A Causal Theory of Double Effect Reasoning

Gerard J. Rothfus

December 2021

Abstract

Trolley problems and like cases show the inadequacy of a purely consequentialist moral theory. Formulating an alternative theory that can match the generality and relative rigor of consequentialism, however, has proved challenging. I offer a new theory that formalizes key insights of traditional double effect reasoning via causal graphs to solve the trolley problem and furnish a plausible rival to consequentialism.

1 Introduction

There is a tragic dilemma nearly every philosopher has faced at some point in her career, though hopefully only in her mind.[1] A runaway trolley carrying five passengers barrels down an unfinished track toward a precipitous cliff. Luckily, you have time to pull a nearby lever to switch the trolley onto an alternate track, averting disaster for its passengers. Unhappily, an oblivious third party has gotten his foot stuck attempting to cross the second track. Pulling the lever means life for the trolley's passengers but death for the man with the trapped foot. Failing to pull the lever reverses these fortunes. While the death of the man caught in the tracks would no doubt be a serious evil, many of us are inclined to say that pulling the lever in this case would be morally permissible, if not obligatory.

Consequentialist moral theory of a classical sort has little difficulty accounting for such a judgment: the good of saving five lives outweighs the harm of ending one. But well known variations on the standard trolley problem make trouble for (act-)consequentialism. Suppose the lever in the scenario described above is removed and in its place stands a man of sufficient mass to stop the trolley and save its ten passengers, were you to push him on to the tracks. Of course, if you were to so push him, the force of the trolley's impact on the man's body would kill him. Its undeniable benefits notwithstanding, many of us inclined to pull the lever in the standard case find the prospect of pushing someone on to the tracks in this variant deeply objectionable. Consequentialism, however, fails to account for this asymmetry in our moral judgments, since the reasons (in terms of prospective goods and bads) that favor/disfavor pulling the lever appear perfectly symmetric to those that favor/disfavor pushing the man.[2]

Since Foot and Thomson made trolley problems like these famous, moral philosophers have been searching for a plausible non-consequentialist moral theory capable of accounting for the intuitive difference between pulling the lever and pushing the man. Along the way, a deluge of judgements about other hypothetical cases (some involving trolleys, some not) have entered into the story as further constraints upon what such a theory must plausibly look like. While the resultant literature has been lively and thought provoking, relatively little (hardly any) effort has gone into developing formal frameworks in which the relevant bevy of cases and

[1] For the original statement of this problem, see Foot 1967, though the version I present here is closer to the variant introduced by Thomson 1976.

[2] A certain kind of deontologist may be inclined to concur with the consequentialist's judgment of moral symmetry while reversing her conclusions, i.e. concluding that pulling the lever is just as wrong as pushing the man.

the principles that distinguish them might be usefully modelled and contrasted. Perhaps as a result, extant alternatives to consequentialism in this area often miss some of its power and generality, neglecting such critical elements of any complete framework for moral reasoning as applicability to decisions under uncertainty and algorithmic implementability in artificial agents.

Some, of course, take the search for a 'solution' to the trolley problem to be a misguided project aimed at systematizing vague and unreliable intuitions.[3] They feel no need to further the venture. But for those of us convinced that we do need a moral theory that handles trolley problems differently than consequentialism, the noted gaps in the literature are well worth filling. I offer my contribution to this program here by introducing a formal framework suitable for modelling a wide range of moral decision problems, including trolley cases, and then articulating a general moral theory that I take to encode some of the key insights of *double effect reasoning (DER)*, while patching up a few of the holes in DER's traditional formulations and shoring up its respectability as a plausible and practicable alternative to purely consequentialist reasoning. What results, I hope, is not simply another proposed solution to the trolley problem, but the first steps towards establishing a more rigorous and plausible method for applying DER-style reasoning in various practical contexts (e.g. medical and machine ethics).

2 Preliminaries: Actions, Causes, Values

Moral theories offer deliberating agents guidance concerning how to act rightly in various decision problems. Abstractly, we could think of such moral theories as (perhaps partial) functions mapping hypothetical decision problems to what the theories view as their morally permissible option sets. To state a moral theory (e.g. consequentialism or DER) in this abstract and idealized way requires first articulating a model of decision problems rich enough to capture all the features of such problems that the theory takes to be relevant for settling questions of permissibility. The model of decision problems that I will suggest as suitable for modelling trolley cases and adequate for an initial statement of an exact theory of DER will take such problems as consisting of five components: (i) a set of possible *actions*, A, (ii) a set of *event-variables*, E, (iii) a *causal ordering*, $\rightarrow$, on $V := \{A\} \cup E$, (iv) a family of *causal dependence functions*, F, specifying the precise causal relations obtaining between causally connected act/event-variables, and (v) a *value function*, v, measuring the comparative strength of the various reasons at play in the problem. Each of these components deserves explanation.

