Ethical Dilemmas in Strategic Games

Pavel Naumov,[1] **Rui-Jie Yew**[2]
[1] King's College, Pennsylvania, USA
[2] Scripps College, California, USA
pgn2@cornell.edu, ryew8098@scrippscollege.edu

Abstract

An agent, or a coalition of agents, faces an ethical dilemma between several statements if she is forced to make a conscious choice between which of these statements will be true. This paper proposes to capture ethical dilemmas as a modality in strategic game settings with and without limit on sacrifice and for perfect and imperfect information games. The authors show that the dilemma modality cannot be defined through the earlier proposed blameworthiness modality. The main technical result is a sound and complete axiomatization of the properties of this modality with sacrifice in games with perfect information.

Introduction

In this paper we study ethical dilemmas faced by agents and coalitions of agents in multiagent systems. As an example, consider the two diagrams in Figure 1. In the situation depicted in the left diagram, an agent must choose between action left (L) and action right (R). These actions will result in the death of Alice and Bob, respectively. The right diagram adds an additional neutral action (N) that results in the system nondeterministically transitioning either in state u or state v and killing Alice or Bob, respectively.

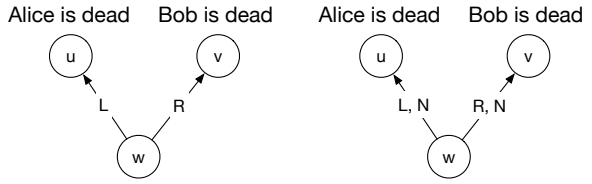

Figure 1: Two situations.

The situations represented by these two diagrams are similar in many respects. In both of them, in state w the agent has a *strategy* to kill Alice (action L) and a strategy to kill Bob (action R). Additionally, in both settings, the agent will be *blamed* for the same outcomes. To claim this, we use an oft-cited (Widerker 2017) definition of blameworthiness through the principle of alternative possibilities: "a person is morally responsible for what he has done only if he could have done otherwise" (Frankfurt 1969). For example, if the system transitions from state w to state u on either of the diagrams, then the agent will be blamed for the death of Alice because the agent had a strategy (action R) to prevent Alice's death. However, the agent is not blamable for the statement "either Alice or Bob is dead", because, in both diagrams, the agent does not have a strategy to prevent the statement from being true.

However, there is a difference in the types of choices the agent must make in these two diagrams. In the left diagram, the agent has to make a hard choice between either consciously killing Alice or consciously killing Bob. On the right diagram, the agent can avoid this choice by selecting action N. We say that, on the left diagram, the agent is facing a moral dilemma between killing Alice and killing Bob, while on the second diagram the agent does not.

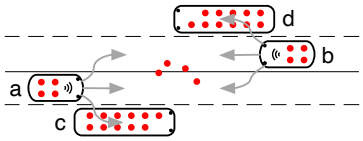

Figure 2: Road traffic situation.

As another example, consider the traffic situation depicted in Figure 2. Here, four pedestrians (red circles in the middle) are stranded on a busy four-lane highway. Self-driving cars a and b are on the path to run them over. It is too late for either of the cars to stop. Car a has three options: to pull left, to keep driving straight, or to pull right into bus c. Similarly, car b can pull left, drive straight, or pull right into bus d.

$a\backslash b$	left	straight	right
left	φ_p	$\varphi_a \wedge \varphi_b$	$\varphi_b \wedge \varphi_d$
straight	$\varphi_a \wedge \varphi_b$	φ_p	$\varphi_b \wedge \varphi_d$
right	$\varphi_a \wedge \varphi_c$	$\varphi_a \wedge \varphi_c$	$\varphi_a \wedge \varphi_b \wedge \varphi_c \wedge \varphi_d$

Table 1: Strategic game between cars a and b.

Table 1 shows different outcomes of this strategic game between players a and b. In this table, letters φ_p, φ_a, φ_b, φ_c,

and φ_d represent the death of the pedestrians, the passengers in car a, the passengers in car b, some of the passengers in bus c, and some of the passengers in bus d, respectively.

In this situation car a faces a choice: it can either pull right or not pull right. In the former case, it is *guaranteed* to kill its own passengers as well as some of the passengers in bus c. In the latter case one of the following is *guaranteed* to happen: either pedestrians will die, cars a and car b will collide, or car b and bus d will collide. In other words, car a is facing a dilemma between an action that will force $\varphi_a \wedge \varphi_c$ and the action that will force $\varphi_p \vee (\varphi_a \wedge \varphi_b) \vee (\varphi_b \wedge \varphi_d)$. We denote this dilemma of car a by formula

$$[a : \varphi_a \wedge \varphi_c, \varphi_p \vee (\varphi_a \wedge \varphi_b) \vee (\varphi_b \wedge \varphi_d)].$$

Similarly, car b is facing dilemma

$$[b : \varphi_p \vee (\varphi_a \wedge \varphi_b) \vee (\varphi_a \wedge \varphi_c), \varphi_b \wedge \varphi_d]. \quad (1)$$

Philosophers distinguish several approaches to morality. Consequential ethicists judge the moral acceptability of actions based on their outcomes. For example, a utilitarian (consequential) ethicist might say that it is morally unacceptable to kill more than a certain number of civil casualties in a military operation. On the other hand, absolute ethicists find certain actions morally unacceptable no matter what their results are. For example, a Kantian ethicist might object to pushing the lever in a trolley dilemma in order to sacrifice one person and save five. Many of such moral constraints can be modeled using the *cost of sacrifice* approach that we propose in this paper. We assign a cost of sacrifice to each action and specify the limit on the acceptable sacrifice for each agent as a subscript of the dilemma modality. For a utilitarian facing an ethical dilemma, the sacrifice is the number of civil casualties. For the absolute ethicist, sacrifice is $+\infty$ for all actions that are not morally acceptable.

The same approach can be used to model constraints imposed by laws, regulations, or company policies. For example, recently introduced German regulations for autonomous vehicles state that, when confronted with the choice between the death of a human being and damage to property, a self-driving car must always choose the latter (Fabio et al. 2017). In this case, cost of an action can be defined as the minimal number of people the action is guaranteed to kill above the unavoidable minimum. For example, if a hypothetical car is choosing between four actions that are guaranteed to kill 5, 9, 5, and 7 people respectively, then the costs of these actions are 0, 4, 0, and 2. The German rule would require a car to select one of the two actions with zero cost.

