Noûs

NOÛS 44:3 (2010) 503–529

Modelling the Moral Dimension of Decisions

Mark Colyvan
University of Sydney

Damian Cox
Bond University

Katie Steele
University of Sydney

Abstract

In this paper we explore the connections between ethics and decision theory. In particular, we consider the question of whether decision theory carries with it a bias towards consequentialist ethical theories. We argue that there are plausible versions of the other ethical theories that can be accommodated by "standard" decision theory, but there are also variations of these ethical theories that are less easily accommodated. So while "standard" decision theory is not exclusively consequentialist, it is not necessarily ethically neutral. Moreover, even if our decision-theoretic models get the right answers vis-à-vis morally correct action, the question remains as to whether the motivation for the non-consequentialist theories and the psychological processes of the agents who subscribe to those ethical theories are lost or poorly represented in the resulting models.

1. Introduction

Decision theory has two components: probabilities and utilities. From the formal point of view, these two components play symmetrical roles in decision theory. For each act–state pair (or outcome) we assign a probability and a utility, then we multiply these together and sum the products across each act. The resulting sum we call the *expected utility* of the act. Standard decision theory then tells us to choose the act with the greatest expected utility (if there is such an act). This is all very familiar. What we draw attention to is that, in terms of the formal decision calculus, probabilities and utilities are treated similarly—they are just two real numbers to be multiplied and then added.

The similarities between probabilities and utilities in the standard decision model run even deeper. Both represent propositional attitudes that are held to be constrained by axiomatic theories: probabilities represent partial beliefs which are thus constrained by the Kolmogorov (1956) axioms, and utilities represent preferences which are axiomatised by von Neumann-Morgenstern (1944) utility theory. These axiomatic theories place minimal structural constraints on beliefs and preferences.[1] The axioms ensure that there are no violations of consistency. For example, the Kolmogorov axioms rule out probabilistic beliefs that don't sum to 1, so if an agent assigns probability p to some proposition Q, then the agent must assign $1–p$ to $\neg Q$. And the von Neumann-Morgenstern axioms insist, for instance, on transitivity of preferences: if an agent prefers A to B and B to C, then the agent ought to prefer A to C. Call an agent whose partial beliefs conform to the Kolmogorov axioms a *Kolmogorov-consistent agent*, and call an agent whose preferences conform to the von Neumann-Morgenstern utility theory a *von-Neumann-Morgenstern-consistent agent*. It is clear that mere Kolmogorov-consistency and von-Neumann-Morgenstern-consistency are, in general, not enough to ensure that an agent is beyond reproach. Compliance with these conditions does not imply responsible decision-making.

Take probabilities first. Consider an agent who assigns probability 1 to the flat earth theory and probability 0 to every other theory about the shape of the earth. If such an agent also obeys Kolmogorov's axioms in all other respects then we can say she is Kolmogorov-consistent, and yet she is a poor or irresponsible *epistemic agent*, for she assigns zero probability to a theory for which there is a great deal of evidence (the roughly-spherical earth theory). Moreover, because she assigns probability 1 to the flat earth theory, no amount of evidence will change this assignment, since updating on new evidence via conditionalisation will never lower the probability of maximal-probability propositions. What are the extra constraints we need to ensure that merely Kolmogorov-consistent agents are responsible epistemic agents? This is going to be a complicated epistemological story, and the details will depend on the particular epistemology to which you subscribe.

Now to utilities. Consider an agent who prefers genocide to murder and prefers murder to a walk in the hills. So long as this agent satisfies transitivity (i.e., prefers genocide to a walk in the hills) and other such structural constraints, the agent is a von-Neumann-Morgenstern-consistent agent. But clearly such an agent is a poor *moral agent*. What are the extra constraints we need to ensure that merely von-Neumann-Morgenstern-consistent agents are responsible moral agents? This is going to be a complicated story about ethics, and, presumably, the details will depend on the particular ethical theory to which you subscribe.

So while decision theory is able to gloss over the details of how probabilities and utilities are assigned (and hence gloss over the thorny issues in epistemology and ethics), a full account of decisions requires both an epistemological and an ethical theory. Moreover, we need the epistemological and ethical theories to be spelled out in ways that enable them to be accommodated in a decision-theory framework. (From what we have said so far it is clear that we are here aiming for the "standard" decision-theory framework, something we will say a bit more about shortly.) Of course a great deal has been written on the epistemic side of this story. Some prominent rules have been put forth that place further constraints on rational belief. The Principle of Indifference (see Hájek 2003a) and Lewis's (1980) Principal Principle are two such examples. But, given the symmetry between the probability side of things and the utility side, it is somewhat surprising that very little has been written about the ethical side of decision theory. This, then, brings us to the central question we will address in this paper: how much freedom do we have in choosing an ethical theory to accompany standard decision theory? It is natural to think that decision theory models a specifically consequentialist way of thinking, with its focus on how much an agent values the possible outcomes associated with an act. But is consequentialism the only game in town, where decision theory is concerned? We investigate whether alternative ethical theories can be at least accommodated in the decision-theory framework. In other words, is a von Neumann-Morgenstern utility function capable of representing the preferences of agents who subscribe to ethical theories other than consequentialism?

Before we move on we should say a little about the significance of what we are trying to do and why we approach it the way we do. In relation to the significance of our project, it might be thought that it is trivial or at least well known that various ethical theories can be modelled in the decision-theory framework.[2] Even if this is right, it is still interesting to see at least some of the details, rather than rest content with the simple knowledge that it can be done. And in fact our project reveals that there are many more complications with modelling the main ethical theories than might first appear to be the case. Moreover, in providing the details, we shed some light on the ethical theories, the decision-theory framework, and the relationship between the two. The question arises as to whether it is appropriate or useful to merely accommodate a wide range of ethical theories within decision theory, or whether we do too much violence to non-consequentialist theories in trying to submit them to this framework. We will return to these issues in the final section.

We set ourselves the task of trying to accommodate the ethical theories in question into *standard* decision theory.[3] Another option might be to modify the standard decision-theory model to introduce, say, two-dimensional utility functions (e.g., Levi 1986, Hájek 2003). While there are no doubt some interesting avenues to be explored here, as well as some very good arguments for exploring them, for the most part we try to restrict ourselves to the standard model. That is, we seek moral constraints on an agent's utility function that are supposed to *supplement* the original von Neumann-Morgenstern axioms.[4] There are several reasons for sticking with the standard model but the most significant is simply that this framework remains the orthodox decision theory, and furthermore, it is thought to underpin other important philosophical positions such as *probabilism* or *Bayesianism*. Moreover, in at least some of its guises, expected utility theory is considered to receive justification (as *the* model of rational choice) from a representation theorem that depends only on ordinal preferences (see, for instance, Savage 1954 or Jeffrey 1983).

Finally, we note that we will be modelling rather specific versions of the ethical theories in question. For instance, we select a particular brand of utilitarianism as our consequentialist representative, and likewise, we run with a particular kind of deontology and virtue theory. We do this because we do not want to delve too far into the debate regarding the essential differences between the various classes of ethical theory (for instance, what distinguishes a consequentialist from a deontological theory). Our modelling task is difficult enough as it is. Of course, even though we pursue specific versions of the main theories, there will still be many details left for the ethicist to fill in. We do not want to take too many stands on substantive issues that go beyond the basic structure of the theories under consideration. In a nutshell, we want to provide some general rules for accommodating arguably quintessential versions of the major ethical theories into standard decision theory.

2. Three Ethical Theories

We examine three of the most important types of theory in contemporary philosophical ethics: utilitarianism[5], deontology and virtue theory. Ethical theories are primarily theories of right action. Although virtue ethics is often taken to be an exception, we treat it here as furnishing a theory of right action. Our challenge is to translate theories of right action into a decision-theoretic framework by representing them as offering diverse accounts of utility. First, however, we outline our general approach to the theories and what we regard as distinctive of each of them.

