Preprint of a paper to appear in a

Festschrift for Krister Segerberg

Right Actions in Perspective

John Horty

Philosophy Department

University of Maryland

College Park, MD 20742

horty@umiacs.umd.edu

www.umiacs.umd.edu/users/horty

Version of: November 27, 2005

Contents

> Working with logical techniques pushes the requirement of rigour so high that pressures of complexity enforce a very narrow focus... Nontechnical philosophers are naturalists who describe what they see with the naked eye. Logicians examine nature through their microscopes or X-ray cameras: what they see is also an aspect of nature, but a different one.
>
> Krister Segerberg, *Getting Started: Beginnings in the Logic of Action*

1 Introduction

I have always liked this passage by Krister Segerberg, because I feel that it provides a good sense of both the limits and the possibilities of formal work in philosophy, as well as both the frustrations and joys of engaging in this kind of work. In this paper, I apply Segerberg's microscopic techniques to a problem in the formal philosophy of action.

Segerberg has himself originated a rich logical theory of action, growing out of his prior work in dynamic logic, and developed in a series of papers ranging from the early [29] and [30] to the more recent [31]. I will be working, however, within a different framework—the framework of "stit semantics" due to Belnap and Perloff, based ultimately on the theory of indeterminism set out in Prior's indeterministic tense logic, and developed in full detail by Belnap, Perloff, and Xu [3].

The issue I want to consider arises when certain normative, or decision theoretic, notions are introduced into this framework: here I will focus on the notion of a *right action*, and so on the formulation of act utilitarianism within this indeterministic setting. The problem is simply that there are two different, and conflicting, ways of defining this notion, both well-motivated, and both carrying intuitive weight.

This is a problem that was pointed out in my [14], but here I address what I now think of as a mistake in that treatment. In that earlier book, in order to explain our conflicting judgments about right actions, I set out two substantially different accounts of the notion, which I labeled as the "orthodox" and "dominance" accounts. But here, there is only one account, only one theory of right actions, and our conflicting intuitions are instead explained by showing how this theory yields different results when actions are evaluated from different perspectives.

The paper is structured as follows. In the first section, I review Prior's indeterministic framework as well as the structures underlying stit semantics. Although these structures were originally introduced for the purpose of interpreting formal languages containing special modal operators—tense operators, agency operators—there is none of that here. The concepts I am concerned with in this paper are defined entirely in terms of the underlying structures themselves. There is no need to introduce or interpret any formal language. In the next two sections, I motivate the two ways of understanding the notion of a right action, and define the corresponding orthodox and dominance act utilitarian theories. In the final section, I then show how these theories can be unified, while still accounting for our different intuitions about right actions.

2 Background

2.1 Individual actions

Prior's theory of indeterminism, set out in his [24], is based on a picture of moments as ordered into a treelike structure, with forward branching representing the openness or indeterminacy of the future and the absence of backward branching representing the determinacy

of the past.

This picture can be represented formally through the postulation of a nonempty set *Tree* of moments together with an ordering $<$ on *Tree* that is transitive and irreflexive, and that satisfies the treelike property according to which, for any m_1, m_2, and m_3 in *Tree*, if $m_1 < m_3$ and $m_2 < m_3$, then either $m_1 = m_2$ or $m_1 < m_2$ or $m_2 < m_1$. A maximal set of linearly ordered moments from *Tree* is a *history*, representing some complete temporal evolution of the world. If m is a moment and h is a history, then the statement that $m \in h$ can be taken to mean that m occurs at some point in the course of the history h, or that h passes through m. Of course, because of indeterminism, a single moment might be contained in several distinct histories. We let $H_m = \{h : m \in h\}$ represent the set of histories passing through m, those histories in which m occurs; and when h belongs to H_m, we speak of a moment/history pair of the form m/h as an *index*.

The set of possible worlds accessible at a moment m can be identified with the set H_m of histories passing through that moment; those histories lying outside of H_m are taken to represent worlds that are no longer accessible. We can therefore identify the *propositions* at m with the subsets of H_m, and where X is such a proposition, we can say that X is true at the history h just in case $h \in X$.

These various ideas can be illustrated as in Figure 1, where the upward direction represents the forward direction of time. This diagram depicts a branching time structure containing five histories, h_1 through h_5. The moments m_1 through m_4 are highlighted; and we have, for example, $m_2 \in h_3$ and $H_{m_4} = \{h_4, h_5\}$. The propositions available at m_4 are the four subsets of H_{m_4}, and each of these is true at every history it contains.

We now turn to the treatment of agency. The goal is to represent the notion that an

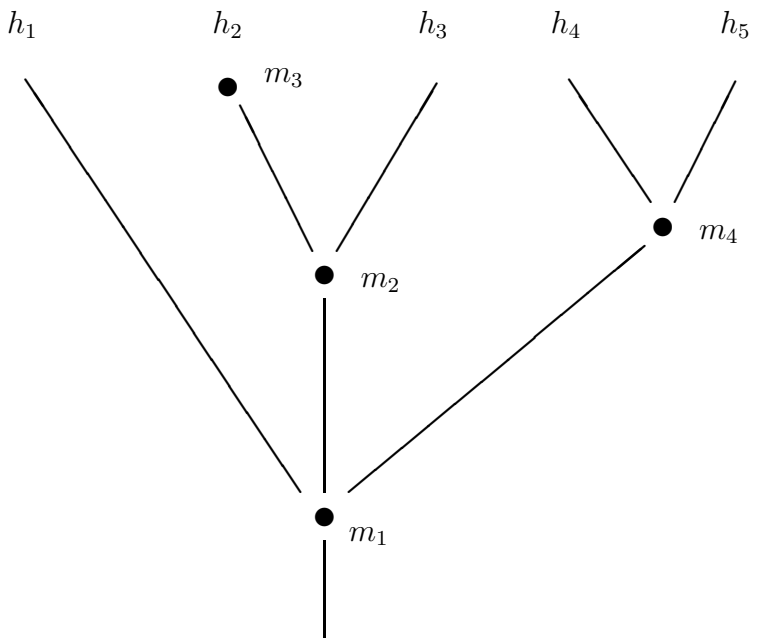

Figure 1: Branching time

agent, through its action, guarantees the truth of some proposition.[1] We must therefore be able to speak of individual agents, and also of their actions or choices; and so the basic framework of branching time is supplemented with two additional primitives.

The first is simply a set *Agent* of agents, individuals thought of as making choices, or acting, in time.

Now what is it for one of these agents to act, or choose, in this way? We idealize by ignoring any intentional components involved in the concept of action, by ignoring vagueness and probability, and also by treating actions as instantaneous. In this rarefied environment, acting can be thought of simply as constraining the course of events to lie within some definite subset of the possible histories still available. When an agent α butters the toast, for example, the nature of its action is to constrain the history to be realized so that it must

[1] In an effort to find language that is both gender neutral and unobtrusive, I assume here that the agents are impersonal acting devices, such as robots, which it is appropriate to refer to using the pronoun 'it'.

lie among those in which the toast is buttered. Of course, such an action still leaves room for a good deal of variation in the future course of events, and so cannot determine a unique history; but it does rule out all those histories in which the toast is not buttered.