According to Car and Driver magazine, Mercedes-Benz manager of driver assistance systems and active safety Christoph von Hugo stated that "If you know you can save at least one person, at least save that one. Save the one in the car. ... If all you know for sure is that one death can be prevented, then that's your first priority." (Taylor 2016). This potential policy for future Mercedes-Benz self-driving vehicles defines the cost of an action as the minimal number of people *inside the vehicle* the action is guaranteed to kill above the unavoidable minimum. The policy also sets the allowed sacrifice in terms of this cost to zero.

Let us now assume that car a (but not car b) in Figure 2 is a self-driving vehicle made by Mercedes-Benz. Under the above policy[1], the car will never choose to pull into bus c. Thus, car a is now facing a vacuous one-option dilemma that any action that the car takes will result in statement $\varphi_p \vee (\varphi_a \wedge \varphi_b) \vee (\varphi_b \wedge \varphi_d)$ being true. We write this as

$$[a : \varphi_p \vee (\varphi_a \wedge \varphi_b) \vee (\varphi_b \wedge \varphi_d)]_{a,b \mapsto 0,+\infty},$$

where sacrifice function $a, b \mapsto 0, +\infty$ assigns the maximal sacrifice that each agent is ready to tolerate. In our case, the limit on the number of people inside the vehicle that car a is ready to sacrifice is 0. Car b does not have any fixed sacrifice limit, which we interpret as the value of the sacrifice function for agent b being $+\infty$. Note that although agent b in this situation does not have a sacrifice limit, the limit on the sacrifice of agent a modifies not only a's dilemma but b's as well. Compare the following statement to statement (1):

$$[b : \varphi_p \vee (\varphi_a \wedge \varphi_b), \varphi_b \wedge \varphi_d]_{a,b \mapsto 0,+\infty}.$$

If self-driving cars a and b decide to cooperate and make a joint decision, then instead of two individual dilemmas they face a single *multiagent* ethical dilemma. Let us first assume that neither of these two vehicles is a Mercedes-Benz. Thus, they can either (i) kill all pedestrians by driving in two different lanes, (ii) kill passengers in cars a and b by sending both vehicles for a head-on collision, (iii) collide car a with bus c, or (iv) collide car b with bus d:

$$[a, b : \varphi_p, \varphi_a \wedge \varphi_b, \varphi_a \wedge \varphi_c, \varphi_b \wedge \varphi_d]_{a,b \mapsto +\infty,+\infty}.$$

Recall that if a is a Mercedes-Benz car, then it is restricted from pulling right into bus c because this action is guaranteed to kill passengers inside car a. Note, however, that, though there is always the chance that car b pulls left and crashes into car a, there is no guarantee that car a will collide with car b. Thus, the same Mercedes-Benz policy does not restrict car a from pulling left. Let us now consider the case where both a and b are Mercedes-Benz vehicles making a joint decision. Does the policy restrict them from a joint decision under which car a drives straight and car b pulls left? In other words, is the policy a restriction on individual actions of Mercedes-Benz cars or a restriction on joint decisions of all Mercedes-Benz vehicles? If the former is true, as it is in the formal model described in this paper, then coalition $\{a, b\}$ is facing a dilemma between killing all pedestrians and a head-on collision: $[a, b : \varphi_p, \varphi_a \wedge \varphi_b]_{a,b \mapsto 0,0}$. If the latter is true, then the two vehicles must either drive straight or both of them must pull left. In any case, the pedestrians will die: $[a, b : \varphi_p]_{a,b \mapsto 0,0}$. Although Christoph von Hugo did not explicitly specify that this policy applies to individual vehicles, we think this is the case. If the policy were to apply to coalitions, then one might face a new version of the trolley dilemma when a fleet of Mercedes-Benz vehicles might choose to sacrifice the life of a passenger in a Mercedes-Benz vehicle in order to safe the lives of two passengers in another Mercedes-Benz vehicle. This seems

[1] Mercedes-Benz later retracted this policy stating that "to make a decision in favor of one person and thus against another is not legally permissible in Germany" (Orlove 2016).

to contradict von Hugo's claim that the first priority should be the prevention of even one death of a passenger in a Mercedes-Benz self-driving vehicle.

Overview

The rest of this paper is organized as follows. First, we describe the syntax and formal semantics of the ethical dilemma modality $[C : \varphi_1, \ldots, \varphi_n]_s$ in a strategic game setting. Then, we review literature on ethical dilemmas and compare the dilemma modality to the earlier studied blameworthiness and coalition power modalities. In particular, we show that the dilemma modality cannot be defined through the blameworthiness modality even in the single-agent setting without sacrifice. We also demonstrate how our definition of ethical dilemma can be extended to games with imperfect information. Finally, we give a complete axiomatization of our modality in the perfect information case. The proof of completeness is in the full version of this paper (Naumov and Yew 2019).

Strategic Game with Normalized Costs

Recall from the introduction that if an autonomous vehicle is confronted with the choice between four actions that are guaranteed to kill 5, 9, 5, and 7 people respectively, then the costs of these actions are 0, 4, 0, and 2. In other words, we assume that costs are "normalized" so that at least one of them is zero.

Definition 1 *A function $f : X \to [0, +\infty]$ is normalized if there is an element $x \in X$ such that $f(x) = 0$.*

The strategic games with normalized costs that we define bellow are very similar to "resource-bounded action frames" used in the semantics of Resource Bounded Coalition Logic (Alechina et al. 2011). By X^Y we denote the set of all functions from set Y to set X. Throughout the paper we assume a fixed set of propositional variables and a fixed set of agents $\mathcal{A}$.