Utilitarianism provides us with the most natural method of generating utility assignments because it expressly incorporates a theory of utility, one that identifies utility with welfare. Welfare is conceptualized by utilitarians in markedly different ways. For example, classical utilitarianism—the theory associated with Jeremy Bentham and John Stuart Mill—identifies welfare with happiness, where the latter is interpreted as a preponderance of pleasure over pain. Preference-satisfaction utilitarianism, by contrast, identifies welfare with the satisfaction of self-regarding preferences (i.e., with a person's preferences about how their life is to go).[6] More objective conceptualizations of welfare have also been proposed in which a mixed basket of objective life conditions are said to contribute in some reasonably well defined way to a person's overall welfare. The important point for our purposes is that every utilitarian theory proposes an account of utility as welfare and introduces a cardinal measure of total utility in a situation. Now, utility assignments of this kind are widely dismissed as too controversial and probably undiscoverable, and for these reasons do not figure much in economic theory and public policy. However, for philosophical purposes, one might nonetheless insist that the notion of a cardinal measure of welfare is coherent and plays a role in determining right action, even though we are usually in a position only to approximately and fallibly identify levels of welfare.

An important feature of utilitarianism is the interpretation of the nature of value it presupposes. For utilitarians, values are attributed to possible states of affairs and furnish reasons to bring them about. Welfare-tracking utility is the sole measure of morally relevant value for utilitarians and they are naturally led towards a maximizing principle with respect to it. To act in a way that promotes something other than maximum expected utility would be to value something other than (welfare-tracking) utility more than one values utility. But since utilitarians recognize no morally relevant value other than utility they generally regard it as a moral error to pursue less than the maximum available expected utility.[7]

Deontologists differ from utilitarians in that they do not offer a theory of outcome utility at all. Rather, deontology introduces a set of moral constraints upon decision-making. These include prohibitions and obligations that often have the effect of undercutting welfare maximization.[8] For example, observing a prohibition against targeting civilians in war may prolong a bloody conflict so that, on any reasonable assessment of welfare, general welfare is greatly diminished as a result.

Prohibitions and obligations need not be absolute; they may be conditional. For example, a deontologist may allow lying under some circumstances—say when an agent is responding to another person who is negotiating in bad faith—and not others. A deontologist might also take the application of duties to be context-sensitive: the duties relevant in one situation may not be relevant in another. For example, a parent may have duties to their child that a stranger does not have. It is also possible for a deontologist to hold certain duties to be defeasible. This is the view that prohibitions and obligations may be defeated by the threat of certain amounts of (consequentialist) disutility. For example, a deontologist might think it permissible to lie or steal when the threat to general well-being, or even to their own survival, is sufficiently high and no better alternatives are available. Prohibitions and obligations may also be ranked, so that, for example, prohibitions against killing outrank prohibitions against stealing, both of which outrank obligations to come to the aid of neighbours in distress.

Deontologists are motivated by a conception of the moral relevance of value that is very different from that of the utilitarian.[9] Where the utilitarian conceives of judgments of value as supplying reasons to promote particular states of affairs, the deontologist is likely to think of morally significant value judgments as directed towards persons, morally requiring that persons be respected. For deontologists, morally valuing someone entails respecting them, not seeking to promote their welfare. To respect a person is not to wish to see more of a certain valuable state of affairs, but, at least in part, to accept that we are legitimately restricted in what we may do to or for the person. Deontologists characteristically direct respect towards persons, but other subjects of respect are also possible in deontology. For example, deontological environmental ethicists may value all living things.[10]

The third ethical theory under view—virtue ethics—neither offers a theory of outcome utility nor a set of explicit moral constraints upon action. Virtue ethics is first and foremost a re-focusing of moral theory away from the concern to provide an account of correct moral decision-making. The central question for the virtue ethicist is not 'what should I do?' but 'what kind of person should I be?' and this latter question is not sufficiently well answered by observing that I should be the kind of person who acts rightly. Nonetheless, a number of contemporary virtue ethicists attempt to show how virtue-theoretic considerations contribute directly to the task of moral decision-making. Virtue-theoretic accounts of right action derive an account of right action in one way or another from an independent account of virtue. The idea is to first determine which character traits or motivations are morally admirable (either intrinsically or because they are essential features of a flourishing life or of a morally good life) and to use this account to describe how a morally good person would act in various challenging situations. On one natural development of this idea, right action is identified with virtuous agency.[11] For such virtue ethicists, to act rightly is to act virtuously and to act virtuously is to manifest a complex inner state: a state involving (at least) morally admirable motivations. An alternative virtue-theoretic approach to right action identifies right action with actions that would, characteristically, be performed by a virtuous agent.[12] Such an indirect, or hypothetical, virtue-theoretic account of right action has the advantage of preserving intuitions about the distinction between doing the right thing (acting rightly) and doing the right thing for the right reason and with the right motivation (acting virtuously). By contrast, direct virtue-theoretic accounts—those identifying right action with virtuous action—enjoin moral agents to pursue virtuous action rather than merely conform to the standards of virtuous action.

Virtues interact with each other in complex ways. It is unlikely that any simple ranking of virtues will capture the variable significance of virtues as they apply in complicated and varied circumstances. When does benevolence—a virtue directed at the well-being of strangers—take precedence over the virtue of caring for loved ones? When is the virtue of integrity outranked by the virtues of practicality and willingness to compromise? In answering questions like these, the virtue ethicist may appeal to their version of what it is for the moral agent concerned to live a good life and the role virtues play in this life, or they may appeal directly to intuitions about the comparative significance of virtues. There is very little precise work on the problem and this reflects real imprecision in our picture of a good life and uncertainty about key moral intuitions. The important point for our purposes is that we take the virtue ethicist to propose some account—albeit a gappy and imprecise account—of how virtues interact which furnishes them with an account of what it is for an action to be the most virtuous possible action in a situation. For example, on a particular occasion a person might face the choice of acting with general benevolence or acting with a special concern for loved ones and the virtue ethicists owes us a way of ranking the virtues in this situation (even if they rank the options equally).

In tying right action to virtuous action, either directly or indirectly, the virtue ethicist identifies valued states of affairs that are to be promoted. For the direct virtue ethicist, an agent's own virtuous agency is to be promoted; for the indirect virtue ethicist, an agent's conforming to virtuous action is to be promoted. Thus virtue ethics identifies states of affairs to be promoted, but unlike utilitarian promotion of welfare, virtue-theoretic values are agent-relative (i.e., indexed to agents) and time-relative (i.e., indexed to the immediate choice situation of the agent). Virtue ethicists do not typically hold a moral agent responsible for the impartial promotion of virtuous action; they have it that individual agents have a special responsibility for their own character and they act rightly only when they act virtuously (or act as the virtuous would act).

The three theories outlined here represent fundamentally different approaches to right action. Utilitarians define right action as the achievement of a goal—maximized (welfare-tracking) utility. Deontologists, by contrast, see right action as a matter of complying with rules of conduct; not promoting the compliance of rules, but actually complying with rules in every choice situation. Virtue ethicists might be said to fall somewhere in the middle: they see right action in terms of the achievement of a goal, but this time an immediate and self-oriented goal. In every choice situation, the virtue ethicist's goal is either to manifest virtue in the situation (in direct versions of virtue ethics) or to outwardly conform to virtuous behaviour in the situation (in indirect versions of the theory).

3. Introducing Ethics to Decision Theory

Now we turn to the task of trying to incorporate these three ethical theories into the standard decision-theory framework. As alluded to in the introduction, we are assuming that there is only one place in the *standard* model where ethics can enter the picture, and that is as additional constraints on admissible utility functions. And here it might seem that there is only one way to proceed, for the deck seems to be stacked in favour of utilitarianism. Indeed, even the language ("*utility* functions" and "expected *utility*") suggests that utilitarian ethics is the only real contender here. Accommodating the other two ethical theories is by no means straightforward, but deontology and virtue ethics should not be dismissed simply because decision theory employs a mathematical function called 'a *utility* function', and this is suggestive of utilitarianism. As we hope to show, we can indeed give the other two a fair run in the framework of decision theory, at least in terms of getting the "right" answers to decision problems. In an important sense we will be trying to *model* the ethical theories in question, rather than give formal presentations of the theories themselves. This distinction will become important in the final section. First we turn to the task of determining, for each of the three ethical theories, what type of constraints should be placed on the utility function, so that decision calculations yield morally correct actions.

Let's start with utilitarianism. As we've already mentioned, representing utilitarianism in the decision theory framework is fairly straightforward. Decision theory already considers the value of each act-state pair, and this value is measured on an interval (as opposed to a merely ordinal) scale.[13] The best act is that which has *maximum* expected value or utility. All this sits very well with utilitarianism, but we shouldn't be fooled into thinking that decision theory's utility functions just are the utilitarian's value functions. For a start, nothing in the von Neumann-Morgenstern utility theory rules out a utility function that assigns greatest value to the act-state pairs that are associated with the greatest harm. Before decision theory's utility functions can be considered utilitarian, we must add something like the following constraints. An outcome, O_{ij}, is the result of the agent performing action a_i while the world is in state s_j. (We will use this notation for acts, states, and outcomes throughout the rest of the paper.) There will, in general, be a range of possible outcomes of performing a_i. Two constraints on a utilitarian utility function are:

(U1) If O_{ij} involves greater total welfare than O_{kl}, then any admissible utility function u must be such that $u(O_{ij}) > u(O_{kl})$.