Our second additional primitive, then, is a device for representing the possible constraints that an agent is able to exercise upon the course of events at a given moment, the actions or choices open to the agent at that moment. These constraints are encoded formally through a function *Choice*, mapping each agent α and moment m into a partition $Choice^m_\alpha$ of the set of histories H_m through m.[2] The idea behind this formalism is that, by acting at m, the agent α selects a particular one of the equivalence classes—or *choice cells*—from $Choice^m_\alpha$ within which the history to be realized must then lie, but that this is the extent of the agent's influence.

If K is a choice cell belonging to $Choice^m_\alpha$, one of the equivalence classes specified by this partition, we speak of K as an *action* available to the agent α at the moment m. We say that α *performs* the action K at the index m/h just in case h belongs to K, and we speak of the set of histories belonging to K as the *possible outcomes* that might result from this action. It is important to notice that all of the information provided by a full index is required in determining whether an agent performs an action: it makes no sense to say that an agent performs an action at a moment, but only at a moment/history pair.

These various concepts relating to choice functions are illustrated in Figure 2, which depicts a structure containing six histories, and in which the actions available to the agent α at three moments are highlighted. The cells at the highlighted moments represent the

[2]The *Choice* function is subject to two technical constraints of "no choice between undivided histories" and "independence of actions," which I will not go into here. The constraints can be found in my [14], and are described in authoritative detail in Belnap, Perloff, and Xu [3].

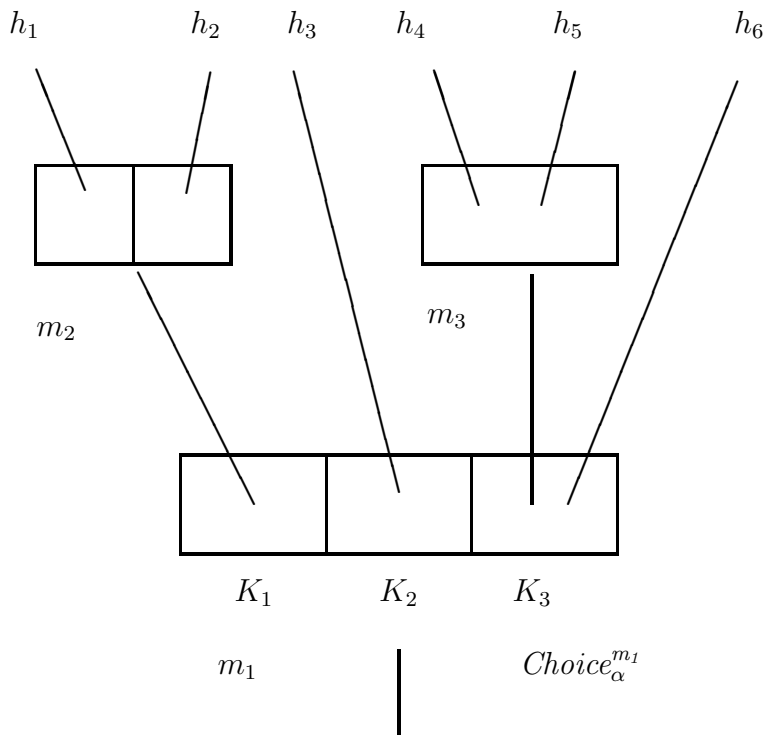

Figure 2: An agent's choices

actions available to α at those moments. For example, there are three actions available to α at m_1—$Choice_\alpha^{m_1} = \{K_1, K_2, K_3\}$, with $K_1 = \{h_1, h_2\}$, $K_2 = \{h_3\}$, and $K_3 = \{h_4, h_5, h_6\}$. Because h_1 and h_2 are still undivided at m_1, these two histories must fall within the same partition there, and likewise for h_4 and h_5. The agent α faces two choices at m_2, but at m_3 it effectively has no choice: histories divide, but there is nothing α can do to constrain the outcome.

Returning to the moment m_1, we can say that α performs the action K_1 at the index m_1/h_2, for example, that it performs the action K_2 at m_1/h_3, and that it performs the action K_3 at m_1/h_6. Again: since the agent performs different actions along different histories through the moment m_1, there is no sense in asking what action it performs at that moment. Finally, we can speak of h_4, h_5, and h_6 as the outcomes that might result from performing

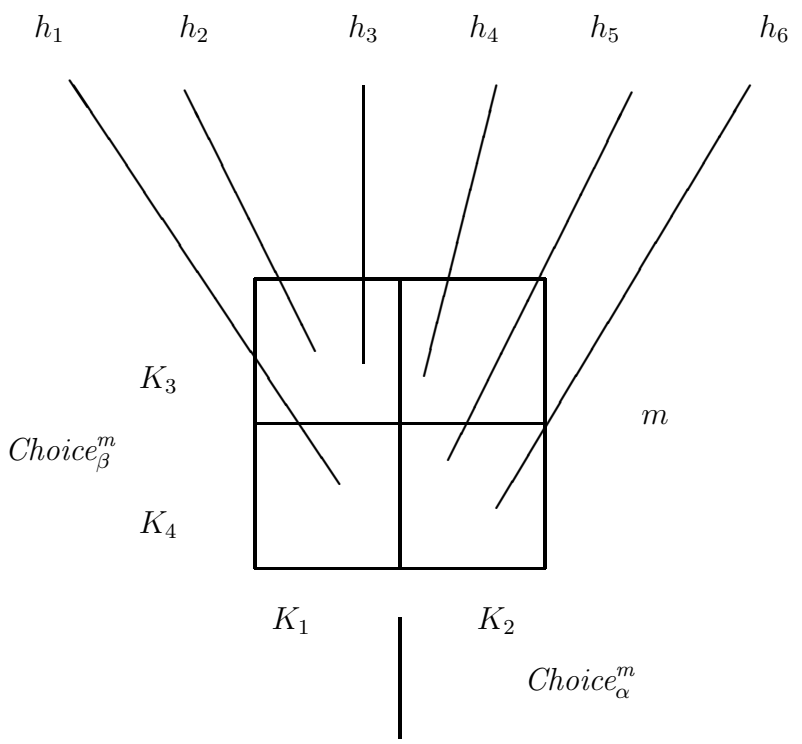

Figure 3: Group actions

the action K_3, for example.

2.2 Group actions

In developing an account of group actions, it is best to begin with an example; so consider the multiple agent situation depicted in Figure 3. Here, the actions open to the agent α at the moment m are depicted by the vertical partitions of H_m; that is, $Choice^m_\alpha = \{K_1, K_2\}$, with $K_1 = \{h_1, h_2, h_3\}$ and $K_2 = \{h_4, h_5, h_6\}$. The actions open to the agent β are depicted by the horizontal partitions; $Choice^m_\beta = \{K_3, K_4\}$, with $K_3 = \{h_2, h_3, h_4\}$ and $K_4 = \{h_1, h_5, h_6\}$.

Now consider the proposition $X = \{h_2, h_3, h_6\}$. It should be clear that, in this situation, neither the agent α nor the agent β acting alone has the ability to see to guarantee the truth of X. Each action available to each of these agents allows for a possible outcome in which

X fails. Still, it seems that the group of agents $\{\alpha, \beta\}$ acting together does have the ability to guarantee the truth of X. If α performs the action K_1 and β performs the action K_3, the group $\{\alpha, \beta\}$ can be said to perform the action $K_1 \cap K_3$, and X holds at each possible outcome of this group action.

As this example suggests, group actions can usefully be defined as patterns of individual actions: an action available to a group of agents can be defined as an intersection of the actions available to the individual agents belonging to that group, one action for each agent.

