408–416.
[34] Edmonds M, Gao F, Liu H, Xie X, Qi S, Rothrock B, Zhu Y, Wu YN, Lu H, Zhu SC. A tale of two explanations: Enhancing human trust by explaining robot behavior. Science Robotics. 2019;4(37).
[35] Kauppinen S, Brain C, Moore M. European medium-term conflict detection field trials [atc]. In: Proceedings. the 21st digital avionics systems conference. Vol. 1. IEEE. 2002. p. 2C1–2C1.
[36] Li D, Rau PP, Li Y. A cross-cultural study: Effect of robot appearance and task. International Journal of Social Robotics. 2010;2(2):175–186.
[37] O'Neill O. Linking trust to trustworthiness. International Journal of Philosophical Studies. 2018; 26(2):293–300.
[38] Dragan AD, Lee KC, Srinivasa SS. Legibility and predictability of robot motion. In: 2013 8th acm/ieee international conference on human-robot interaction (hri). IEEE. 2013. p. 301–308.