Definition 2 *A game is a tuple $(W, M, \Delta, |\cdot|, \pi)$, where*

1. *W is a set of states,*
2. *Δ is a set of "actions",*
3. *$M \subseteq W \times \Delta^{\mathcal{A}} \times W$ is a relation, called "mechanism",*
4. *$|d|_w^a \in [0, +\infty]$ is the "cost" of action $d \in \Delta$ for $a \in \mathcal{A}$ in state $w \in W$, such that $|d|_w^a$ is **normalized** as a function of action d for any fixed values $a \in \mathcal{A}$ and $w \in W$,*
5. *$\pi(p)$ is a subset of W for each propositional variable p.*

We refer to functions in set $\Delta^{\mathcal{A}}$ as complete action profiles of the game. Informally, mechanism M captures the rules of the game. Namely, $(w, \delta, u) \in M$ if under complete action profile δ the game can transition from state w to state u. Our semantics is slightly more general than in (Alechina et al. 2011) because we assume that mechanism is a relation and not necessarily a function. In other words, we allow a complete action profile to transition the game into one of several different states. Our approach also allows some complete action profiles to result in no next state at all. We interpret this as a termination of the game. We normalize the costs of actions in order to avoid a situation when, for a given sacrifice, an agent would not have any actions to choose from. Note that Definition 2 allows actions with infinite costs. We further discuss such actions in the conclusion.

Syntax

In this paper we assume a fixed set $\mathcal{A}$ of agents. By a coalition we mean any nonempty subset of $\mathcal{A}$. By a sacrifice function we mean an arbitrary function from set $\mathcal{A}$ to set $[0, +\infty]$. It represents the maximal cost of the sacrifice that each individual agent is ready to bear.

The language Φ of our logical system is defined by the grammar $\varphi := p \mid \neg\varphi \mid \varphi \to \varphi \mid [C : X]_s$, where C is a coalition, X is a nonempty finite set of formulae, and s is a sacrifice function. We read $[C : X]_s$ as "coalition C under sacrifice constraints defined by function s has a dilemma between consciously forcing one of the statements in set X to be true". For the sake of simplicity, we abbreviate $[C : \{\varphi_1, \ldots, \varphi_n\}]_s$ as $[C : \varphi_1, \ldots, \varphi_n]_s$. We assume that Boolean connectives $\wedge$ and $\vee$ as well as constants truth $\top$ and false $\bot$ are defined as usual. By $\wedge X$ and $\vee X$ we denote the conjunction and the disjunction of all formulae in X respectively. As usual, $\wedge\varnothing$ and $\vee\varnothing$ are defined to be $\top$ and $\bot$, respectively.

Semantics

Throughout this paper, we write $f =_X g$ if $f(x) = g(x)$ for each $x \in X$. We also use shorthand notation captured in the following definition.

Definition 3 *For any game, any complete action profile δ, any state w, and any sacrifice function s, we write $|\delta|_w \le s$ if $|\delta(a)|_w^a \le s(a)$ for each agent $a \in \mathcal{A}$.*

By a strategy of a coalition C in a given game we mean any function from the set $\Delta^{\mathcal{A}}$ that assigns an action to each member of the coalition.

Now, we introduce a key definition of this paper. Its part (4) specifies the formal meaning of the multiagent dilemma modality $[C : X]_s$. Item 4(a) states that any strategy of coalition C forces a specific statement $\varphi \in X$ to be true. Item 4(b) states that X is a minimal set with such property.

Definition 4 *For each game $(W, \Delta, |\cdot|, M, \pi)$, each state $w \in W$, and each formula $\varphi \in \Phi$, the satisfaction relation $w \Vdash \varphi$ is defined recursively:*

1. *$w \Vdash p$, if $w \in \pi(p)$, where p is a propositional variable,*
2. *$w \Vdash \neg\varphi$, if $w \nVdash \varphi$,*
3. *$w \Vdash \varphi \to \psi$, if $w \nVdash \varphi$ or $w \Vdash \psi$,*
4. *$w \Vdash [C : X]_s$, if*
   - *(a) for any strategy $t \in \Delta^C$ of coalition C there is a formula $\varphi \in X$ such that for any action profile $\delta \in \Delta^{\mathcal{A}}$ and any state $u \in W$ if $|\delta|_w \le s$, $t =_C \delta$, and $(w, \delta, u) \in M$, then $u \Vdash \varphi$,*
   - *(b) for any nonempty subset $Y \subsetneq X$ there is a strategy $t \in \Delta^C$ of coalition C such that for any formula $\varphi \in Y$ there is an action profile $\delta \in \Delta^{\mathcal{A}}$ and a state $u \in W$ where $|\delta|_w \le s$, $t =_C \delta$, $(w, \delta, u) \in M$, and $u \nVdash \varphi$.*

We added the minimality condition 4(b) to the above definition in order to eliminate arbitrary irrelevant alternatives being added to set X. We believe that with this condition the definition better reflects our intuition of what a dilemma is. Without item 4(b) the definition would capture the notion of *weak dilemma* that we discuss later.

Recall that we allow a game to terminate as a result of agents' actions. For example, suppose that in a state w an agent a has three actions d_1, d_2, d_3 all of which have a cost of 1. Let action d_1 transition the system into a state in which statement φ_1 is true, action d_2 transition the system into a state in which statement φ_2 is true, and action d_3 be an action that terminates the game. Then, $w \Vdash [a:\varphi_1,\varphi_2]_{a\mapsto 1}$ is true, because each action of agent a predetermines a specific φ_i to be true in each outcome state. In other words, being able to terminate the system does not provide a way for an agent to "escape" the dilemma.

We allow set X in statement $[C:X]_s$ to be singleton. In such a case, $[C:X]_s$ is not a dilemma in the common sense of the world, but a "necessary" modality.

Literature Review

The dilemmas that we study in this paper are usually referred to in the literature as variations of the "trolley dilemma". The original trolley dilemma is proposed in (Foot 1967) as a dilemma faced by an agent who must choose between allowing five people to die and killing one person to prevent the death of those five. The distinction between letting one die and killing someone is also emphasised in (Thomson 1976, 1984) as well as in (Bruers and Braeckman 2014). Navarrete et al. study the same distinction in a virtual reality environment (2012).