In order to develop this suggestion, it is convenient to reify patterns of action by defining an *action selection* function at a moment m as a function assigning to each agent some action available to that agent at m—that is, a function s mapping each agent α into some member of $Choice^m_\alpha$. Each of these action selection functions represents a possible pattern of action at the moment m, a selection of an available action for each agent. These patterns of action can be collected together into the set $Select_m$, containing the various action selection functions at m. And where Γ is a group of agents, the set $Choice^m_\Gamma$ of action available to the group at the moment m—the patterns of action available to the members of the group—can then be defined as follows:

$$Choice^m_\Gamma = \{\bigcap_{\alpha \in \Gamma} s(\alpha) : s \in Select_m\}.$$

And it should be clear that this definition says what it should: the set of actions available to the group Γ is identified with the set of intersections of actions available to the agents belonging to that group, one action for each agent.

3 The dominance account

3.1 Our question

With this much of the formal framework in place, we now add one final primitive: a function *Value* mapping each history into a real number representing the overall value of that history, however that is conceived. This new primitive is illustrated in Figure 4, where the numbers written beside histories indicate the values assigned to those histories, so that, for example, $Value(h_1) = 10$.

Now that values have been assigned to the various histories consistent with an agent's actions—the various possible outcomes of those actions—we can turn to the central question of this paper: How, in this indeterministic setting, can we characterize the act utilitarian notion of a *right action* for the agent to perform?

According to the standard formulation of act utilitarianism, an action is defined as right if there is no action among the available alternatives with better consequences, and wrong otherwise.[3] In the present framework, it is easy enough to define the alternatives available to an agent α at a moment m; these are simply the actions belonging to $Choice^m_\alpha$. And our *Value* function, of course, provides a straightforward ranking of possible outcomes. But in a setting that is genuinely indeterministic, how can we define the notion of an action's consequences?

The problem that a robust indeterminism presents for the characterization of an action's consequences—and so for a definition of act utilitarianism—was noted some time ago by Prior, in his contribution to a symposium on the topic:

[3]Perhaps the most careful formulation of act utilitarianism can be found in Bergström [4], which aims to develop in a precise way the theory described in Chapters 1 and 2 of Moore [22].

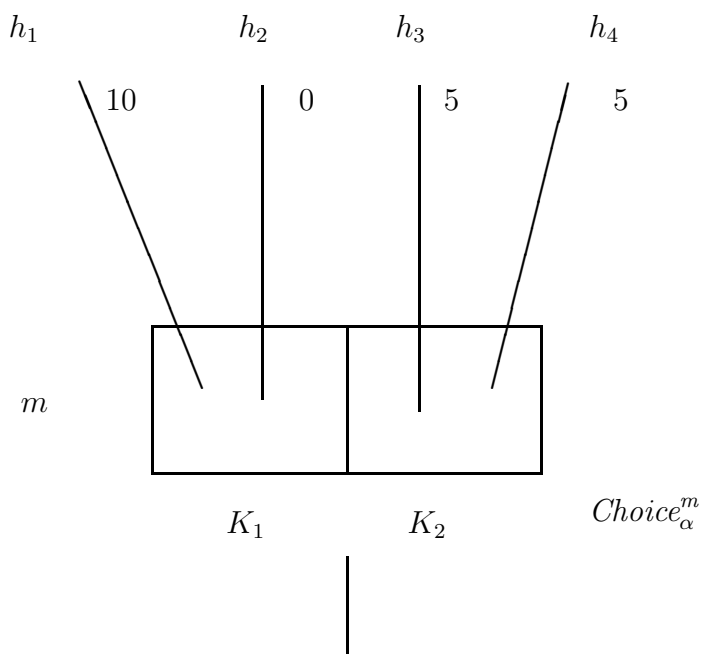

Figure 4: Histories with values

> Suppose that determinism is *not* true. Then there may indeed be a number of alternative actions which we could perform on a given occasion, but none of these actions can be said to have any "total consequences," or to bring about a definite state of the world which is better than any other that might be brought about by other choices . . . it's not merely that one cannot calculate the totality of what will happen if one decides in a certain way; the point is rather that there *is* no such totality. [23, pp. 91–92]

And the general point is clear enough. In the case of Figure 4, for example, the agent must choose between two available actions. The choice of K_2 leads invariably to an outcome whose value is 5, while the choice of K_1 leads to an outcome whose value is either 10 or 0, depending on whether things evolve along the lines of h_1 or h_2. But since, if K_1 is selected, it is then indeterminate whether h_1 or h_2 will be realized, how can we possibly say which of the two

actions, K_1 or K_2, has the better consequences?

In response to this problem, Prior himself offers the standard suggestion of appealing to probabilistic information, such as a probability distribution on the histories that might result from an action. Using this information, we could assign an expected value to each of the actions available to an agent, and the ordering of actions based on their expected values would then allow us to define a form of act utilitarianism that did not, in fact, rely on some definite notion of an action's consequences: an action could be defined as right whenever there is no alternative with greater expected value.

This approach—leading to a theory that might be described as *expected value act utilitarianism*—is, of course, very natural when the required probability distribution can be found. But there are many situations in which this information is either unavailable or meaningless; this is true, particularly, when the outcome resulting from an agent's action depends, not simply on a roll of the dice, but on the independent choice of another free agent. In the literature on decision theory, a situation in which the actions available to an agent might lead to their various possible outcomes with known probability is described as a case of *risk*, while a situation in which the probability with which the available actions might lead to their various possible outcomes is either unknown or meaningless is described as a case of *uncertainty*.[4]

[4]A discussion of this terminology can be found, for example, in Sections 2.1 and 13.1 of Luce and Raiffa [19]. Of course, the legitimacy of the distinction between uncertainty and risk is itself an issue: following Ramsey [25] and Savage [28], many writers in the Bayesian tradition assume that an agent's assessment of the possible outcomes in a given situation can always be represented through a probability measure, so that uncertainty always reduces to risk. However, there is an important tradition of resistance to the assimilation of uncertainty and risk in a single numerical measure. A classic paper in this tradition is Ellsberg [7]; for more recent work on decision theory in situations that mix elements of risk and uncertainty, see the papers

Our concern here is with situations involving uncertainty, rather than risk, and we proceed by adapting a standard treatment of these situations from decision theory: since an ordering based on expected value is not possible, we instead define a notion of dominance that can be used to order the actions available to an agent.

3.2 Dominance act utilitarianism

We begin with a preference ordering on propositions, arbitrary sets of histories through a moment.

> Preferences ordering on propositions: Let X and Y be propositions at a moment. Then $X \leq Y$ (Y is *weakly preferred* to X) if and only if $\mathit{Value}(h) \leq \mathit{Value}(h')$ for each $h \in X$ and each $h' \in Y$; and $X < Y$ (Y is *strongly preferred* to X) if and only if $X \leq Y$ and it is not the case that $Y \leq X$.

The idea is that, if Y is weakly preferred to X, each history from Y is at least as valuable as any history from X, so that we are sure to do at least as well in a history at which Y holds as we would in a history at which X holds. If Y is strongly preferred to X, then not only is each history from Y is at least as valuable as any history from X, but some history from Y is better than some history from X, so that we are not only sure to do at least as well with Y as with X, we might do better.

In the current framework, the actions available to an agent at a moment are reified as sets of histories through that moment. Each action is therefore a proposition, and so it is tempting to imagine that the dominance relations among actions might be identified with the preference orderings defined for propositions more generally. This idea is plausible, and

contained in Parts II and IV of Gärdenfors and Sahlin [9].