(U2) If O_{ij} involves the same total welfare as O_{kl}, then any admissible utility function u must be such that $u(O_{ij}) = u(O_{kl})$.

These constraints are enough to prevent perverse anti-utilitarian functions that assign greater value to the greatest harm to the greatest number. For instance, they are sufficient for ruling out perverse unethical functions that value genocide over a walk in the hills. There are still substantive details that need to be dealt with, most of which involve the details of the version of utilitarianism to be represented, and, in particular, the account of welfare that is in play. We set these aside, though, for there does not seem to be any serious impediment to this kind of approach to the incorporation of utilitarianism into decision theory. Indeed, adding the above constraints (U1 and U2) is very natural, and something like this is likely to have been in the back of the minds of many decision theorists.

Next consider deontology, and how it might be incorporated into decision theory. The first point to observe is that whereas utilitarians evaluate outcomes wholly in terms of their preferred characterisation of welfare represented by an outcome, a deontologist evaluates morally salient outcomes in terms of the actions that produced them. We can say that in both cases, the subject of evaluation is an outcome of an action, but insist that for the utilitarian, outcomes are individuated with respect to differences in welfare, and for the deontologist, in morally salient situations, the nature of the action that produces a given outcome is essential to the outcome itself. For example, unlike the utilitarian, the deontologist regards a "pleasant and harmonious evening" brought about via the prohibited act of lying as a different outcome to an otherwise identical "pleasant and harmonious evening" brought about via an honest act, even though the same amount of welfare is supposedly produced in each case. In fact, here we will depict duties as not merely *contributing* to the character of the outcomes they produce; according to our model, the nature of a duty more or less *determines* the utility of any outcome resulting from that duty. Indeed, in the first instance, at least, we depict a version of deontology that treats any outcome resulting from the one prohibited or obligatory act as having the same utility.

While our characterisation of outcomes in terms of the acts that produced them might seem somewhat non-standard, this is a widely accepted decision-modelling move. Jeffrey's (1983) decision theory in fact explicitly defines outcomes as a conjunction of an act, a, and a potential state-of-the-world s: the completely specified outcome is $(a \,\&\, s)$.[14] As will become apparent, our initial attempt at formulating constraints on the deontologist's utility function is such that the agent must be indifferent between $(a_D \,\&\, s_1)$ and $(a_D \,\&\, s_2)$ for any duty a_D and states-of-the-world s_1 and s_2. Let us now drop this Jeffrey formalism, however, and return to our original terminology for outcomes. The point is just that we are not doing anything terribly controversial by requiring the outcomes of duties to be classified with reference to the act that produced them.

As mentioned in our characterisation of deontology in Section 2, a deontologist may think of obligations and prohibitions as conditional upon types of situation encountered. For example, a deontologist might consider lying prohibited, except for cases when they think another person is negotiating with the deciding agent in bad faith. The clearest way to introduce this type of conditionality into our model is at the level of act description. A deontologist may say either that lying is prohibited except in cases of bad faith negotiations or they may refine their description of the relevant act/duty. Under such a description, instead of lying *per se* being prohibited conditionally, lying of a certain kind is prohibited unconditionally. If we allow this way of refining act/duty descriptions, we need not worry about a deontologist's utility function having to accommodate duties whose claims on the agent are conditional upon these kinds of situational details.[15]

Here is a preliminary attempt to describe the constraints prohibitions and obligations make on utility functions, specified in terms of types of acts yielding an outcome:

(D1*) If O_{ij} is the result of an (absolutely) prohibited act, then any admissible utility function u must be such that $u(O_{ij}) = -\infty$.

(D2*) If O_{ij} is the result of an (absolutely) obligatory act, then any admissible utility function u must be such that $u(O_{ij}) = +\infty$.

What goes in favour of D1* and D2* is that these constraints make the utility of morally salient outcomes completely dependent on the action that produces them. Moreover, D1* and D2* set obligations (prohibitions) apart from other acts by ensuring that they have absolute maximum (minimum) expected utility.

As one might suspect, the introduction of infinite (dis)utilities into one's decision model has some troubling implications for choice. While infinite (dis)utilities might seem the appropriate way to represent how duties completely trump ordinary acts, the problem is that an infinite (dis)utility completely "swamps" any probability associated with it. For instance, if killing is assigned infinite disutility, then any of my actions that might lead to me killing (no matter how unlikely) will presumably also have infinite disutility; any such action will thus be prohibited. One might initially think that this is not such a bad feature of the model—perhaps the deontologist really does regard any act leading to a chance of killing as a prohibited act, just like the act of killing itself. The problem is that most ordinary agents will attribute some probability, however small, to their killing in the future, no matter what they do now. In fact, as Hájek (2003) nicely points out in his discussion of infinite utilities and Pascal's wager, if we stipulate that an agent's belief function be regular,[16] then they *must* assign non-zero probability to all logical possibilities, and this means to their killing in the future, no matter what act they now pursue. The upshot here (if we ignore obligatory acts for the time being) is that every act would appear to be prohibited.

We can expand on the above problem so that it is made to look even more serious. It might not cripple one's decision theory if every act was prohibited but there were yet ways to distinguish between prohibitions. The effect of using infinite (dis)utilities for prohibitions, however, is to place all acts on a par. The act of killing itself will have just the same expected utility (negative infinity) as some other act that has a mere one-in-a-million chance of killing.

Surely the deontologist does not want us to be indifferent between two such acts.[17] Furthermore, recall that we have been ignoring the possibility of obligatory acts in this discussion. Presumably, any act will have some positive probability of leading to a prohibition and also some positive probability of leading to an obligation. Given the use of infinite (dis)utility in D1* and D2*, this means that the expected utility of all acts will be undefined. Needless to say, this hardly makes for a satisfactory decision theory!

We might seek to resurrect our deontological decision theory, while retaining the use of infinite (dis)utilities, by refining what constitutes a "duty". Perhaps the above problem just goes to show that duties must be both agent *and* time relative. There is nothing ground-breaking in asserting that duties must be agent relative; indeed, this seems almost definitional of a "duty". An agent is not supposed to have the same concern for the acts of another as for their own acts. To give a typical example, if lying is a prohibited act, then under normal circumstances one should not lie to prevent others from lying. If they do not have the same significance as lying oneself, others' acts of lying should not be represented in the decision model as prohibited acts that yield outcomes with infinite disutility. This means that we need not worry about the prohibited/obligatory acts of others "swamping" our own decision problem. In our discussion above, what was causing all the problems was rather the agent's own future actions. The issue was that any ordinary act an agent chooses to perform *now* has some positive probability of leading to them performing any kind of duty *in the future*. It would seem then that we could circumvent the whole swamping problem by simply imposing strict time relativity on the operation of duties. The idea would be that duties pertain only to the present time, and the predicted actions of future time-slices of the agent are treated just like the predicted actions of any other person.

Imposing strict time relativity on the operation of duties might seem intuitively questionable, but one could argue that this is simply in keeping with the deontological insistence that duties—moral rules—are to be complied with. Agents abiding by such an ethical theory do not seek to promote maximal compliance with rules of conduct and they are not prepared to violate rules in order to promote such compliance. Typically, such an agent would not be prepared to kill just in order to prevent future killings (by themselves or by others). Prohibitions and obligations are seen to apply, rather, only to moral agents in their immediate choice situation. Or so the argument might go. Note also that such a deontological model can still capture many intuitions about our responsibility for our future actions by introducing future-oriented duties, such as the duty not to place oneself in moral danger or a prohibition against reckless behaviour.[18]

Unfortunately, even with the time relativity stipulation, the swamping problem associated with infinite (dis)utilities does not completely disappear. A related issue is that our deontological decision theory is not consistent with the von Neumann-Morgenstern preference axioms. In particular, the continuity axiom is violated.[19] Both of these issues revolve around the status of mixed acts.[20] How does the deontologist want to rank a mixed act that could yield (with, say, equal probability) either the satisfaction of an obligation or a morally neutral action? If the obligation is associated with outcomes of infinite utility (as per D1* and D2*), then the mixed act will also have infinite expected utility. In fact, any mixture involving the possible satisfaction of an obligation and the possible performance of a morally neutral act will have infinite expected utility, contrary to continuity. The question is whether this will be palatable to the deontologist.