At the same time, others shift the focus of the trolley dilemma away from the distinction between letting things happen and making things happen. Marczyk and Marks empirically study whether perceived moral permissibility changed when the person making a decision in the trolley dilemma stands to benefit from or be harmed by one of the outcomes (2014). Pan and Slater analyse participants' ethical reasoning when they were confronted with the trolley dilemma through an online survey versus through immersive virtual realities (2011). Chen et al. examine the differences in brain activity of Chinese undergraduates who experienced the great Sichuan earthquake when confronted with trolley dilemmic situations where they must choose to rescue one of two relatives and one of two strangers (2009). Indick et al. investigate how the gender of a person affects the decision that she makes in the trolley dilemma-like settings (2000). Bleske-Rechek et al. observe that people are less likely to sacrifice the life of one person for the lives of five if the one person is young, a genetic relative, or a current romantic partner (2010). In a related work, Kawai, Kubo, and Kubo-Kawai show that most people are inclined to sacrifice an older person over a younger one (2014). In this paper, we also consider trolley-like dilemmas in this broader sense.

Although we are not aware of any works treating dilemma as a modality, there are papers that use existing logical formalism to capture ethical dilemmas. Berreby, Bourgne, and Ganascia use simplified event calculus to model dilemmas within answer set programming (2015). Horty suggests using nonmonotonic logic for reasoning about moral dilemmas (1994). Bonnemains, Saurel, and Tessier propose formal notations for capturing different ethical norms that can be used in dilemmic settings (2018).

Finally, in this paper we use the cost of a sacrifice as a *constraint* on agent's available actions. In a related work, Halpern and Kleiman-Weiner propose to use the cost of a sacrifice as a *degree* of blameworthiness (2018).

Ethical Dilemma vs Blameworthiness

In this section we compare the ethical dilemma modality with blameworthiness modality. We show that the notion of ethical dilemma proposed in this paper cannot be expressed through blameworthiness, as defined through the principle of alternative possibilities: "a person is morally responsible for what he has done only if he could have done otherwise" (Frankfurt 1969). In other words, we say that an agent (or a coalition of agents) is responsible for statement φ if φ is true and the agent had a strategy to prevent φ. Several formal semantics for blameworthiness as a modality have been proposed in (Naumov and Tao 2019, 2020a,b).

The ethical dilemma modality, just like most other modalities in logic, captures a property of a state. Blameworthiness is not a property of a state, but rather of a transition between states: statement φ is true at a *current* state u, but the agent had a strategy to prevent it in the *previous* state w. As a result, if the language contains blameworthiness modality, the definition of satisfaction relation $\Vdash$ given in Definition 4 should be modified to be a ternary relation $(w,u) \Vdash \varphi$ between two states and a formula.

The goal of this section is to show that the dilemma modality cannot be defined through blameworthiness modality. To do this, we first translate the definition of ethical dilemma given in Definition 4 into the setting of the two-state satisfaction relation $(w,u) \Vdash \varphi$. While doing this, we omit the sacrifice subscript, assume that the set of agents $\mathcal{A}$ contains a single *fixed* agent a, and the set of propositional variables contains a single variable p. We do this with the intent to show a *stronger* result that the dilemma modality is not expressible through blameworthiness modality even in this simple case. In this single-agent setting, we denote coalition strategies and action profiles simply by the action of that fixed agent a.

In this section we consider language Φ_0 described by the grammar $\varphi := p \mid \neg\varphi \mid \varphi \to \varphi \mid [a:X] \mid \mathsf{B}_a\varphi$, where X is any nonempty finite set of formulae in the language Φ_0. We read $\mathsf{B}_a\varphi$ as "agent a is blamable for φ". The formal semantics for this language is given below.

Definition 5 *For each game (W,Δ,M,π), any states $w,u \in W$, and each formula $\varphi \in \Phi_0$, the satisfaction relation $(w,u) \Vdash \varphi$ is defined recursively:*

1. $(w,u) \Vdash p$, *if* $u \in \pi(p)$,
2. $(w,u) \Vdash \neg\varphi$, *if* $w \nVdash \varphi$,
3. $(w,u) \Vdash \varphi \to \psi$, *if* $w \nVdash \varphi$ *or* $w \Vdash \psi$,
4. $(w,u) \Vdash [a:X]$, *if*

(a) for any action $t \in \Delta$ there is $\varphi \in X$ such that for any state $u' \in W$ if $(w, t, u') \in M$, then $(w, u') \Vdash \varphi$,

(b) for any nonempty set $Y \subsetneq X$ there is an action $t \in \Delta$ such that for any formula $\varphi \in Y$ there is a state $u' \in W$ where $(w, t, u') \in M$, and $(w, u') \nVdash \varphi$,

5. $(w, u) \Vdash \mathsf{B}_a\varphi$, if $(w, u) \Vdash \varphi$ and there is $t \in \Delta$ such that for any state $u' \in W$ if $(w, t, u') \in M$, then $(w, u') \nVdash \varphi$.

Note that items 1 through 4 above are straightforward modifications of corresponding items in Definition 4 for a single-agent no-sacrifice language Φ_0. Item 5 captures the principle of alternative possibilities in the same way as in (Naumov and Tao 2019).

In addition to language Φ_0, we also consider a fragment $\Phi_0^{-[\,]}$ of Φ_0 that does not use the ethical dilemma modality.

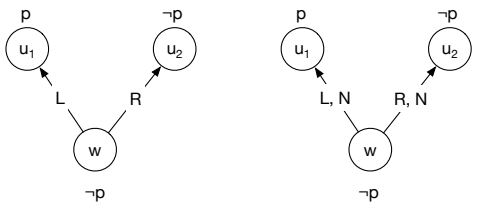

Figure 3: A game.

To show that ethical dilemma modality cannot be defined through the blameworthiness modality, we construct two single-player games that are indistinguishable in language $\Phi_0^{-[\,]}$ but are distinguishable in language Φ_0. The two games are depicted in Figure 3. These are essentially the same as in our introductory example in Figure 1. We will refer to these games as "left" and "right" games. Both games have three states: w, u_1, and u_2. In both games, propositional variable p is true in state u_1 only. In other words, $\pi_l(p) = \{u_1\} = \pi_r(p)$, where π_l and π_r are valuation functions for the left and the right games respectively. The set of actions in the left game consists of two actions: L and R. The right game includes action N in addition to actions L and R. The mechanisms M_l and M_r of the left and the right games respectively are shown in the diagrams using directed edges. For example, the edge from state w to state u_1 is labeled with action L on both diagrams. This means that $(w, L, u_1) \in M_l$ and $(w, L, u_1) \in M_r$. We will refer to the satisfaction relations for the left and the right games as $\Vdash_l$ and $\Vdash_r$ respectively.