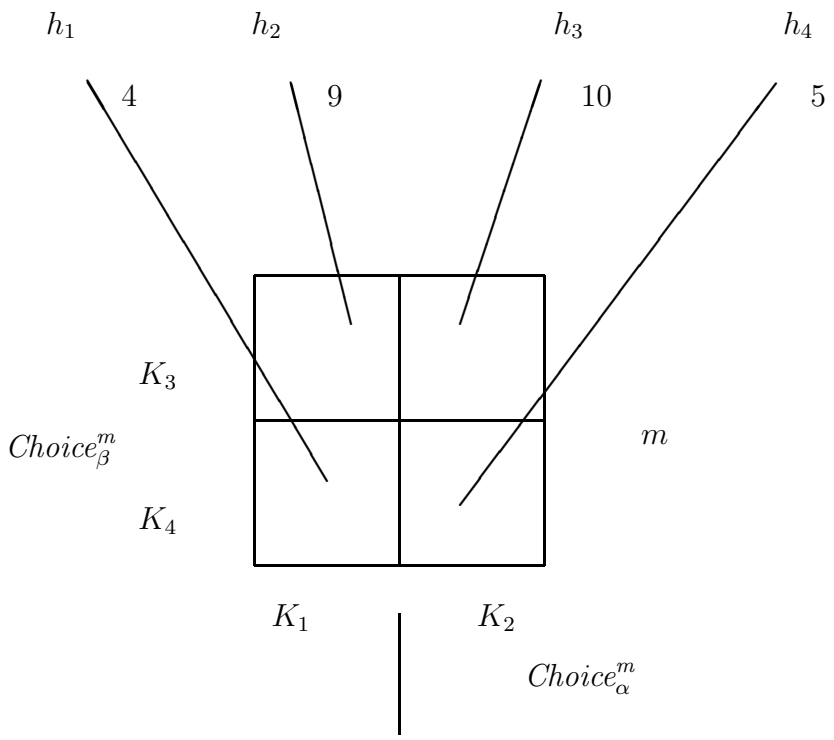

Figure 5: The coin example

there are a number of examples in which it seems to yield the correct results, including the earlier Figure 4, where it tells us that neither of the two actions, K_1 or K_2, dominates the other. However, the suggestion of simply identifying the dominance orderings over an agent's actions with the preference orderings on propositions fails in more complicated cases.

To see this, consider Figure 5, depicting a situation of simultaneous choice by two agents, and interpreted as follows. We suppose that the agent α is holding a nickel in its hand, and that at the moment m, the agent is faced with a choice between two actions: placing this nickel on a certain table either heads up, performing the action K_1, or tails up, performing the action K_2. At the same moment, the agent β must likewise choose between placing a dime onto the table either heads up or tails up, performing either the action K_3 or the action K_4. If α places the nickel on the table heads up, then the resulting utility is 9 if β places

the dime heads up and 4 if β places the dime tails up; but if α places the dime on the table tails up, the resulting utility is 10 if β places the dime heads up and 5 if β places the dime tails up.

In this situation, neither of the two actions open to α is even weakly preferred to the other in the sense of the propositional ordering, since each contains an outcome more valuable than some outcome belonging to the other. Nevertheless, there is a persuasive argument in favor of the conclusion that K_2 is a better action than K_1 for α to perform: The agent β must place the dime on the table either heads up or tails up, performing either K_3 or K_4. So suppose, first, that β places the dime heads up, performing K_3. In that case, it is clearly better for α to place the nickel on the table tails up, performing K_2 rather than K_1, since the unique history h_3 belonging to $K_2 \cap K_3$ is more valuable than the unique history h_2 belonging to $K_1 \cap K_3$. Next, suppose that β places the dime tails up, performing K_4. Then it is again better for α to place the nickel on the table tails up, again performing K_2 rather than K_1, since the unique history h_4 belonging to $K_2 \cap K_4$ is more valuable than the unique history h_1 belonging to $K_1 \cap K_4$. In each of these two cases, then, it is better for α to perform K_2 rather than K_1, and since these cases exhaust the possibilities, a pattern of reasoning sometimes described as the *sure-thing principle* suggests that K_2 is simply a better action than K_1 for α to perform.[5]

[5]This pattern of reasoning is first explicitly characterized as the "sure-thing principle" in Savage [28], but the principle appears already in some of Savage's earlier work, such as [27, p. 58], where he writes concerning situations of uncertainty that "there is one unquestionably appropriate criterion for preferring some act to some others: If for every possible state, the expected income of one act is never less and is in some cases greater than the corresponding income of another, then the former act is preferable to the latter."

The key to applying sure-thing reasoning in a given situation lies in identifying an appropriate partition of the possible outcomes into a set of states (sometimes called "states of nature" or "conditioning events"), against the background of which the actions available to an agent can then be evaluated through a state-by-state comparison of their results. This is often a difficult task, but we simplify in the current setting, not only by supposing that probabilistic information is unavailable, but also by imagining that the only sources of causality present are the actions of the various agents.

Given these assumptions, it is natural to identify $State_{\alpha}^{m}$—the set of states confronting an agent α at the moment m—with the possible patterns of action that might be performed at that moment by all other agents. In the case of Figure 5, for example, if we assume that α and β are the only two agents—that is, $Agent = \{\alpha, \beta\}$—then $State_{\alpha}^{m}$ can be identified with $Choice_{\beta}^{m}$, the set $\{K_3, K_4\}$ of actions available to β. Although we concentrate here on simple cases like this, with two agents at most, the definition of a state is more general. Where *Agent* contains an arbitrary group of agents, the set of agents other than α is $Agent - \{\alpha\}$, of course, and we can then define the set of states confronting α at m by stipulating that:

$$State_{\alpha}^{m} = Choice_{Agent-\{\alpha\}}^{m}.$$

Having characterized the states facing an agent, we can now define a dominance ordering on the actions available to the agent through a state-by-state comparison of their results; and as an initial step, we must first specify a standard for comparing the possible results of two actions against the background of a particular state. The example depicted in Figure 5 is deceptively simple in this regard, for in this situation, once a particular state from $State_{\alpha}^{m}$ is fixed, each action available to α then determines a unique outcome, so that these actions can simply be ranked along with their outcomes.

In the more general case, of course, even against the background of a fixed state, the actions available to an agent may determine only sets of outcomes, or propositions, rather than unique outcomes—but here, we can nevertheless compare the results of different actions in a state by appealing to the preference ordering defined earlier on propositions. Where S is a state belonging to $State^m_\alpha$, and where K and K' are actions available to α at m, we can say that the results of K' are at least as good as those of K in the state S whenever the proposition $K' \cap S$, determined by performing the action K' in the state S, is weakly preferred to the proposition $K \cap S$, determined by performing K in S, and likewise, that the results of K' are better than those of K in S whenever the proposition $K' \cap S$ is strongly preferred to the proposition $K \cap S$.

With these various concepts in place, we are now in a position to define a dominance ordering on the actions available to an agent at a moment.

> Dominance ordering on actions: Let α be an agent and m a moment, and let K and K' be members of $Choice^m_\alpha$. Then $K \preceq K'$ (K' *weakly dominates* K) if and only if $K \cap S \leq K' \cap S$ for each state $S \in State^m_\alpha$; and $K \prec K'$ (K' *strongly dominates* K) if and only if $K \preceq K'$ and it is not the case that $K' \preceq K$.