Perhaps we could accept a deontological decision theory that violates the continuity axiom of EU theory. Of course, this would amount to not merely supplementing the von Neumann-Morgenstern axioms with extra deontological constraints; it is rather a departure from von Neumann and Morgenstern's theory. But there is scope to argue that such a departure is very minimal and entirely reasonable. Indeed, a number of authors argue that continuity is not so much an ideal of rationality, but rather a mathematical idealisation that permits a more elegant representation of an agent's preferences.[21] Moreover, there are arguably plausible preference structures that violate continuity as a matter of principle—consider any lexical ordering that privileges certain types of goods above others.[22]

Regardless of the status of continuity, however, there is still the issue raised above about whether D1* and D2* adequately reflect how a deontologist wants to deal with mixed acts. We may avoid the kind of swamping problem that results in *all* acts having undefined expected utility, but under D1* and D2*, mixed acts will nonetheless have some unusual properties. For instance, any mixed act involving both an obligation and a prohibition will have undefined expected utility. More seriously, the agent is directed to be indifferent between an obligation/prohibition and any mixed act that assigns only miniscule probability to that obligation/prohibition (assuming the other acts in the mixture are ordinary non-moral acts). The deontologist needs to consider whether or not this should be a feature of their theory.

As well as any problems with infinite utilities, a significant problem with our way of modelling deontology so far is that all prohibited (or obligatory) acts are on a par. Murder, if it is prohibited, is no better or worse than genocide, if this too is prohibited. In effect, we have only modelled absolutely binding prohibitions and obligations and we have not introduced means of comparing the claims of one duty against another. We will thus offer an alternative set of constraints on the deontologist's utility function that will accommodate such distinctions. Note that we are aiming, in the first instance at least, for a model that respects the von Neumann-Morgenstern axioms, despite our previous comments about the potential dispensability of continuity. As such, we will not consider models that introduce multiple dimensions of utility, or models that incorporate the "hyperreals" or the "surreals" (see, for instance, Conway 1976) in place of the reals as the range of the utility function. We do not deny that these may be very fruitful avenues for modelling deontological reasoning, if one is willing to forgo the continuity axiom.[23]

We will pursue deontological constraints on utilities that remain more or less in the vicinity of D1* and D2*. The idea, intuitively speaking, is to assign very large utilities and disutilities in place of infinite (dis)utilities. The important thing about duties, from a deontologist's point of view, is that one is rationally bound to them over and above other possibilities for action. So, for that matter, what we really require is that the utilities of outcomes associated with obligations (prohibitions) be sufficiently larger (smaller) than the utilities of ordinary outcomes. To be more precise, the absolute difference between the utilities of outcomes associated with an obligation (prohibition) and a permissible act should be much greater than the absolute difference between any two permissible acts.[24] Our revised set of deontological constraints must therefore include the following:

(D1) If O_{ij} is the result of a prohibited act and O_{kl} and O_{mn} result from any two permissible acts, then any admissible utility function u must be such that $u(O_{ij}) < u(O_{kl})$ and $|u(O_{ij}) - u(O_{kl})| \gg |u(O_{kl}) - u(O_{mn})|$

(D2) If O_{ij} is the result of an obligatory act and O_{kl} and O_{mn} result from any two permissible acts, then any admissible utility function u must be such that $u(O_{ij}) > u(O_{kl})$ and $|u(O_{ij}) - u(O_{kl})| \gg |u(O_{kl}) - u(O_{mn})|$.

The assignment of finite utilities to the outcomes of duties also allows us to introduce a ranking among obligations and prohibitions. Say there is a ranking of prohibition types $P_1, P_2, \ldots P_n$, such that P_k outranks (or is more pressing than) P_{k+1} and a ranking of obligation types, or positive duties, D_1, $D_2, \ldots D_m$, such that D_l outranks D_{l+1}, then:

(D3) If O_{ij} is the result of a prohibited act of type P_r and O_{kl} and O_{mn} result from prohibitions of type P_s, such that P_r outranks P_s, then any admissible utility function u must be such that $u(O_{ij}) < u(O_{kl})$ and $|u(O_{ij}) - u(O_{kl})| \gg |u(O_{kl}) - u(O_{mn})|$.

(D4) If O_{ij} is the result of an obligatory act of type D_r and O_{kl} and O_{mn} result from obligatory acts of type D_s, such that D_r outranks D_s, then any admissible utility function u must be such that $u(O_{ij}) > u(O_{kl})$ and $|u(O_{ij}) - u(O_{kl})| \gg |u(O_{kl}) - u(O_{mn})|$.

Clearly D1–D4 allow us to represent more and less binding duties, but do they provide for a better treatment of mixed acts? To a large extent at least, these revised constraints on the deontologist's utility function do answer to our earlier problems. The outcomes of duties are still set apart from ordinary act outcomes in terms of the size of their (dis)utility. Moreover, duties of varying importance are similarly set apart from one another. But it is also the case that mixed acts with varying probability distributions will not be ranked together. For example, the mixed act that yields 0.9 probability of performing an ordinary, morally neutral, act and only 0.1 probability of satisfying an obligation will be less preferred than the mixed act yielding a much higher probability of satisfying the obligation. This should be good news for the deontologist. Of course, the constraints D1–D4 do not specify *just how much* difference there should be between the utilities of duties of varying importance, and between duties and ordinary acts. But that is a detail for the deontologist to work out.

It might yet be argued, however, that our deontological decision theory, captured by the addition of constraints D1–D4 on an agent's utility function, has some unusual features. We have thus far merely extended von Neumann and Morgenstern's theory rather than departed from it, but there is scope to argue that the latter route must be pursued if we want to faithfully model deontological decision-making. In particular, our preservation of the continuity axiom could be questioned. Perhaps the deontologist does not want any mixture involving a duty to be ranked amongst other ordinary acts, no matter how small the probability of the duty in the mixed act. Likewise, the deontologist might not want any mixture involving a prohibition, no matter how small the probability of the prohibition, to be ranked above an ordinary permissible act. A related oddity in our model is that it does not rule out a 50–50 mixture of a prohibition and an obligation having the same expected utility as an ordinary act. We would not want to revert back to infinite (dis)utilities to address these issues, because, as discussed, this move introduces the problem of not being able to discriminate between duties of varying rank, nor mixtures with varying probability distributions over duties. But there are other modelling possibilities that we have only briefly mentioned that could depict an absolute discontinuity between duties and ordinary acts. For instance, one could appeal to a lexical decision theory that effectively employs different dimensions of utility to deal with the duty and ordinary aspects of acts.

While we recognise the potential qualities of deontological decision theories that violate continuity, we will not pursue such alternatives here. The reason for this is that we think our model, which introduces the constraints D1–D4, is at least plausible. In fact, we hold that the issue of mixed acts is not at all clear-cut when it comes to deontological ethics. Ethical discussions are rarely conducted in probabilistic terms, and so it follows that matters such as the status of mixed acts tend to be overlooked. The question of whether duties should be both agent and time relative is another issue that is not typically addressed by deontologists. The constraints D1–D4 do not in fact require time relativity of duties because there is not the same swamping problem associated with infinite (dis)utilities. Nonetheless, the deontologist might want to stipulate that duties be time relative. Let us just say that a valuable aspect of the modelling process is that it focuses attention on such questions and thus sheds light on the details of ethical theories.

In any case, it might be argued that we have not yet provided the means to represent all the demands of deontology. A difficult issue yet unaccounted for is the idea of *defeasible* obligations and prohibitions, that is, obligations and prohibitions that may be overturned by consequentialist considerations when the stakes are sufficiently high. This is perhaps best considered another way in which the description of duties is context-sensitive. For instance, lying under circumstances in which only a modest amount of consequentialist good would otherwise be forsworn is prohibited, whereas lying under circumstances in which excessive consequentialist good would otherwise be forsworn is just an ordinary permissible act.[25] The trouble is, if we want to be more precise about the circumstances in which lying is prohibited, then we will need to do the consequentialist utility calculations.