The next lemma shows that the left and the right games are not distinguishable in language $\Phi_0^{-[\,]}$.

Lemma 1 *$(w, u) \Vdash_l \varphi$ iff $(w, u) \Vdash_r \varphi$ for any state $u \in \{u_1, u_2\}$ and formula $\varphi \in \Phi_0^{-[\,]}$.*

PROOF. We prove the statement of the lemma by structural induction on formula φ. To prove the statement in case when formula φ is propositional variable p, note that $\pi_l(p) = \{u_1\} = \pi_r(p)$, see Figure 3. Thus, $(w, u_1) \Vdash_l p$ iff $(w, u_1) \Vdash_l p$ by item 1 of Definition 5.

If φ is a negation or an implication, the desired follows from the induction hypothesis and items 2 and 3 of Definition 5 in the standard way. Suppose now that formula φ has the form $\mathsf{B}_a\psi$.

($\Rightarrow$) : Let $(w, u) \Vdash_l \mathsf{B}_a\psi$. Thus, by item 5 of Definition 5,

$$(w, u) \Vdash_l \psi \tag{2}$$

and there is an action $t \in \{L, R\}$ such that $(w, u') \nVdash_l \psi$ for any state $u \in \{u_1, u_2\}$ such that $(w, t, u) \in M_l$. Observe that $\{(w, t, u) \in M_r \mid t \in \{L, R\}\} = M_l$, see Figure 3. Thus, $(w, u') \nVdash_l \psi$ for any state $u \in \{u_1, u_2\}$ such that $(w, t, u) \in M_r$. Hence, by the induction hypothesis, $(w, u') \nVdash_r \psi$ for any state $u \in \{u_1, u_2\}$ such that $(w, t, u) \in M_r$. At the same time, also by the induction hypothesis, statement (2) implies that $(w, u) \Vdash_r \psi$. Therefore, $(w, u) \Vdash_r \mathsf{B}_a\psi$ by item 5 of Definition 5.

($\Leftarrow$) : Assume that $(w, u) \Vdash_r \mathsf{B}_a\psi$. Thus, by item 5 of Definition 5,

$$(w, u) \Vdash_r \psi \tag{3}$$

and there is an action $t \in \{L, N, R\}$ such that $(w, u') \nVdash_r \psi$ for any state $u \in \{u_1, u_2\}$ such that $(w, t, u) \in M_r$. If $t \neq N$, then the prove is similar to the one for the case ($\Rightarrow$).

Assume now that $t = N$. In other words, $(w, u') \nVdash_r \psi$ for any state $u \in \{u_1, u_2\}$ such that $(w, N, u) \in M_r$. Hence, $(w, u') \nVdash_r \psi$ for any state $u \in \{u_1, u_2\}$, see Figure 3. Thus, by the induction hypothesis, $(w, u') \nVdash_l \psi$ for any state $u \in \{u_1, u_2\}$. In particular, $(w, u') \nVdash_l \psi$ for any state $u \in \{u_1, u_2\}$ such that $(w, L, u) \in M_l$. At the same time, by the induction hypothesis, statement (3) implies that $(w, u) \Vdash_l \psi$. Therefore, $(w, u) \Vdash_l \mathsf{B}_a\psi$ by item 5 of Definition 5. ⊠

The next two lemmas show that the left and the right models are distinguishable in the language that contains ethical dilemma modality.

Lemma 2 *$(w, u) \Vdash_l [a{:}p, \neg p]$ for any state $u \in \{u_1, u_2\}$.*

PROOF. We verify the two conditions of item 4 of Definition 5 separately.

Condition (a): Consider any $t \in \{L, R\}$. Without loss of generality, let $t = L$. Consider any state $u' \in \{w, u_1, u_2\}$ where $(w, L, u') \in M_l$. To verify the condition, it suffices to show that $(w, u') \Vdash_l p$.

Indeed, assumption $(w, L, u') \in M_l$ implies $u' = u_1$, see Figure 5. Thus, $u' \in \pi_l(p)$, see Figure 5. Then, $(w, u') \Vdash_l p$ by item 1 of Definition 5.

Condition (b): Consider any nonempty set $Y \subseteq \{p, \neg p\}$. Without loss of generality, assume that $Y = \{p\}$. Let $t = R$. To verify the condition, it suffices to prove that there is a state $u' \in \{w, u_1, u_2\}$ such that $(w, t, u') \in M_l$ and $(w, u') \nVdash p$. Indeed, $u_2 \notin \pi_l(p)$, see Figure 5. Thus, $u_2 \nVdash_l p$ by item 1 of Definition 5. At the same time, $(w, R, u_2) \in M_l$, see Figure 5. ⊠

Lemma 3 *$(w, u) \nVdash_r [a{:}p, \neg p]$ for any state $u \in \{u_1, u_2\}$.*

PROOF. We will show that condition 4(a) of Definition 5 does not hold. Indeed, consider strategy $t = N$ and any formula $\varphi \in \{p, \neg p\}$. To show that the condition does not hold, it suffices to find state $u' \in \{w, u_1, u_2\}$ such that $(w, N, u') \in M_r$ and $u' \nVdash_r \varphi$. Without loss of generality,

let $\varphi = p$. Note that $u_2 \notin \pi_r(p)$, see Figure 5. Thus, $u_2 \nVdash_r p$ by item 1 of Definition 5. At the same time $(w, N, u_2) \in M_r$, see Figure 5. ⊠

The next theorem follows the three previous lemmas.