The idea is that, K' weakly dominates K, then the results of performing K' are at least as good as those of performing K in every state, so that, no matter which state is realized, the agent is sure to do at least as well with K' as with K. If K' strongly dominates K, then not only are the results of performing K' at least as good as those of performing K in every state, but there is some state in which K' yields better results, so that the agent is sure to do at least as well with K' as with K, and might do better.

Let us now return to our central question: how, in this indeterminist setting, can we define the utilitarian notion of a right action? The dominance account provides an answer that is both precise and intuitively plausible.

We begin by defining the set $\mathit{Optimal}^m_\alpha$ containing the *optimal actions* available to an agent α at a moment m, those actions available to the agent that are not strongly dominated by any others, as follows:

$$\mathit{Optimal}^m_\alpha = \{K \in \mathit{Choice}^m_\alpha : \neg\exists K' \in \mathit{Choice}^m_\alpha (K \prec K')\}.$$

It is then natural formulate a theory that might be characterized as *dominance act utilitarianism* simply by identifying the right actions available to an agent at a moment with the optimal actions.

> Dominance act utilitarianism: Let α be an agent and m a moment, and suppose $K \in \mathit{Choice}^m_\alpha$. Then the action K is *right* at the moment m if and only if $K \in \mathit{Optimal}^m_\alpha$, and *wrong* otherwise.

The theory can be illustrated by returning to our earlier examples. In the case of Figure 4, we have $\mathit{Optimal}^m_\alpha = \{K_1, K_2\}$, so that both actions available to the agent at the moment m are right. In the case of Figure 5, we have $\mathit{Optimal}^m_\alpha = \{K_2\}$, so that K_2 is right and K_1 is wrong.

4 The orthodox account

4.1 An example

The theory of dominance act utilitarianism is, I suspect, not too surprising. It is, perhaps, even obvious: the underlying idea of dominance and optimality are familiar from decision

theory, generalized only slightly to allow for the fact that an action in a state yields a proposition, rather than a a unique outcome. What may be surprising, however, is the fact that the treatment of utilitarianism within the ethical literature does not follow this dominance account at all, but is based on an entirely different approach, which I will refer to, in deference to the literature, as the *orthodox* account.

In order to illustrate this orthodox account, let us consider an example that has figured prominently in the discussion of different forms of utilitarianism. Although the example was first introduced by Gibbard [10], and was elaborated on shortly thereafter by Sobel [32], we take the later but more extensive discussion by Regan [26] as our primary source:

> Suppose that there are only two agents in the moral universe, called Whiff and Poof. Each has a button in front of him which he can push or not. If both Whiff and Poof push their buttons, the consequences will be such that the overall state of the world has a value of ten units. If neither Whiff nor Poof pushes his button, the consequences will be such that the overall state of the world has a value of 6 units. Finally, if one and only one of the pair pushes his button (and it does not matter who pushes and who does not), the consequences will be such that the overall state of the world has a value of 0 (zero) units. Neither agent, we assume, is in a position to influence the other's choice. [26, p. 19]

In the present framework, this example can be depicted as in Figure 6, where α represents Whiff, β represents Poof, and m is the moment at which each of these two agents must choose whether or not to push his button.[6] The action K_1 represents Whiff's option of pushing its

[6]Regan does not actually require that these choices must be simultaneous (though simultaneity is part of Gibbard's earlier description), but he does require the choices to be independent, and we guarantee independence through simultaneity.

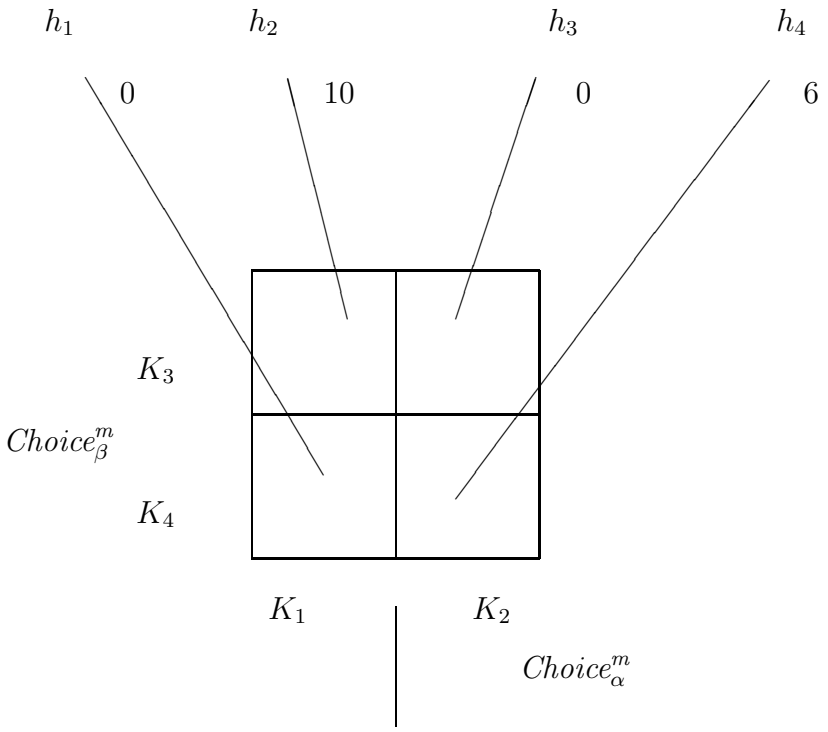

Figure 6: Whiff and Poof

button, and K_2 the option of refraining; likewise, K_3 and K_4 represent Poof's options of pushing or refraining; and the possible outcomes resulting from the choices by these agents are represented by the histories h_1 through h_4, which are assigned the values indicated in Regan's description.

Now, when the example is set out in this way, it is easy to see that both agents must satisfy our previous theory of dominance act utilitarianism no matter what they do. Neither action available to either agent is dominated, and so we have both $Optimal^m_\alpha = \{K_1, K_2\}$ and $Optimal^m_\beta = \{K_3, K_4\}$. Since both of the actions K_1 and K_2 available to Whiff are optimal, both are right according to the dominance theory; and both of the actions K_3 and K_4 available to Poof are right as well.

The theory of dominance act utilitarianism, then, yields results that are at least definite in this case, even if not particularly constraining: each of the two agents can satisfy the theory

by selecting either of the available actions. However, Regan's own conclusions—based on his own theory of act utilitarianism or, as he calls it, AU—are strikingly different:

> Now, if we ask what AU directs Whiff to do, we find that we cannot say. If Poof pushes, then AU directs Whiff to push. If Poof does not push, then AU directs Whiff not to push. Until we specify how Poof behaves, AU gives Whiff no clear direction. The same is true, *mutatis mutandis*, of Poof. [26, p. 18]

In saying that act utilitarianism gives Whiff no clear direction, Regan does not mean only that this theory, like the dominance theory, classifies multiple actions as right, allowing the agents to choose among them. Instead, he means that, on the basis only of the information provided so far, the theory is simply unable to generate any results at all: no actions can be classified either as right or as wrong. In order to arrive at a situation in which act utilitarianism is able to yield definite results, Regan feels that it is necessary to supplement the description of the example provided so far, and depicted in Figure 6, with additional information concerning the actions actually performed by the individuals involved:

> If we shift our attention to patterns of behavior for the pair, we can decide whether each agent satisfies AU in any specified pattern. [26, p. 18]

And he illustrates the kind of reasoning allowed by this additional information as follows:

> Suppose, for example, Whiff and Poof both push their buttons. The total value thereby achieved is ten units. Does Whiff satisfy AU? Yes. The only other thing he might do is not push his button. But under the circumstances, which include the fact that Poof pushes his button, Whiff's not pushing would result in a total utility of zero. Therefore Whiff's pushing his button has at least as

good consequences as any other action available to him under the circumstances. Therefore, it is right according to AU. [26, pp. 18–19]

4.2 Orthodox act utilitarianism

Evidently, Regan is unwilling to classify the actions available to Whiff and Poof as either right or wrong absolutely, but only as right or wrong under conditions determined by the actions of the other, that is, by the states in which these agents find themselves.[7] It therefore seems natural to represent the theory of act utilitarianism guiding Regan's judgments by first introducing a notion of conditional optimality.