In what follows, we will suggest a way to deal with defeasibility. Our proposal is rather complicated, however, and does not sit easily with standard decision theory. This might lead one to conclude that defeasibility is not in fact a core component of deontology. Or else it might point to an incompatibility between standard decision theory and deontology. We save this discussion for later. First, the details: it seems most plausible to say that moral duties are variably defeasible. The circumstances in which a deontologist might be prepared to license the telling of a lie are more widespread than the circumstances in which they would licence the torture of innocents, for example. We suggest that the deontologist select a utility function (from amongst the possible set of utility functions satisfying D1–D4 that are each positive linear transformations of each other), and nominate degrees of utility/disutility appropriate to the exercise of every obligation and prohibition—limits beyond which they are prepared to give up the claim of that duty. This could be recorded as a function f from prohibited act-types and obligatory act-types to defeasibility limits. It is a function from act-types to degrees of utility. These defeasibility utilities will be relatively high for prohibited act-types, and as stated above, they will be higher the more serious the prohibition. By "relatively high", we mean high within the range of utilities for ordinary act outcomes. At the other end of the spectrum, the defeasibility utilities for obligatory act-types will be relatively low (within the range of ordinary act outcome utilities), and they will be lower the more pressing the obligation. Even in this more complex setting, the representation of deontological preferences should still be unique up to positive linear transformation. Note that the deontological utility function and accompanying defeasibility function come as a pair—positive linear transformations of the utility function are permissible, but must be accompanied by a similar transformation of the defeasibility function if the agent's overall preference ordering is to go unchanged.

The idea is that acts are classified as obligations or prohibitions only if their outcomes do not breach the defeasibility limit specific to that duty. So what is it exactly for a duty to breach its defeasibility limit? We need to consider what the utilities for the act's outcomes would be in consequentialist terms, i.e., what the outcome utilities would be if the act were not a duty but just an ordinary act (we will call this notion of utility "consequentialist utility"). It is arguably the *expected* consequentialist utility of the potential duty (which will denote by $\mathrm{EU_C}$), rather than the consequentialist utility of any of its outcomes in isolation, that should be assessed relative to the defeasibility threshold. And (again arguably) the most plausible model of this makes defeasibility dependent on the nature of the alternative available acts. In this way, defeasibility is not about breaching some absolute utility value (supplied by the function f), but rather is an issue of the harm done or the goods foresworn by acting in the ordinarily obligatory manner. In certain situations, for example, lying may be permissible, not because the expected consequentialist utility of lying exceeds some absolute expectation, but because all other available options are just so much worse, from a consequentialist point of view, than lying.

To model our version of defeasibility, we introduce two extra constraints—D5 and D6—on the deontologist's utility function. Consider obligations first, or all the acts that look like obligations. Let a_o be some act that is individuated both in terms of its character (it is possibly a specific kind of obligation) as well as in terms of its structural relations to the other acts available in the choice setting.[26] The relevant defeasibility limit (given a particular utility representation) is given by $f(a_o)$. The n alternative available acts in the choice setting are denoted by a_i (where $0 \leq i \leq n$). Recall that $\mathrm{EU_C}$ refers to the expected utility of a potential duty in consequentialist terms, i.e., what the expected utility of the act would be if it were just an ordinary act.

(D5) If, according to initial calculations $\mathrm{EU_C}(a_o) < \sup\{\mathrm{EU}(a_i)\} - f(a_o)$ then the defeasibility limit is surpassed and a_o is in fact just an ordinary act; the outcomes of a_o should be assigned regular consequentialist utilities. Otherwise, if the inequality does not hold, then a_o really is an obligation and its outcomes should be assigned utilities in accordance with D1–D4.

Defeasible prohibitions work similarly. Let a_p be some act that looks like a particular kind of prohibition. The relevant defeasibility limit (given a particular utility function) is thus given by $f(a_p)$. Again, the n alternative available acts in the choice setting are denoted by a_i (where $0 \leq i \leq n$).

(D6) If, according to initial calculations $\mathrm{EU_C}(a_p) > \sup\{\mathrm{EU}(a_i)\} + f(a_p)$ then the defeasibility limit is surpassed and a_p is in fact not a prohibition at all, but just an ordinary act whose outcomes should be assigned regular consequentialist utilities. Otherwise, if the inequality is not satisfied, a_p really is a prohibition and its outcomes should be assigned utilities in accordance with D1–D4.

There is scope for disagreement as to whether defeasibility really is an essential component of deontology. Moreover, there is ample room for dispute about how any such defeasibility condition actually works, but we have given here a plausible model that may be adjusted as desired. As one might suspect due to its complexity, the model offered is a significant deviation from the standard decision-theoretic framework because it makes the classification of acts a two-step process that can be made precise only by reference to a background consequentialist utility representation. Thus, we must acknowledge that the inclusion of defeasible obligations/prohibitions jeopardises the compatibility of deontology and standard decision theory.

Next we turn to virtue ethics. This is more difficult still since, as we pointed out in section 2, virtue ethics is concerned as much with the motivations of the agents as with their actions or with the outcomes of those actions. But we can make some headway here. A concrete example might be helpful at this stage. Consider the choice of whether to donate to charity or not. A consequentialist is only interested in the outcomes of the acts in question—it doesn't matter whether a donation is motivated by generosity, by the desire to be seen to be generous, or by guilt. All that matters, from the perspective of the consequentialist, are the consequences of the act in question.[27] But the motivations make all the difference in the world for the virtue ethicist. Presumably, the act of charity stemming from generosity is what a virtuous person would aim at doing (at least for direct virtue theorists, who identify right action with virtuous action). This, then, suggests a way of incorporating virtue ethics into decision theory: we discriminate actions a little more finely than normally so that we distinguish actions with different motivations, and assign different utilities according to the virtuousness of the act's motive. If an act has a virtuous motivation, we say that it expresses the virtue in question. Thus, a generous act of charity expresses the virtue of benevolence. A cynical act of charity does not.

A satisfactorily complete virtue theory should provide us with the means of discriminating actions in terms of their expression of virtue. Indirect virtue theories, i.e., those that characterize right action in terms of actions that a hypothetical virtuous agent *would* undertake, will describe the hypothetical expression of virtue. Actions would then be ranked in terms of the extent of their match to the motivations of ideal virtuous agents. Here we consider only direct versions of virtue theory. Though more complex, indirect virtue theories can be accommodated within the general framework we describe. In any situation an agent will confront a finite number of available actions, $a_1, a_2, \ldots a_i, \ldots a_n$. We use our virtue theory to rank these actions in terms of their expression of virtuous motivation. Minimal constraints that virtue theories impose upon the relevant utilities are given by V1 and V2 (where we read '$O_{ij} \equiv O_{hk}$' as outcomes O_{ij} and O_{hk} are equivalent, in the sense that they are indistinguishable outcomes in all but their virtue-theoretic motivations).

(V1) If a_i is more virtuous in terms of motivation than a_h and $O_{ij} \equiv O_{hk}$, then $u(O_{ij}) > u(O_{hk})$.

(V2) If a_i and a_h are equally virtuous in terms of motivation and $O_{ij} \equiv O_{hk}$, then $u(O_{ij}) = u(O_{hk})$.

Such virtue-theoretic constraints on the utility function come down to this: the utility of the outcome of a virtuously motivated act will always exceed the utility of that very outcome produced by less virtuous means. A charitable gift may be a valuable outcome on this way of viewing things, but well-motivated charitable gift-giving invariably possesses higher utility. Equivalent outcomes of equally virtuous actions have identical utility. Say that fairness and generosity are equally virtuous motivations. The outcome of a charitable donation given out of a sense of fairness would then attract the same utility as an identical outcome motivated by generosity. (It is more likely, however, that virtues are not all equal like this. See below.)

V1 and V2 account for the effect that motivations of varying virtue have on the choice-worthiness of acts, but the typical virtue theorist also cares about the outcomes of acts. As it stands, our model permits perverse preferences with respect to outcomes. For example, a fanatical Nazi might regret the coldness and harshness required to pursue the destruction of European Jewry, and yet hold this goal to be sufficiently important that his strong motivational scruples are overwhelmed. We therefore need to introduce further constraints on the utility functions of virtuous agents. A complete and adequate virtue theory will probably furnish these constraints in terms of the legitimate ends of virtuous agency.[28] This might be characterised in terms of the pursuit of agent-neutral goods, or in neo-Aristotelian versions, as a nested series of agent-specific ends leading to the condition of eudaimonia, or the leading of a good and fulfilled life. Let us describe the former case. One example of this is a version of morality as universal benevolence.[29] The primary virtuous motivations here are benevolence and generosity, and outcomes are otherwise assessed in the familiar utilitarian way. We introduce two additional constraints on the utility function to model this utilitarian component:

(V3) Let acts a_i and a_k each express the same motivation m. If O_{ij} involves greater total welfare than O_{kl}, then any admissible utility function u must be such that $u(O_{ij}) > u(O_{kl})$.