Theorem 1 *Ethical dilemma modality* $[\,]$ *is not definable in language* $\Phi_0^{-[\,]}$. ⊠

Ethical Dilemma vs Coalition Power

Marc Pauly proposed a logic of coalition power that captures properties of modality "coalition C has a strategy to achieve φ" (Pauly 2001, 2002). His approach has been widely studied in the literature (Goranko 2001; van der Hoek and Wooldridge 2005; Borgo 2007; Sauro et al. 2006; Ågotnes et al. 2010; Ågotnes, van der Hoek, and Wooldridge 2009; Belardinelli 2014; Goranko, Jamroga, and Turrini 2013; Naumov and Ros 2018). Alur, Henzinger, and Kupferman introduced Alternating-Time Temporal Logic (ATL) that combines temporal and coalition modalities (2002). Goranko and van Drimmelen gave a complete axiomatization of ATL (2006). (Alechina et al. 2011) introduce resource-bounded coalitional logic (RBCL). A logical system with a modality labeled by budget and profit is introduced in (Cao and Naumov 2017).

The dilemma modality $[C : X]_s$, even without the sacrifice subscript s, cannot be expressed in the original logic of coalition power. This can be shown using the same two models from Figure 3 that we used to prove Theorem 1. However, this modality, *without the sacrifice subscript*, can be expressed via *socially friendly coalition power* modality introduced in (Goranko and Enqvist 2018). Its authors proposed several versions of socially friendly modality. The basic one, $[C](\varphi; \psi_1, \ldots, \psi_n)$ stands for "coalition C has an action profile that guarantees φ and enables the complementary coalition $\overline{C}$ to realize any one of $\psi_1, \ldots, \psi_k$ by a suitable action profile". Our modality $[C\!:\!\varphi_1, \ldots, \varphi_n]$ without the sacrifice function is expressible through socially friendly modality as $[C](\top; \varphi_1, \ldots, \varphi_n) \wedge \bigwedge_{D \subsetneq C} \neg[D](\top; \varphi_1, \ldots, \varphi_n)$.

Unlike ours, the logical system proposed in (Goranko and Enqvist 2018) does not consider cost of actions. Thus, our modality $[C\!:\!X]_s$ with the sacrifice function s is not expressible in their system. They sketch the proof that their axiomatization of socially friendly modality is complete, but, unlike us, do not claim strong completeness. The completeness proofs here and in (Goranko and Enqvist 2018) use different constructions – see our discussion in (Naumov and Yew 2019). Additionally, none of the axioms in (Goranko and Enqvist 2018) is similar to our main axiom, the Combination axiom. Also, recall that the mechanism in Definition 2 is nondeterministic. This means that statement $[C\!:\!\varphi_1, \ldots, \varphi_2]$ does *not* imply that the complement of coalition C has a strategy to force each of the statements $\varphi_1, \ldots, \varphi_2$. Goranko and Enqvist's statement $[C](\top, \varphi_1, \ldots, \varphi_n)$ does imply this.

Ethical Dilemma and Imperfect Information

Recall our introductory example in which an agent is facing a dilemma because she has to make a hard choice between consciously killing Alice and consciously killing Bob. As we discuss there, the agent does not face a dilemma if she can avoid the hard choice by using action N and leaving the outcome up to chance. The other case when the agent does not have to make a hard choice between consciously killing Alice and consciously killing Bob is when she is *unaware* of the possible outcomes of her actions.

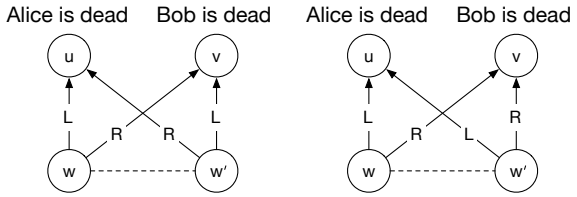

Figure 4: Two settings with imperfect information.

Consider, for example, the *left* diagram in Figure 4. This diagram depicts an imperfect information game with states w and w' indistinguishable to the agent. In state w the agent has a choice between action L and action R. The first of these actions results in Alice's death, the second in Bob's death. However, the agent does not know which action results in whose death because she cannot distinguish state w from state w' where the same actions have the opposite effect. Thus, by choosing one of the two actions in state w, the agent does not make a hard choice between consciously killing Alice and consciously killing Bob. We say that she does not face a dilemma in this setting. At the same time, the agent does face a dilemma in the setting depicted in the *right* diagram in Figure 4 because in both indistinguishable states the actions lead to the same outcome.

To formally define ethical dilemma modalities in imperfect information game settings, one needs to add an indistinguishability equivalence relation $\sim_a$ between states to Definition 2 of the game. Furthermore, because this definition allows costs of actions to vary from state to state, we need to assume that the cost of the action to an agent a is the same in all a-indistinguishable states. In other words, we need to assume that the cost of the action is *known* to the agent.

After the above changes are done to Definition 2, one can modify item 4 of Definition 4 to capture ethical dilemma in imperfect information setting as as shown below. We write $w \sim_C u$ if $w \sim_a u$ for all agents $a \in C$.

Definition 6 *For each game* $(W, \{\sim_a\}_{a\in\mathcal{A}}, \Delta, |\cdot|, M, \pi)$ *with imperfect information, each state* $w \in W$*, and each formula* $\varphi \in \Phi$*, the satisfaction relation* $w \Vdash \varphi$ *is defined recursively:*

4. $w \Vdash [C\!:\!X]_s$*, if*
   - *(a) for any strategy* $t \in \Delta^C$ *of coalition* C *there is a formula* $\varphi \in X$ *such that for any action profile* $\delta \in \Delta^{\mathcal{A}}$ *and any states* $w', u \in W$ *if* $w \sim_C w'$*,* $|\delta|_w \leq s$*,* $t =_C \delta$*, and* $(w', \delta, u) \in M$*, then* $u \Vdash \varphi$*,*
   - *(b) for any nonempty subset* $Y \subsetneq X$ *there is a strategy* $t \in \Delta^C$ *of coalition* C *such that for any formula* $\varphi \in Y$ *there is an action profile* $\delta \in \Delta^{\mathcal{A}}$ *and states* $w', u \in W$

where $w \sim_C w'$, $|\delta|_w \leq s$, $t =_C \delta$, $(w', \delta, u) \in M$, *and* $u \nVdash \varphi$.

Later in this paper we propose a sound and complete logical system for ethical dilemma modality with sacrifice in a perfect information setting. A logical system that describes an interplay between distributed knowledge and blameworthiness in an imperfect information setting is introduced in (Naumov and Tao 2020b). We leave the development of a similar system for knowledge and dilemmas for the future.