We introduce such a notion in three steps. First, taking X as a proposition, the set of actions available to an agent α at m under the condition that X holds—expressed here as $Choice^m_\alpha/X$—is simply the set containing those actions open to α at m that are consistent with X:

$$Choice^m_\alpha/X = \{K \in Choice^m_\alpha : K \cap X \neq \emptyset\}.$$

The next step is to generalize our earlier treatment of dominance to include conditional dominance.

Conditional dominance ordering on actions: Let α be an agent and m a moment, and let K and K' be members of $Choice^m_\alpha$, and X a proposition.

[7] Gibbard adopts a similar viewpoint in his original discussion of this example, evaluating each agent's selection only under an assumption about the action selected by the other [10, p. 215]. And Sobel defends Gibbard's strategy as follows: "It is perhaps natural to feel that Gibbard's first case is objectionable just because it includes assumptions concerning what agents will and would do. But this can be no objection since it is obvious that such assumptions are essential to the application of AU; without such assumptions the dictates of AU could not be determined ..." [32, p. 152].

Then $K \preceq_X K'$ (K' *weakly dominates* K *under the condition* X) if and only if $K \cap X \cap S \leq K' \cap X \cap S$ for each state $S \in State^m_\alpha$; and $K \prec_X K'$ (K' *strongly dominates* K *under the condition* X) if and only if $K \preceq_X K'$ and it is not the case that $K' \preceq_X K$.

This conditional analysis follows the pattern of the absolute treatment set out earlier, except that, in comparing the results of two actions K and K' in a given state S, our attention is now restricted only to those outcomes that are consistent with the background proposition X.

And finally, having generalized both choice and dominance to the conditional setting, we can now combine these ideas to arrive at a concept of conditional optimality. Again taking X as a proposition, we define the set of optimal actions available to α at m under the condition X—expressed as $Optimal^m_\alpha/X$—to be the set of those actions available to α at m under the condition X that are not strongly dominated under this condition by any other such action:

$$Optimal^m_\alpha/X = \{K \in Choice^m_\alpha/X : \neg\exists K' \in Choice^m_\alpha/X\,(K \prec_X K')\}.$$

It is worth noting explicitly that the conditional notions of choice, dominance, and optimality introduced here are, in fact, generalizations of our earlier concepts. When the background condition X is identified with the trivial proposition H_m—that is, $H_m \subseteq X$—each of these three conditional notions coincides with its absolute counterpart. In particular, we have:

$$Optimal^m_\alpha/H_m = Optimal^m_\alpha.$$

Now that the notion of conditional optimality has been introduced, it remains only to define the propositions on which we conditionalize.[8] Just as $Choice^m_\alpha/X$ represents the set

[8]These definitions may seems to be needlessly general, but please bear with me; the generality will help us later.

of actions available to α at m that are consistent with X, we can likewise define

$$\mathit{State}^m_\alpha/X = \{K \in \mathit{State}^m_\alpha : K \cap X \neq \emptyset\}$$

as the set of states confronting α at m that are consistent with X. And in this case, it is also convenient to represent the proposition formed by taking the union of these states—the proposition, that is, according to which one of these states holds—written $\mathit{Choice}^m_\alpha(X)$ and defined as follows:

$$\mathit{State}^m_\alpha(X) = \bigcup \mathit{State}^m_\alpha/X.$$

In the special case in which $X = \{h\}$ is a maximally specific proposition, containing only a single history, we write $\mathit{State}^m_\alpha/h$ and $\mathit{State}^m_\alpha(h)$ for convenience; and here, $\mathit{State}^m_\alpha/h$ is a unit set containing the unique state consistent with that history, and $\mathit{State}^m_\alpha(h)$ is simply this unique state itself.

With these concepts before us, we can now define a form of act utilitarianism designed to model the orthodox notion found in the work of Gibbard, Sobel, Regan, and others.

> Orthodox act utilitarianism: Let α be an agent and m a moment, and suppose $K \in \mathit{Choice}^m_\alpha$. Then the action K is *right* at the index m/h if and only if $K \in \mathit{Optimal}^m_\alpha/\mathit{State}^m_\alpha(h)$, and *wrong* otherwise.

What the definition tells us, then, is simply that the action K is right at the index m/h whenever K is optimal under the condition specified by the state containing the history h.

This version of act utilitarianism can be illustrated by returning to the Whiff and Poof example, Figure 6, and considering, for example, the index m/h_2, where both Whiff and Poof push their buttons. At this index, the situation confronting Whiff, determined by Poof's action, is K_3; that is, $\mathit{State}^m_\alpha(h_2) = K_3$. We therefore have $\mathit{Optimal}^m_\alpha/\mathit{State}^m_\alpha(h_2) =$

$Optimal^m_\alpha/K_3$. And it is easy to verify also that $Optimal^m_\alpha/K_3 = \{K_1\}$, so that the action K_1 is classified as right at m/h_2. In the same way, however, we can see that $Optimal^m_\alpha(h_1) = \{K_2\}$, so that the action K_1 is classified as wrong at the index m/h_1.

As this example shows, the orthodox classification of actions as right or wrong—in contrast to the dominance account— depends on a full index, not just a moment. Here, the same action, K_1, is classified as right at the index m/h_2 but wrong at the index m/h_1; although Whiff performs the same action at each of these two indices, this agent satisfies orthodox act utilitarianism at the first, performing an action that is classified as right, but not at the second. It is as Regan says: we cannot define which of an agent's actions are right or wrong until we know the state confronting that agent, the actions performed by the other agents involved.

5 Double time reference

5.1 A problem

We now have two accounts of right action before us, orthodox and dominance. In order to compare these accounts, I now want to introduce another example, which I have found to be especially helpful in highlighting their differences.[9] Imagine that two drivers are traveling toward each other on a one-lane road, with no time to stop or communicate, and with a single moment at which each must choose, independently, either to swerve or to continue along the road. There is only one direction in which the drivers might swerve, and so a collision can be avoided only if one of the drivers swerves and the other does not; if neither

[9]The example is due to Goldman [11], but also discussed by Humberstone in [15], a paper that sets out in a different context some of the fundamental ideas underlying the orthodox account.