(V4) Let acts a_i and a_k each express the same motivation m. If O_{ij} involves the same total welfare as O_{kl}, then any admissible utility function u must be such that $u(O_{ij}) = u(O_{kl})$.

V1 to V4 might be considered sufficient for representing the demands of at least one plausible version of virtue theory. One might want to say more, however, about the contribution that virtuous motivation should make to an outcome's utility. It is plausible to think that any such contribution should be systematic across the various motivations. Having said that, not all virtues are equally significant and so they should each make a distinctive contribution to the utility of the outcomes of actions that express it.[30] In view of this, we might introduce a virtue-specific additive factor that represents the difference between an outcome's utility when it is produced with neutral motivation and the same outcome's utility when it is produced by the motivation in question. Thus we introduce a function, V, from motivations to additive factors, which features in the following constraint (where we again read '$O_{ij} \equiv O_{hk}$' as outcomes O_{ij} and O_{hk} are equivalent, in the sense that they are indistinguishable outcomes in all but their virtue-theoretic motivations):

(V5) If O_{ij} is produced with motivation m_i where $V(m_i) = \zeta$, and if $O_{ij} \equiv O_{hk}$ and O_{hk} is produced via virtue-neutral means, i.e., with $V = 0$, then $O_{ij} = O_{hk} + \zeta$.

The basic idea here is that there is utility arising from the motivation of the action in question, and this utility needs to be taken into account. Although there are various ways this "extra" utility or disutility might be accommodated, we take it on board by *adding* it to the non-virtue-theoretic value of the outcome.[31] The sign of ζ matters: when the motivation in question is virtuous, ζ will be positive; when the motivation in question is vicious, ζ will be negative; and, obviously, when the motivation is virtue-theoretically neutral, ζ will be zero. But this does not amount to there being a privileged zero for our utility function u (something ruled out by the standard axioms of utility theory). The utility function in question does not have privileged zero, it is just the virtue-theoretic additive factor (the motivation function V) that has such a feature. And here a privileged zero is entirely appropriate, with the interpretation of the zero a standard part of virtue theory; it is a virtue-neutral motivation and represents the boundary between virtuous and vicious motivations.[32]

There are still substantive issues of what the various virtues are and how to resolve different virtues in terms of their desirability and the contribution they make to the choice-worthiness of acts/outcomes. We set such complications aside for now. The details of these interactions is the business of virtue ethics, and it is simply not our task to prejudice the question of how to resolve open or difficult questions within ethical theories. Our task is merely to show how each theory, *once suitably spelled out by the advocates of the ethical theory in question*, might be modelled in an appropriate fashion in the decision-theory framework. So bearing this in mind, we take it that the above axioms constitute a plausible start to the problem at hand.

4. Adequacy of the Models

It is common to distinguish two quite different kinds of model in science: descriptive models and explanatory models. A descriptive model is a model that's empirically adequate in the sense that it gets the data right (or nearly right). An explanatory model needs to shed light on the underlying reasons for the way the system in question behaves as it does.[33] We don't propose that the distinction between these two types of models is sharp—it certainly is not—but it does provide a useful way to think about the purpose and role of theoretical models.[34] With this distinction in mind, let's turn to the adequacy of the three models of ethical decision making we have presented in this paper.

The formal constraints on the utility function that we've proposed above, we take it, amount to reasonable ways of representing familiar versions of the three ethical theories in question. At least the formal constraints are a credible first shot at representing versions of the theories in question. But it is important to note that all we have done is provide a framework that is able to model the preferences of the utilitarian, deontologist, and virtue theorist, as well as the outcomes of their decision making processes; we have not attempted to model their underlying thought processes and motivations. Nor have we modelled the justifications moral theories furnish agents. Moral theories not only aim at specifying moral behaviour, they aim to supply justifications for moral behaviour. Only our model of utilitarianism furnishes the means of morally justifying actions. According to utilitarianism, a moral agent is always justified in optimally promoting general welfare, which is tracked by expected utility calculations employing an adequate utilitarian utility function. A utilitarian can thus use the fact that the expected utility of a particular action is greatest of all current options to justify their performing it.

For deontologists, however, appeal to the effect of the enormous disutility of prohibited options, which are given arbitrary precision by nominating a specific disutility, has no justificatory power. These precise measures of disutility do not reflect deontological proposals about the precise relative disvalue of a prohibited act, for example. In our model, the disutility of a prohibited act applies only to an agent's current options. But if a person disvalued a prohibited kind of act in general, say lying, and used this valuation as a basis of moral decision-making, the disutility of lying should affect the utility of all outcomes involving lying, not just those involving the agent lying now. Although deontological determinations of right action can be modelled in terms of the pursuit of optimal expected utility, it does not follow that deontologists are motivated to optimise expected utility. Nor are they inclined, or equipped, to justify actions by appealing to the optimisation of expected utility. Deontologists, typically, think of morality as providing a series of constraints on behaviour based on their interpretation of what it is to respect another person, or on intuitions of rightness that are independent of their conception of the good.

Unlike deontologists, virtue theorists share an overall teleological approach with utilitarians. Virtue-theoretic considerations demand that virtuously derived outcomes have enhanced utility and viciously derived outcomes have diminished utility. However, a virtue-theorist is ill equipped to justify their decisions in terms of the utilities thus specified. This is because, as with deontology, virtue-theoretic modifications of utility functions apply only to the options currently faced by a decision maker. They do not describe the kind of agent-neutral value assessments (e.g., malicious actions make the world poorer) that make for plausible justifications of value promotion. The virtue-theorist may be modelled as optimising expected utility under an appropriate description of this utility, but they are not equipped to justify their actions in terms of the promotion of general values expressed by these utilities. Virtue-theoretic justification of action must take a different form: appealing to the importance of self-respect, or the special responsibility each person has for their own character and its expression in action.

So a case might be made for the utilitarian model being an explanatory model but the other two are only descriptive (rather than explanatory), in the sense that they describe (ideal) deontological and virtue-theoretic agents, respectively. The fact that these models are only descriptive does not mean, however, that we should be dismissive of them. If virtue theory and deontology can be described within the standard decision-theoretic framework then that is a non-trivial and interesting finding. But we need to be careful not to over-interpret the models and thus overstate their significance. In all three models we've represented the decision-making process in terms of the maximisation of expected utility (with ethical constraints on the utility function). But we should not read too much into the name 'utility function' or 'expected utility'. These are both just formal features of the model and may have nothing to do with utility in the usual sense. Indeed, this is so even for standard decision theory (without ethical constraints). The utility function is best regarded as an uninterpreted mathematical function constrained by the von Neumann-Morgenstern axioms. Typically the utility function is interpreted in the obvious way, as a measure of agent-neutral and time-insensitive values, but this is a further move—arguably a move away from a descriptive to an explanatory model—and this move can be resisted.

A related point is that we should not conclude from the fact that all three ethical theories are represented as maximising some quantity, that they are all consequentialism in disguise.[35] Virtue theory and, in particular, deontology had to be shoehorned into the consequentialist framework of decision theory. As we've argued, we are not claiming to have provided explanatory models of these two, nor explanations of the behaviour of the deontological and virtue agents. Nor have we claimed to faithfully represent the justifications available to such agents. Indeed, our models either misrepresent or make opaque such justifications. It would thus be a mistake to press further claims about deontology and virtue theory that depend on our having captured the motivations for the theories in question.[36] As with all models, it is important to remember that these are just models and that there is danger in reading off too much from the model.[37] Having said this, however, there is still a significant issue of how explanatory the models we've presented are. We leave this issue for another occasion.