First: *actions*. The most obvious constituent of a decision problem is the set A of actions (here assumed finite) that an agent facing the decision problem is free to choose amongst. (We might just as well refer here to *acts* or *options*.)[4] For ease of formal modelling, I take actions to be not (actual or possible) events, as would perhaps be ontologically appropriate, but rather propositions.[5] So, for example, my action of pulling a given lever is identified in the present model with the proposition that I pull the lever. This modelling choice will allow us to conveniently connect actions with other propositions via standard Boolean connectives. It also naturally allows us to state a critical assumption regarding A: namely, that it constitutes a logical partition, i.e. the members of A are pairwise incompatible, while their disjunction has the logical force of a tautology. This amounts to assuming that in a well-formed decision problem, an agent must, as a matter of necessity, choose to perform exactly one action, i.e. opt to make exactly one member of A true. Such actions may, of course, include what we would intuitively

[3] See, e.g. the arguments in Greene 2013.

[4] The sort of actions intended here are something like *basic actions* in Anscombe 1957's sense, e.g. things like pulling a lever or swinging a bat but not writing a book or fighting a war. The rough and ready idea is that one action corresponds to one act of the will, i.e. an action is something within an agent's power to realize by a single decision. Of course, many decision problems involve multiple choice points and hence multiple action sets. The model discussed here ought to eventually be generalized to handle such *dynamic choice problems* in future work; see §10.

[5] I follow the lead of Jeffrey 1965/1983 in this regard.

call 'doing nothing', i.e. refusing to willfully direct one's bodily or mental movements in any particular direction. I will often refer to the action of doing nothing as the *null act*, denoted $a_0 \in A$.[6]

Second: *event-variables*. In any given decision problem, there are various morally relevant features of the world external to the agent's possible willings that need to be represented in any adequate model of the problem. For example, in the standard trolley problem, we may need to represent the trajectory of the trolley, the position of the man with the trapped foot, etc. I propose doing so via what I call *event-variables*, each of which represents the different possible ways a particular aspect of the world might turn out. Event-variables may be modelled, like A, as partitions of propositions. Thus, where h is the proposition that the man stuck on the track is hit, one event variable in the standard trolley proble may be: $H = \{h, \overline{h}\}$.[7] The set of all such event variables will be denoted as E. It is crucial that E be rich enough to include a representation of all morally relevant considerations that tell fundamentally (i.e. non-instrumentally) for and against each available action in the decision problem at hand. Again employing the standard trolley problem as an example, assuming that the lives of the passengers and the man on the track are the relevant goods at stake in your decision, each of these person's living/dying must be represented as possibilities via event-variables like H.[8] As with A, I assume E to be finite for simplicity. For convenience, I will write V as a shorthand for $\{A\} \cup E$. Finally, I assume that all the variables in V are logically independent in the sense that every possible way of jointly specifying the true members of the variables V is logically coherent.

Third: *causal ordering*. In the statement of any decision problem, there will typically be (implicitly or explicitly) assumed causal relations obtaining amongst the variables represented in V. For example, whether the man on the track gets hit, H, causally depends upon my decision whether or not to pull the lever, A. Intuitively, one variable X directly causally depends upon another Y just in case which member of X turns out to be true is (partly) causally determined by which member of Y turns out to be true (and this effect is not screened off by any intermediary variables in the model). Formally, we represent this direct causal ordering via an acyclic binary relation, $\rightarrow$, on V, corresponding to a directed acylic graph (or DAG) in which the members of V are treated as vertices. Following standard graph-theoretic terminology, I will refer to the set of variables that immediately precede a given variable X relative to $\rightarrow$ as the *parents of* X or $par(X)$ and to the set of all variables that precede X along some chain of $\rightarrow$ as the *ancestors of* X or $an(X)$. I assume, as seems appropriate in a model of free decision making, that $par(A) = \emptyset$, i.e. actions are treated as uncaused in the model.[9]

Fourth: *causal dependence functions*. $\rightarrow$ gives the direct causal dependence ordering amongst the members of V, but it doesn't specify the precise nature of the causal dependence amongst these variables. For example, in the trolley problem, $A \rightarrow H$ signifies that my decision to pull the lever or not in some way bears upon whether the man on the track gets hit, but it doesn't say, for example, whether pulling the lever will lead to him getting hit or not pulling the lever will lead to him getting hit. I will encode such detailed information about how exactly one variable $X \in V$ (where $par(X) \neq \emptyset$) causally depends upon its parents in a causal dependence function, $f_X : N_X \rightarrow X$, where N_X is the set of maximally consistent conjunctions formable from the members of $par(X)$. For example, in an overly simplified version of the trolley problem in

[6] This is purely for notational convenience, nothing in the substantive theory I develop here relies upon the existence of a privileged null act.