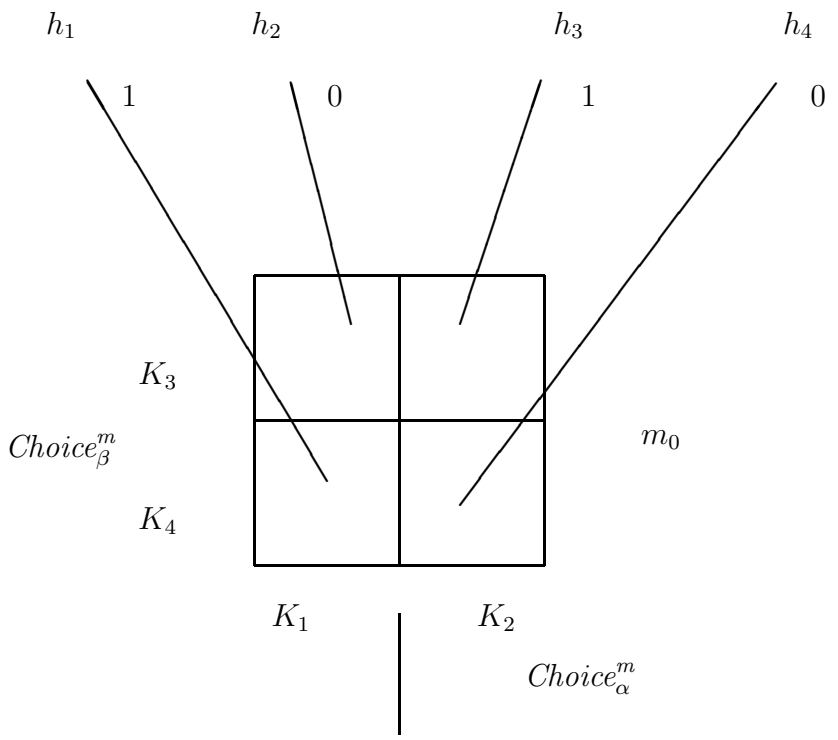

Figure 7: The driving example

swerves, or both do, a collision occurs. This example is depicted in Figure 7, where α and β represent the two drivers, K_1 and K_2 represent the actions available to α of swerving or continuing along the road, K_3 and K_4 likewise represent the swerving or continuing actions available to β, and m represents the moment at which α and β must make their choice. The histories h_1 and h_3 are the ideal outcomes, resulting when one driver swerves and the other does not; collision is avoided. The histories h_2 and h_4, resulting either when both drivers swerve or both continue along the road, represent nonideal outcomes; collision occurs.

Now imagine that what actually happens is that both agents continue along the road, so that the resulting outcome is the history h_4, in which there is a collision. Suppose that, looking back at the situation from some later moment belonging to h_4—perhaps while recovering in its hospital bed—the agent α says to itself: I performed the wrong action; it

would have been right to swerve. And let us ask: is what the agent says correct, or not? The answer, I think, is that we can legitimately understand this statement either as correct or as incorrect, and that the contrast between these two different readings can be captured by appeal to our distinction between the orthodox and dominance accounts of right action.

On the one hand, it is clear from the standpoint of the later moment that, if the agent had swerved, there would have been no collision. The agent would not now be in the hospital recovering from its injuries, and would surely be better off. Therefore, the agent was wrong not to swerve; it would have been right to serve. This way of understanding the agent's statement is captured by the orthodox account, which classifies K_1 as right and K_2 as wrong as the index m/h_4. On the other hand, the past tense in the agent's statement, uttered at some later moment along the history h_4, refers back to the earlier moment m; and it is hard to see how we could have said at this earlier moment—with the four histories each lying ahead as future possibilities—that it would be right for the agent to swerve and wrong not to. But then, if we could not have said at the time that it was wrong not to swerve, how can we say that now? How can it be wrong to have done something when, at the time, it was not wrong to do it? This way of understanding the agent's statement is captured by the dominance account, which classifies both K_1 and K_2 as right as the moment m.

Although I do not have space (or time) to justify this claim here, I believe that the contrast apparent in this example between the two different ways of evaluating the agent's action can be seen as underlying many of the debates in utilitarian theory that commanded so much attention during the 1970's and 1980's—particularly the debates over the "actualist" and "possibilist" positions regarding the relations between an agent's present obligations and future choices.[10] The current proposal has the benefit, then, of providing a rigorous

[10]This problem was originally presented in a trio of papers: Goldman [11], Sobel [33], and Thomason [34].

explication of these two different ways of understanding our normative evaluation both in the driving example presented here and also, I believe, in other cases from the literature on utilitarian theory as well.

But this benefit comes with a cost. The cost is that, while the proposal does succeed in explicating this contrast, it does so only by postulating a semantic ambiguity in words that do not really seem to be ambiguous at all. In the driving case, we are not forced into the artificial position of classifying the agent's action either as unequivocally right or as unequivocally wrong, ignoring the pull of the opposing account. And we are not forced to contradict ourselves by describing the action as both right and wrong; we can say, instead, that the action was wrong in the orthodox sense but right in the dominance sense. In doing so, however, we are relying on the assumption that these words, "right" and "wrong," carry two different senses, two different meanings. Is this plausible? Does is really seem that these words are *semantically* ambiguous?

I now want to show that there is a better way. We can preserve the benefits of the account presented here, allowing appeal to both the orthodox and dominance perspectives in evaluating an agent's actions, without postulating semantic ambiguity, by relying instead on a pragmatic difference.

5.2 Perspectival act utilitarianism

The idea is that actions are to be evaluated as right or wrong, not just at a moment, but at one moment from the perspective of another. As far as I know, the key component of this

Further discussion can be found, for example, in Bergström [5], Carlson [6], Feldman [8], Goldman [12], Greenspan [13], Humberstone [15], Jackson[16], Jackson [17], Jackson and Pargetter [18], McKinsey [21], and Zimmerman [35].

idea—the appeal to "double time reference"—was first set out systematically by Belnap [1], with an emphasis on the assessment of speech acts, particularly the speech act of assertion.[11] It was later developed in a somewhat different way by MacFarlane [20], who is concerned with the role of perspective in the assessment of a statement's content: what is said, rather than the act of saying it.

Let us take m as the *moment of action* and m' as the moment from which the action selected at m is appraised—the *moment of appraisal*, which we can sensibly assume to be comparable to m in the treelike ordering of moments: either later than, earlier than, or identical with m. In that case, $\mathit{State}^m_\alpha/H_{m'}$ can be thought of as the set of states confronting the agent at m, as judged from the standpoint of m'. And as we have seen, $\mathit{State}^m_\alpha(H_{m'})$ is simply the proposition that one of these states holds. What perspectival utilitarianism tells us, then, is that an action available to α at the m of action is right from the standpoint of the moment m' of appraisal just in case the action is optimal given the states that the agent is confronting at m, as judged from the standpoint of m'.

> Perspectival act utilitarianism: Let α be an agent and m and m' moments such that either $m < m'$ or $m' < m$ or $m = m'$, and suppose $K \in \mathit{Choice}^m_\alpha$. Then the action K is *right* at m from the standpoint of m' if and only if $K \in \mathit{Optimal}^m_\alpha/\mathit{State}^m_\alpha(H_{m'})$, and *wrong* otherwise.