There is also the issue of whether the models we've developed are consistent, in the sense that they deliver consistent advice.[38] This is a big issue. A full response would, presumably, involve providing consistency proofs, or at least relative consistency proofs, of all the models we discuss in this paper. We won't do that here. Instead, we will say a few words by way of alleviating such concerns. The first thing to note in relation to this issue is that the resulting model will be inconsistent if the ethical theory is inconsistent—at least if the model is doing its job. So for example, consider a virtue theory where the ranking of motivations in a particular context is non-transitive. Generosity is more virtuous than loyalty, loyalty is more virtuous than courage and yet courage is more virtuous than generosity. Such a theory will demand inconsistent actions of moral agents in certain situations. But any such inconsistency would be a feature of the ethical theory, and since our task has been to faithfully (or at least as faithfully as possible) model the ethical theories in question within standard decision theory, the resulting model will, if successful, be inconsistent. This is as it should be. The more worrying kind of inconsistency is inconsistency introduced by accident, as it were: inconsistency introduced as a result of adding the further constraints.

Recall that our project, so far as possible, is that of adding further constraints to standard decision theory so that the resulting theory respects the ethical sensitivities of the various ethical theories. The kind of further constraints we have added are supposed to fill in the details of the decision theory and are motivated by the ethical theories in question. There are no further constraints we've added for any other reasons. So if the resulting model turns out to be inconsistent, that would suggest that either the ethical theory itself is inconsistent (as considered in the paragraph above) or that the ethical theory is inconsistent with standard decision theory. For the most part, we have aimed for models that do not conflict with the axioms of von Neumann-Morgenstern utility theory, and we have been upfront about any conflicts that have arisen. In particular, the deontologist has to make some choices about what their ethical theory demands; these choices might favour compatibility with standard decision theory, or they might not. For instance, if the deontologist is prepared to accept continuity between duties and ordinary acts, and if they are prepared to give up on the notion of defeasible duties, then deontological constraints can be added to standard decision theory without any inconsistency. Otherwise, there will be some incompatibility between the two. Either way, the conclusions are interesting.

Finally, we say a few words about why we've approached the task of this paper by placing further constraints on the utility function rather than on the preference structure. After all, it might be argued that the axiomatisation of preferences is the more fundamental in the von Neumann-Morgenstern theory. Moreover, both the deontologist and the virtue ethicist might complain about the numerical character of the utility function and about taking such a rich mathematical structure for granted in representing their ethical theories. The deontologist and the virtue theorist do not countenance such numerical representations and so, it seems, we misrepresent them right from the start. They might be more sanguine about taking preferences as basic, since these do not have the numerical character of utility functions. We have a couple of things to say in response to such concerns. First, we admit that placing further constraints on the axioms for preferences may well be a fruitful way to approach this problem. We are not claiming that the approach of this paper is the only way to achieve a reconciliation of ethics and decision theory. Pointing out that there may be other ways to approach the task in question does nothing to undermine our project. Indeed, were such a preference-based approach to be carried out, it would be fascinating to compare it to the approach we've suggested in this paper. As for the charge that we've misrepresented the deontologist and the virtue ethicist from the start by starting with the utility function, we point out that for these two ethical theories to be reconciled with standard decision theory (or at least most of the axioms of standard decision theory), somewhere along the line they will need to buy into utility functions. It strikes us as irrelevant whether utility functions are bought at the start or later on.[39] Finally, we reiterate our earlier remarks about the nature of the models we are proposing here; they are intended to be descriptive, in the sense that they faithfully represent the ethical decisions and not the moral psychology of the agents making the decisions.

In this paper we have shown that, despite initial appearances, deontology and virtue theory can be accommodated in something like the standard decision-theory framework, and thus expected utility theory need not be thought of as a tool available only to the consequentialist. As it stands, decision theory is silent on ethical matters, but with a little work, even the standard model can be made to accommodate *versions* of each of the major ethical theories. Of course, this does not settle the issue as to whether standard decision theory is entirely neutral with respect to ethical considerations. As indicated, some variations of our models depict versions of the major ethical theories that are not in fact compatible, to greater or lesser extent, with standard decision theory. In such cases we must ask whether there is a fault in the ethical theory in question, or whether it is standard decision theory that requires revision. This only serves to highlight the importance of investigating the connections between ethics and decision theory. We have seen that ethical theories can be clarified via decision-theoretic modelling. Less obvious, perhaps, is that the assumptions of standard decision theory are also challenged by differing ethical conceptions of value and right action.[40]

Notes

[1] One might argue about whether the respective axioms are in fact minimal constraints on beliefs and preferences. We will not directly pursue such issues here, but our project certainly raises the question as to whether von Neumann-Morgenstern utility theory is too restrictive a theory of rational preference. Nor will we say much about alternative decision theories. Indeed "standard" decision theory (let alone non-standard varieties) could well be considered Savage's (1954) theory, or alternatively, Jeffrey's (1983) theory, rather than that of von Neumann and Morgenstern. Or perhaps some version of causal decision theory should be considered "standard" these days. There are some significant differences between these theories. For our purposes, however, these differences do not matter, so we will stick to the von Neumann-Morgenstern axiomatisation to give focus to our discussion.

[2] Indeed, there are some interesting results along these lines. See Oddie and Milne (1991). Others who have considered the issue of the relationship between ethics and decision theory include Broome (1991), Colyvan *et al.* (2001), Dreier (2004), Jackson (2001), Jackson and Smith (2006), Louise (2004), and Sen (1977 & 1997).

[3] Recall our discussion of what we mean by "standard" decision theory in footnote 1.

[4] We say "supposed to supplement the original von Neumann-Morgenstern axioms" because occasionally we are forced to deviate from standard decision theory, and when we do so we will acknowledge this.

[5] We recognize that utilitarianism is just one class of consequentialist moral theory. We focus on utilitarianism because it is arguably the dominant consequentialist theory, and serves as a useful point of comparison with the other ethical theories.

[6] Preference utilitarianism introduces a cardinal measure of social utility because it involves (at least) comparing total numbers of satisfied preferences.

[7] For this discussion, we set aside a couple of variants of utilitarianism. One is the class of utilitarian theories that advise agents to pursue the maximum *possible* utility (no matter how improbable the relevant outcome is) as opposed to maximum *expected* utility. We also set aside satisficing versions of utilitarianism, i.e., versions in which moral agents aim for sufficient levels of utility.

[8] Alongside obligations and prohibitions, deontologists sometimes also posit permissions or prerogatives. We set aside these aspects of deontology here.

[9] The following characterization is a simplification. Some deontological approaches rest on direct intuitions about duties and about the priority of the right over the good rather than on accounts of respect. See Ross (1967).

[10] See, for example, Paul Taylor (1986).

[11] Slote (2001) develops virtue ethics along these lines.

[12] Rosalind Hursthouse (1991 and 1999) develops such a view.

[13] That is, the utility functions employed in decision theory respect distance between the various values.

[14] At least, this is how Joyce (1999) depicts Jeffrey's decision model. The value of an act, $V(a)$, is given by:

$$V(a) = \sum_{s} P(s \mid a) u(a \,\&\, S)$$

[15] The conditionality just discussed is a special case of context-sensitivity. It is plausible that not only the nature, but also the relative demands of duties, may vary in different contexts. For example, rankings of duties in order of importance may vary from one context to another. We ignore this further complication in our model.

[16] A "regular" probability function is one that assigns probability 1 only to logical truths and 0 only to logical contradictions. Some argue that a rational agent's belief function must be regular. See Hájek (2003, pp. 31–32) for a discussion of regularity in relation to Pascal's wager and the use of infinite utilities to model that problem.

[17] Hájek (2003) brings due attention to this problem in his discussion of Pascal's wager. He refers (p. 34) to a "Requirement of Distinguishable Expectations"—in Pascal's case, it should be the case that if one act is more likely to secure God's favour than another act, then the former is more desirable.

[18] It may also be the case that certain actions may be ruled against on consequentialist grounds. For example, my buying a gun would not be prohibited just because it raises the chance of my future self engaging in the prohibited act of killing. Rather, my buying a gun might simply be ruled against because it raises the probability of low utility outcomes such as deaths by shooting.

[19] The continuity or Archimedean axiom rules out infinite utilities. Informally, the axiom states that for any three acts p, q, and r where p is preferred to q and q is preferred to r, there is some mixed act comprised of p and r that is indifferent to q. Refer to Resnik (1987, p. 91) for a formal statement of the axiom. It is clear that if act p had infinite utility or act r had negatively infinite utility, then continuity would not hold because no mixture of p and r would correspond to an act with intermediate finite utility.

[20] We thank David Gray for drawing this issue of mixed acts to our attention.

[21] Joyce (1999, p. 82) distinguishes between the axioms of expected utility theory in this way.

[22] Note that one might argue for a version of consequentialism that involves a lexical ordering of values.