Weak Dilemma

In the next section we state the axioms of our logical system that capture the properties of modality $[C:X]_s$. When stating these axioms, it will be convenient to define $[\![C:X]\!]_s$ as an abbreviation for formula $\bigvee_{\varnothing \neq Z \subseteq X} [C:Z]_s$. In other words, $[\![C:X]\!]_s$ means that each action profile of coalition C forces a specific formula in set X to be true, but set X is *not necessarily* a minimal such set. We call expression $[\![C:X]\!]_s$ a *weak dilemma*. Alternatively, $[\![C:X]\!]_s$ could be defined by omitting condition 4(b) from Definition 4.

Axioms

In this section we list and discuss the axioms and inference rules of our logical system. The first of these axioms uses the notation $X \otimes Y$. For any two sets of formulae X and Y, let $X \otimes Y$ be the set of formulae $\{\varphi \wedge \psi \mid \varphi \in X, \psi \in Y\}$.

In addition to propositional tautologies in language Φ, our logical system contains the following axioms:

1. Combination: $[C:X]_s \to ([C:Y]_s \to [\![C:X \otimes Y]\!]_s)$,
2. Monotonicity: $[C:X]_{s'} \to [\![D:X]\!]_s$, where $C \subseteq D$ and $s \leq s'$,
3. Minimality: $[C:X]_s \to \neg[C:Y]_s$, where $Y \subsetneq X$,
4. No Alternatives: $[C:X]_s \to [D:X]_s$, where $|X| = 1$.

We write $\vdash \varphi$ if formula $\varphi \in \Phi$ is derivable in our logical system using the Modus Ponens, the Necessitation, and the Substitution inference rules

$$\frac{\varphi, \quad \varphi \to \psi}{\psi} \qquad \frac{\varphi}{[C:\varphi]_s} \qquad \frac{\{\varphi \to \tau(\varphi) \mid \varphi \in X\}}{[C:X]_s \to [\![C:\tau(X)]\!]_s},$$

for each function τ that maps set Φ into set Φ. If $\vdash \varphi$, then we say that formula φ is a theorem of our system. We write $X \vdash \varphi$ if formula φ is provable from all theorems of our logical system and an additional set of formulae X using the Modus Ponens inference rule only.

The Combination axiom states that if each action profile of coalition C forces a specific formula in set X to be true and a specific formula in set Y to be true, then each action profile of coalition C forces a specific formula in set $X \otimes Y$ to be true. Indeed, if a particular action profile forces $\varphi \in X$ to be true and $\psi \in Y$ to be true, then this profile also forces $\varphi \wedge \psi$ to be true. A hypothetical Combination axiom with the single bracket modality in the conclusion is not sound.

The Monotonicity axiom states that if each action profile of coalition C forces a specific formula in set X to be true under a more relaxed constraint s' on sacrifice, then each action profile of a larger coalition D forces a specific formula in set X to be true under a stronger constraint s. A hypothetical Monotonicity axiom with single bracket modality in the conclusion is also not sound. The Minimality axiom captures the minimality requirement of item 4(b) in Definition 4.

The No Alternatives axiom deals with the extreme case of a singleton set $X = \{\varphi\}$. Note that statement $[C:\varphi]_s$ means that statement φ is predetermined to be true under any action profile of coalition C as long as actions of *all agents* are constrained by s. In other words, φ is true as long as actions of all agents are constrained by s. Since the last statement does not depend on the coalition C, we may conclude that validity of statement $[C:\varphi]_s$ does not depend on the choice of coalition C. This observation is captured in the No Alternatives axiom.

The Necessitation rule states that if formula φ is true in all states of all games, then statement φ is predetermined to be true under any action profile of coalition C and any constraint s. Note that in this case the minimality condition 4(b) of Definition 4 is vacuously satisfied because singleton set $\{\varphi\}$ has no nonempty proper subsets.

The Substitution rule says that if $[C:X]_s$ and statement φ in set X is replaced with a logically weaker statement $\tau(\varphi)$, then each action profile of coalition C still forces a specific formula in the set $\tau(X)$ to be true, but $\tau(X)$ is not necessarily the smallest such set. An example of an instance of this rule is

$$\frac{\neg\neg\varphi \to \varphi, \qquad \psi \to (\chi \to \psi)}{[C:\neg\neg\varphi, \psi]_s \to [\![C:\varphi, \chi \to \psi]\!]_s}.$$

Note that X and $\tau(X)$ are sets, not lists. Thus, set $\tau(X)$ might have fewer elements than set X:

$$\frac{\varphi \to (\varphi \vee \psi), \qquad \psi \to (\varphi \vee \psi)}{[C:\varphi, \psi]_s \to [\![C:\varphi \vee \psi]\!]_s}.$$

Theorem 2 (strong soundness) *If* $X \vdash \varphi$ *and* w *is a state of a model such that* $w \Vdash \chi$ *for each formula* $\chi \in X$, *then* $w \Vdash \varphi$. ⊠

The proof of the following completeness theorem can be found in (Naumov and Yew 2019).

Theorem 3 (strong completeness) *For any set of formulae* X *and any formula* φ, *if* $X \nvdash \varphi$, *then there is a game and a state* w *of this game such that* $w \Vdash \chi$ *for each formula* $\chi \in X$ *and* $w \nVdash \varphi$.

Conclusion

The contribution of this paper is three-fold. First, we introduce a formal semantics for ethical dilemmas in a strategic game setting expressed through the modality $[C:X]_s$. Second, we show that this modality is not definable through the blameworthiness modality. Finally, we give a complete axiomatization of the properties of the dilemma modality.

Our completeness result is the *strong* completeness theorem with respect to the proposed semantics. We believe that the standard filtration technique could be used to prove *weak* completeness with respect to the class of finite games. This would imply decidability of our logical system, assuming the sacrifice function is rational-valued functions.

References

Ågotnes, T.; Balbiani, P.; van Ditmarsch, H.; and Seban, P. 2010. Group announcement logic. *Journal of Applied Logic* 8(1):62 – 81.

Ågotnes, T.; van der Hoek, W.; and Wooldridge, M. 2009. Reasoning about coalitional games. *Artificial Intelligence* 173(1):45 – 79.