This new account subsumes the orthodox account, as we can see by returning to the driving example, again supposing that neither driver swerved, the crash occurs, and the agent α is looking back on the incident from some the standpoint of some later moment lying on the history h_4. Let m_1 be this later moment. We can the see that $\mathit{State}^m_\alpha(H_{m_1}) = K_4$,

[11] Further discussion can be found at various points throughout Belnap et al. [3] (see index entries under "double time reference"), and an informal presentation appears in Belnap [2].

and also that $Optimal^m_\alpha/K_4 = \{K_1\}$, so that from the standpoint of m_1, the agent can reach the orthodox judgment that the action K_2 was wrong and K_1 would have been right. On the other hand, suppose that, at the crucial moment, both drivers swerve, another crash occurs, things proceed along the history h_2, and α is now looking back at the event from some later moment, say m_2, belonging to this history. Then since $State^m_\alpha(H_{m_2}) = K_3$ and $Optimal^m_\alpha/K_3 = \{K_2\}$, the judgment is reversed: from the standpoint of m_2, the action K_1 was wrong and K_2 would have been right. Two different perspectives, two different judgments: the appraisal of an action varies with the standpoint from which it is appraised.

Our new account, then, allows us to recover the orthodox account, but interestingly, it subsumes the dominance account as well. Suppose the moment of appraisal m' is identical with the moment m of action, or indeed, earlier: $m' = m$ or $m' < m$. Then it is easy to see, since each member of $State^m_\alpha$ contains some history from $H_{m'}$, that $State^m_\alpha/H_{m'} = State^m_\alpha$, and therefore, since $State^m_\alpha$ partitions the set H_m, that $State^m_\alpha(H_{m'}) = H_m$. As noted earlier, the set $Optimal^m_\alpha/H_m$, containing those actions that are optimal at m given the trivial proposition, coincides with the set $Optimal^m_\alpha$ itself. It therefore follows that

$$\begin{aligned} Optimal^m_\alpha/State^m_\alpha(H_m) &= Optimal^m_\alpha/H_m \\ &= Optimal^m_\alpha, \end{aligned}$$

so that the set of actions at m that are right from the standpoint of m' according to the perspectival account coincides with the set of actions that are right according to the dominance account. In the case of Figure 7, for example, we have $Optimal^m_\alpha/State^m_\alpha(H_{m'}) = \{K_1, K_2\}$: when the actions available to α at m are appraised from the standpoint of m itself, or an earlier moment, both actions are right.

To sum up: The theory of perspectival act utilitarianism set out here allows us see how we can say, in the driving example, for instance, that the agent's actions at the crucial

moment might legitimately be viewed as both right and wrong. But it does not do so by postulating two separate senses of the words "right" and "wrong"—an orthodox and a dominance sense—captured by two separate utilitarian theories. Instead, the explanation of the different judgments is pragmatic, reflecting different relations between the moment at which the agent is forced to select an action and the moment from which the available actions are appraised. If the moment if appraisal is strictly later than the moment of action, then the perspectival theory agrees with orthodox act utilitarianism. But if the appraisal takes place at the very moment of action, or earlier, then the perspectival theory agrees with dominance act utilitarianism. The difference between our orthodox and dominance intuitions is not a substantial difference that needs to be explained by postulating two separate utilitirian theories, but only a matter of perspective.

References

[1] Nuel Belnap. Double time references: speech act reports as modalities in an indeterministic setting. In F. Wolter, H. Wansing, M. de Rijke, and M. Zakharyaschev, editors, *Advances in Modal Logic, Volume 3*, pages 1–21. CSLI Publications, 2001.

[2] Nuel Belnap. Future contingents and the battle tomorrow. Manuscript, Philosophy Department, University of Pittsburgh, 2004.

[3] Nuel Belnap, Michael Perloff, and Ming Xu. *Facing the Future: Agents and Choices in Our Indeterministic World*. Oxford University Press, 2001.

[4] Lars Bergström. *The Alternatives and Consequences of Actions*, volume 4 of *Stockholm Studies in Philosophy*. Almqvist and Wiksell, 1966.

[5] Lars Bergström. Utilitarianism and future mistakes. *Theoria*, 43:84–102, 1977.

[6] Erik Carlson. *Consequentialism Reconsidered*, volume 20 of *Theory and Decision Library, Series A: Philosophy and Methodology of the Social Sciences*. Kluwer Academic Publishers, 1995.

[7] Daniel Ellsberg. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75:643–669, 1961.

[8] Fred Feldman. *Doing the Best We Can: An Essay in Informal Deontic Logic*. D. Reidel Publishing Company, 1986.

[9] Peter Gärdenfors and Nils-Eric Sahlin, editors. *Decision, Probability, and Utility: Selected Readings*. Cambridge University Press, 1988.

[10] Allan Gibbard. Rule-utilitarianism: merely an illusory alternative? *Australasian Journal of Philosophy*, 43:211–220, 1965.

[11] Holly Goldman. Dated rightness and moral imperfection. *The Philosophical Review*, 85:449–487, 1976.

[12] Holly Goldman. Doing the best one can. In A. I. Goldman and J. Kim, editors, *Values and Morals*, pages 185–214. D. Reidel Publishing Compamy, 1978.

[13] Patricia Greenspan. Oughts and determinism: a response to Goldman. *Philosophical Review*, pages 77–83, 1978.

[14] John Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

[15] I. L. Humberstone. The background of circumstances. *Pacific Philosophical Quarterly*, 64:19–34, 1983.

[16] Frank Jackson. On the semantics and logic of obligation. *Mind*, 94:177–195, 1985.

[17] Frank Jackson. Understanding the logic of obligation. In *Proceedings of the Aristotelian Society, Supplementary Volume 62*. Harrison and Sons, 1988.

[18] Frank Jackson and Robert Pargetter. Oughts, options, and actualism. *Philosophical Review*, 99:233–255, 1986.

[19] R. Duncan Luce and Howard Raiffa. *Games and Decisions*. John Wiley and Sons, 1957.

[20] John MacFarlane. Future contingents and relative truth. *Philosophical Quarterly*, 53:321–336, 2003.

[21] Michael McKinsey. Levels of obligation. *Philosophical Studies*, 35:385–395, 1979.

[22] G. E. Moore. *Ethics*. Oxford University Press, 1912.

[23] Arthur Prior. The consequences of actions. In *Proceedings of the Aristotelian Society, Supplementary Volume 30*. Harrison and Sons, 1956.

[24] Arthur Prior. *Past, Present, and Future*. Oxford University Press, 1967.

[25] Frank Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and Other Logical Essays*, pages 156–191. Routledge and Kegan Paul, 1931. Originally published in 1926.

[26] Donald Regan. *Utilitarianism and Co-operation*. Clarendon Press, 1980.

[27] Leonard Savage. The theory of statistical decision. *Journal of the American Statistics Association*, 46:55–67, 1951.

[28] Leonard Savage. *The Foundations of Statistics*. John Wiley and Sons, 1954. Second revised edition published by Dover Publications, 1972.

[29] Krister Segerberg. The logic of deliberate action. *Journal of Philosophical Logic*, 11, 1982.

[30] Krister Segerberg. Bringing it about. *Journal of Philosophical Logic*, 18:327–347, 1989.

[31] Krister Segerberg. Outline of a logic of action. In F. Wolter, H. Wansing, M. de Rijke, and M. Zakharyaschev, editors, *Advances in Modal Logic, Volume 3*. CSLI Publications, 2002.

[32] J. Howard Sobel. Rule-utilitarianism. *Australasian Journal of Philosophy*, 46:146–165, 1968.

[33] J. Howard Sobel. Utilitarianism and past and future mistakes. *Nous*, 10:195–219, 1976.

[34] Richmond Thomason. Deontic logic and the role of freedom in moral deliberation. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 177–186. D. Reidel Publishing Company, 1981.

[35] Michael Zimmerman. Where did I go wrong? *Philosophical Studies*, 59:55–77, 1990.