[7] I will typically write partitions, whether action sets or event-variables, with uppercase letters and propositions with lowercase letters.

[8] Note: there is no requirement here that event-variables or their members uniquely correspond to reasons. What is required is simply that every reason be represented in the value of some variable. It is thus fine if a single variable value represents multiple goods (e.g. the survivial of all of a trolley's several passengers), provided the obtaining of these goods suitably covary in the problem.

[9] There are clear affinities between this feature of the present model and the causal modelling frameworks developed in, among others, Pearl 2009 and Spohn 2012.

which H only depends upon A and where $a \in A$ is the action of pulling the lever, we might have f_H defined by: $f_H(a) = h, f_H(a_0) = \overline{h}$. If $par(X) = \emptyset$ for $X \in E$, we take a somewhat different interpretation of f_X: in this case, $f_X : \{X\} \to X$ simply indicates the value of X known to obtain. The family of all such causal dependence functions for a given decision problem is $F = \{f_X\}_{X \in E}$. (Note that f_A is left undefined since A need have neither causal parents nor an antecedently known value.) With $\to$ and F in place, we have a precise specification of how the values of root variables lacking parents fix the values of all causally downstream variables.[10]

Lastly: *the value function*. We have thus far assumed that the various reasons that tell for and against particular courses of action (i.e. prospective goods and bads) have been represented via the set of event variables, E. However, nothing in the model so far indicates the relative strength of competing reasons, nor even the direction in which any determination of event-variables bears upon the advisability of available actions. For this purpose, we may introduce a real-valued value function, v, intended to measure the strength of the reasons provided by the various non-moral goods and bads contemplated in a given decision problem. It is tempting to define v directly on $\cup V$, that is, to make it a direct measure of the goodness/badness of the possible ways the different variables in V might turn out. On such an approach, it would be enticing to then think we could figure out the overall value of the state induced by performing a particular act a by summing up the values of each of the variable states brought about by a. However, taking this approach would involve us in difficult axiological questions about the additivity of value. While I have no principled objection to such an approach, my goal here is simply to introduce a general logic of right conduct compatible with various theories of value, and for this purpose it is desirable to commit ourselves only to the minimal axiological assumptions needed to apply the logic.

Hence, the approach I will instead take is to understand v as representing the comparative value of the *worlds* generated by a decision problem. A world specifies, for every variable in V, which of its members is true and can thus be thought of as a maximally consistent conjunction formable from members of $\cup V$. The set of all such worlds (given a fixed decision problem) will be denoted 'W'. Thus, $v : W \to \mathbb{R}$ measures the comparative non-moral desirability of maximally specific ways a particular decision problem might conceivably unfold.[11,12] For the case of decision under certainty (our current focus), it will suffice for our purposes that value be measured on a merely ordinal scale, i.e. the only morally significant information that need be encoded by v is the ordering it imposes on worlds. However, when we turn to consider decision under uncertainty, it will be necessary to view v as encoding cardinal information as well. Standard representation theorems from decision theory (duly reinterpreted) are capable of rendering these assumptions reasonable, in my view, so I won't have more to say here about the measurement scale of v.[13]

[10] We will generalize the present model to consider cases where an agent may be subjectively uncertain of the causal structure of the problem she faces in §8. The treatment there may also suggest a way to handle cases where the causal dependence functions might be objectively indeterministic, though I won't address that case here.

[11] Note that most of these conceivable ways for the problem to unfold will not, however, be genuinely possible, given $\to$ and F.

[12] I take it that the values encoded by v should plausibly be understood as agent-relative in the sense that two different agents facing largely continuous decision problems (e.g., in which $A, E, \to$, and F are the same) might nonetheless appropriately employ different value functions in arriving at their respective decisions because their individual life histories may properly render certain goods more pertinent for one agent than for another. E.g. a parent may reasonably weight the good of their child significantly higher than another agent might weight the good of the same child. Again, however, I wish to avoid making too many controversial axiological assumptions in this paper, and nothing in the formal theory I sketch or in the examples I discuss in the paper hinges upon viewing v as agent-relative vs agent-neutral.