Alechina, N.; Logan, B.; Nguyen, H. N.; and Rakib, A. 2011. Logic for coalitions with bounded resources. *Journal of Logic and Computation* 21(6):907–937.

Alur, R.; Henzinger, T. A.; and Kupferman, O. 2002. Alternating-time temporal logic. *Journal of the ACM* 49(5):672–713.

Belardinelli, F. 2014. Reasoning about knowledge and strategies: Epistemic strategy logic. In *Proceedings 2nd International Workshop on Strategic Reasoning, SR 2014, Grenoble, France, April 5-6, 2014*, volume 146 of *EPTCS*, 27–33.

Berreby, F.; Bourgne, G.; and Ganascia, J.-G. 2015. Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for programming, artificial intelligence, and reasoning*, 532–548. Springer.

Bleske-Rechek, A.; Nelson, L. A.; Baker, J. P.; Remiker, M. W.; and Brandt, S. J. 2010. Evolution and the trolley problem: People save five over one unless the one is young, genetically related, or a romantic partner. *Journal of Social, Evolutionary, and Cultural Psychology* 4(3):115.

Bonnemains, V.; Saurel, C.; and Tessier, C. 2018. Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology* 20(1):41–58.

Borgo, S. 2007. Coalitions in action logic. In *20th International Joint Conference on Artificial Intelligence*, 1822–1827.

Bruers, S., and Braeckman, J. 2014. A review and systematization of the trolley problem. *Philosophia* 42(2):251–269.

Cao, R., and Naumov, P. 2017. Budget-constrained dynamics in multiagent systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 915–921.

Chen, P.; Qiu, J.; Li, H.; and Zhang, Q. 2009. Spatiotemporal cortical activation underlying dilemma decision-making: an event-related potential study. *Biological Psychology* 82(2):111–115.

Fabio, U. D.; Broy, M.; Brüngger, J.; Eichhorn, U.; Grunwald, A.; Heckmann, D.; Hilgendorf, E.; Kagermann, H.; Losinger, A.; Lutz-Bachmann, M.; Lütge, C.; Markl, A.; Müller, K.; and Nehm, K. 2017. Automated and connected driving. Technical report, Ethics Commission, German Federal Ministry of Transport and Digital Infrastructure. https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile.

Foot, P. 1967. The problem of abortion and the doctrine of the double effect. *Oxford Review* (5).

Frankfurt, H. G. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66(23):829–839.

Goranko, V., and Enqvist, S. 2018. Socially friendly and group protecting coalition logics. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, 372–380. International Foundation for Autonomous Agents and Multiagent Systems.

Goranko, V., and van Drimmelen, G. 2006. Complete axiomatization and decidability of alternating-time temporal logic. *Theoretical Computer Science* 353(1):93 – 117.

Goranko, V.; Jamroga, W.; and Turrini, P. 2013. Strategic games and truly playable effectivity functions. *Autonomous Agents and Multi-Agent Systems* 26(2):288–314.

Goranko, V. 2001. Coalition games and alternating temporal logics. In *Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge*, 259–272. Morgan Kaufmann Publishers Inc.

Halpern, J. Y., and Kleiman-Weiner, M. 2018. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Horty, J. F. 1994. Moral dilemmas and nonmonotonic logic. *Journal of philosophical logic* 23(1):35–65.

Indick, W.; Kim, J.; Oelberger, B.; and Semino, L. 2000. Gender differences in moral judgement: is nonconsequential reasoning a factor. *Current Research in Social Psychology* 5(20):285–298.

Kawai, N.; Kubo, K.; and Kubo-Kawai, N. 2014. "granny dumping": Acceptability of sacrificing the elderly in a simulated moral dilemma. *Japanese Psychological Research* 56(3):254–262.

Marczyk, J., and Marks, M. J. 2014. Does it matter who pulls the switch? perceptions of intentions in the trolley dilemma. *Evolution and Human Behavior* 35(4):272–278.

Naumov, P., and Ros, K. 2018. Strategic coalitions in systems with catastrophic failures (extended abstract). In *Proceedings of the 16th International Conference on Principles of Knowledge Representation and Reasoning*, 659–660.

Naumov, P., and Tao, J. 2019. Blameworthiness in strategic games. In *Proceedings of Thirty-third AAAI Conference on Artificial Intelligence (AAAI-19)*.

Naumov, P., and Tao, J. 2020a. Blameworthiness in security games. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.

Naumov, P., and Tao, J. 2020b. An epistemic logic of blameworthiness. *Artificial Intelligence* 283. 103269.

Naumov, P., and Yew, R.-J. 2019. Ethical dilemmas in strategic games. *arXiv:1911.00786*.

Navarrete, C. D.; McDonald, M. M.; Mott, M. L.; and Asher, B. 2012. Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem". *Emotion* 12(2):364.

Orlove, R. 2016. Now Mercedes says its driverless cars won't run over pedestrians, that would be illegal. *Jalopnik*. https://jalopnik.com/now-mercedes-says-its-driverless-cars-wont-run-over-ped-1787890432.

Pan, X., and Slater, M. 2011. Confronting a moral dilemma in virtual reality: a pilot study. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, 46–51. British Computer Society.

Pauly, M. 2001. *Logic for Social Software*. Ph.D. Dissertation, Institute for Logic, Language, and Computation.

Pauly, M. 2002. A modal logic for coalitional power in games. *Journal of Logic and Computation* 12(1):149–166.

Sauro, L.; Gerbrandy, J.; van der Hoek, W.; and Wooldridge, M. 2006. Reasoning about action and cooperation. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '06, 185–192. New York, NY, USA: ACM.

Taylor, M. 2016. Self-driving Mercedes-Benzes will prioritize occupant safety over pedestrians. *Car and Driver*. https://www.caranddriver.com/news/a15344706/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/.

Thomson, J. J. 1976. Killing, letting die, and the trolley problem. *The Monist* 59:204–217.

Thomson, J. J. 1984. The trolley problem. *Yale LJ* 94:1395.

van der Hoek, W., and Wooldridge, M. 2005. On the logic of cooperation and propositional control. *Artificial Intelligence* 164(1):81 – 119.

Widerker, D. 2017. *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*. Routledge.