[23] Hájek (2003) explores both of these modifications to EU theory when considering how to revise Pascal's decision problem so as to avoid the problems associated with attributing infinite utilities to outcomes.

[24] We are much obliged to an anonymous referee for a suggestion that led us to formulate the deontological constraints in this particular way.

[25] Compare with the sort of contextual distinction that was made earlier between ordinary lying and lying under conditions of bad faith.

[26] One might think there will be a regress problem associated with defining acts relative to other acts, but we do not think this is the case here. Firstly, only potential duty acts need reference other acts in the choice setting. And secondly, we can conceptualize the defining/individuating of acts in a structuralist sense, rather than as a sequential (and recursive) process.

[27] This is a little simplistic. Charity motivated by guilt might have different consequences from charity motivated by generosity, but let's set such complications aside for now.

[28] Note that a complete deontological ethic would presumably include some consequentialist constraints as well. We did not include any such constraints in our presentation of deontology because they are quite distinct from the constraints pertaining to duties. In the case of virtue theory, on the other hand, the consequentialist constraints are more intimitely bound up with the constraints pertaining to virtue, and so we think it necessary to state them explicitly.

[29] Michael Slote (2001) attributes such a theory to James Martineau.

[30] As in the case of deontological theories we discuss above, we ignore the complication of context-sensitivity here. It seems plausible to say that virtues are not ranked absolutely, but are variably appropriate to distinct situations. In one situation, a sense of fairness might be the most appropriate—and thus virtuous—motivation to act upon; in another, a sense of generosity may be of greater moral significance. A fully developed virtue theory should be able to specify what it is that triggers this variability.

[31] There is an intuition some people have that the contribution of the virtuous or vicious motivation should be proportional to the stakes. This intuition pushes for multiplicative factor rather than an additive one as we have adopted here. Although there is something to be said for the multiplicative approach, it does face some technical problems, which, in our view, results in a more radical departure from standard decision theory. This is an interesting result, and shows that some versions of virtue theory are much less compatible with standard decision theory than others.

[32] The representation of virtue-theory preferences should still be unique up to positive linear transformation. As per the deontology model, the virtue-theory utility function and accompanying motivation function come as a pair—positive linear transformations of the utility function are permissible, but must be accompanied by an identical linear transformation of the motivation function, with the exception that, in the latter case, the zero-point must remain unchanged, i.e., the additive constant for the motivation function transformation must be zero.

[33] A couple of examples might help. A purely mathematical description of the growth of a population, in terms of, say, the logistic equation, may be empirically adequate in that such a model makes correct predictions about the abundance of the population in question. But without a story about why the population abundance can be described by the logistic equation, the model fails to be explanatory, at least with regard to causal processes. On the other hand, an explanatory model might lay bare the underlying biology and thus be causally explanatory (a story about carrying capacity, birth and death rates and so on), but may fail to deliver the predictive success of the logistic model.

[34] We put aside the issue of the place of normative models. All the models under discussion in this paper are normative (since they involve both ethics and rational decision making), but the descriptive–explanatory distinction is supposed to cut across the normative–descriptive distinction. (It is unfortunate that the word 'descriptive' is used in both these contexts—to contrast with both 'normative' and 'explanatory'.)

[35] Oddie and Milne (1991) and Louise (2004) draw conclusions along these lines from the representability of ethical theories in a consequentialist framework.

[36] Such as that the theories are really just consequentialism after all, that they are lacking motivation, or that they have implausible motivations.

[37] We do not, for example, conclude that fluids are incompressible because our model of fluid flow assumes this, or that Sydney has no hills because our street directory of Sydney has no hills.

[38] We thank an anonymous referee of this journal for raising this issue.

[39] If the worry is that hard-line deontologists and virtue ethicists will not buy into utility functions at *any* stage, then the game is over. The prospect of accommodating such versions of deontology and virtue ethics within decision theory seems hopeless.

[40] We'd like to thank audiences at the 2005 Australasian Association of Philosophy Conference at the University of Sydney, a workshop at the University of Queensland in 2005, a philosophy seminar at the Australian National University in 2005, a philosophy seminar at the University of Colorado at Boulder in 2006, and the 2006 CMU-Pittsburgh graduate conference. We are also grateful to Selim Berker, David Braddon-Mitchell, James Chase, Peter Forrest, David Gray, Drew Khlentzos, Julian Lamont, Jennie Louise, Gary Malinas, Graham Oddie, and Martin Rechenauer for very helpful conversations on the issues addressed in this paper or for comments on earlier drafts of the paper. We are especially indebted to Alan Hájek for many insightful comments on earlier drafts. These comments resulted in a number of significant improvements and prevented several serious errors.

References

Broome, J. 1991. "The Structure of Good: Decision Theory and Ethics" in M. Bacharach and S. Hurley (eds.) *Foundations of Decision Theory*, Blackwell, Oxford, 123–146.

Colyvan, M., Regan, H. M. and Ferson, S. 2001. "Is it a Crime to Belong to a Reference Class?", *Journal of Political Philosophy*, 9: 168–181. Reprinted in H.E. Kyburg and M. Thalos (eds.), *Probability is the Very Guide of Life*, Open Court, Chicago, 2003, 331–347.

Conway, J. H. 1976. *On Numbers and Games*, Academic Press, London.

Dreier, J. 2004. "Decision Theory and Morality" in A. R. Mele and P. Rawling (eds.), *Oxford Handbook of Rationality*, Oxford University Press, Oxford, chap. 9.

Gillies, D. 2000. *Philosophical Theories of Probability*, Routledge, London.

Hájek, A. 2003. “Waging War on Pascal’s Wager”, *Philosophical Review*, 113: 27–56.

——— 2003a. “Interpretations of Probability” *The Stanford Encyclopedia of Philosophy*, (Summer 2003 edition) E. N. Zalta (ed.), URL=<http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/>.

Hursthouse, R. 1991. “Virtue Theory and Abortion”, *Philosophy and Public Affairs*, 20: 223–246.

——— 1999. *On Virtue Ethics*, Oxford University Press, Oxford.

Jackson, F. 2001. “How Decision Theory Illuminates Assignments of Moral Responsibility”, in N. Naffine, R.J. Owens, and J. Williams (eds.), *Intention in Law and Philosophy* Ashgate, Aldershot, pp. 19–36.

Jackson, F. and Smith, M. 2006. “Absolutist Moral Theories and Uncertainty”, *The Journal of Philosophy*, 103: 267–283.

Jeffrey, R. C. 1983. *The Logic of Decision*, 2nd edition, University of Chicago Press, Chicago.

Joyce, J. 1999. *The Foundations of Causal Decision Theory*, Cambridge University Press, Cambridge.

Kant, I. 1996. *Practical Philosophy*, translated and edited by M. Gregor, Cambridge University Press, Cambridge.

Kaplan, M. 1996. *Decision Theory as Philosophy*, Cambridge University Press, Cambridge.

Kolmogorov, A. N. 1956. *Foundations of the Theory of Probability*, 2nd English edition, Chelsea Publishing Company (first published in 1933).

Levi, I. 1986. *Hard choices: decision making under unresolved conflict*, Cambridge University Press, Cambridge & New York.

Lewis, D. 1980. “A Subjectivist’s Guide to Objective Chance” in R. Jeffrey (ed.), *Studies in Inductive Logic and Probability II*, University of California Press, Berkeley, 263–293.

Louise, J. 2004. “Relativity of Value and the Consequentialist Umbrella”, *The Philosophical Quarterly*, 54: 518–536.

Oddie, G. and Milne, P. 1991. “Act and Value: Expectation and the Representability of Moral Theories”, *Theoria*, 57: 42–76.

Resnik, M. 1987. *Choices: an introduction to decision theory*, University of Minnesota Press, Minneapolis.

Ross, W. D. 1967. *The Right and the Good*, Clarendon Press, Oxford.

Savage, L. J. 1954. *The Foundations of Statistics*. Wiley, New York.

Sen, A. 1977. “Rational Fools: A Critique of the Behavioural Foundations of Economic Theory”, *Philosophy and Public Affairs*, 6: 317–344.

Sen, A. 1997. “Maximization and the Act of Choice”, *Econometrica.* 65: 745–779.

Slote, M. 2001. *Morals from Motives*, Oxford University Press, Oxford.

Taylor, P. 1986. *Respect for Nature: A Theory of Environmental Ethics*, Princeton University Press, Princeton.

Von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.