[13] There is, I grant, one aspect of modelling relative value of worlds by means of a single ordinal value function that is objectionable to my mind. In particular, it rules out value incommensurability or cases where the values of two worlds are objectively incomparable. However, this problem is neatly solved by the trick (well known to advocates of various *imprecise decision theories*) of replacing v with a *set* of value functions. Determinate value relations between worlds could then be modelled via agreement among the members of such a set, with the possibility of disagreement making room for incommensurability. At the cost of clumsier notation, the moral theory developed in this paper could be stated and defended just as well with v replaced in this way to make room for incommensurability. I avoid doing

As noted above, I will not assume that v is necessarily additive in the sense that the ordering it imposes on W could be seen as generated by summing up a set of independent values defined on the individual members of $\cup V$. However, it is critical for applying DER that we be able to recognize individual events as good or bad in themselves, beyond assigning value to complete worlds that include them. Hence, I will assume the following *separability* condition that will allow us to view the members of any individual variable as determinately ordered from better to worse in themselves. In particular, I will assume that, for any $w, w' \in W, X \in V$, and $x_1, x_2 \in X$, $v(w(X = x_2)) > v(w(X = x_1))$ iff $v(w'(X = x_2)) > v(w'(X = x_1)$ (and the same holding when '>' is replaced by '='), where $w(X = x)$ is the world that agrees with w except perhaps for the fact that it renders x the true member of X. Note that separability allows us to derivatively infer from v, for any $X \in V$, an ordinal value scale v_X measuring the comparative value of the members of X. While not quite as strong as outright additivity, separability is nonetheless still a relatively atomistic assumption that might be questioned in certain value contexts. However, it is, I think, fairly plausible in the context of many of the important life or death decision problems that form our focus in this essay and in which the various goods and bads at stake do seem plausibly separable in the relevant sense. I will thus assume separability going forward and defer more substantive consideration of its defensability as a general axiological principle to future work.

Stepping back, we have presented a framework where decision problems are modelled as quintuples: $\langle A, E, \rightarrow, F, v\rangle$. A moral theory, recall, can then be viewed as a function mapping each such quintuple to the subset of A consisting of the morally permissible actions available to the agent facing the problem. Our task is now to consider how we might construct a plausible theory, so construed.

3 Consequentialism

Consequentialism offers a particularly simple rule for generating moral verdicts given decision problems. Note that, given a generic decision problem, $\langle A, E, \rightarrow, F, v\rangle$, F uniquely determines which world each action in A will lead to. Hence, we can define $W(a)$ to be that world. We can then state Consequentialism as:

> **Consequentialism:** Given a fixed decision problem, the morally permissible actions are those that maximize $Con(a) := v(W(a))$.

The consequentialist's advice to a moral agent is simple: just look at all the worlds that might eventuate from your various possible actions and opt to perform one of the actions that leads to a world of greatest non-moral value. Unfortunately, the straightforwardness of this advice is offset by its moral turpitude.

4 Trolley Problems

Trolley problems bring out some of the flaws in **Consequentialism**. The variation involving pushing someone onto the tracks in order to halt the trolley suffices to bring out this point. It may be more instructive though to consider a particularly clear cut alternative version of the trolley problem.

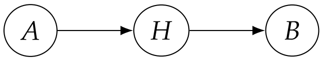
> **Dr. Evil's Trolley Problem:** Dr. Evil has five innocent victims strapped to a ticking time bomb. He has carefully arranged matters so that the bomb will go off in five minutes (killing all five victims) unless you pull a lever redirecting an empty trolley onto a track where, alas, an innocent third party (undeservedly despised by Dr. Evil) has gotten his foot stuck. Dr. Evil has set up a perfectly reliable measurement

so here only to avoid paying the notational cost.

> device that will defuse the bomb if and only if it records that the man with the stuck foot has been killed.

We might plausibly model this decision scenario via $\langle A, E, \rightarrow, F, v\rangle$, where:

- $A = \{a, a_0\}$, a being the act of pulling the lever and a_0 being the null act of doing nothing.
- $E = \{H, B\}$, with $H = \{h, \overline{h}\}$ being the event-variable corresponding to whether or not the stuck man is hit by the trolley and killed and $B = \{b, \overline{b}\}$ being the event-variable corresponding to whether or not the bomb detonates and kills the five trapped victims.
- $\rightarrow$ is given by the causal graph:



- $F = \{f_H, f_B\}$, where $f_H(a) = h$, $f_H(a_0) = \overline{h}$ and $f_B(h) = \overline{b}$, $f_B(\overline{h}) = b$.
- Values might be given by lives saved, resulting in $v(ahb) = 0, v(ah\overline{b}) = 5, v(a\overline{h}b) = 1, v(a\overline{h}\overline{b}) = 6$. (Replacing a with a_0 resulting in the same values.)

Consequentialism's verdict in this case is clear: $Con(a) = v(ah\overline{b}) = 5 > 1 = v(a_0\overline{h}b) = Con(a_0)$. Redirecting the trolley to hit the man with the stuck foot results in a better world than not pulling the lever, so consequentialists would have us run over Dr. Evil's unhappy enemy for the greater good. Many of us find this sort of application of consequentialist reasoning seriously problematic. It would be immoral to carry out Dr. Evil's dirty work in this trolley problem, even granting the great good that could come from it. So, **Consequentialism** is morally inadequate and we must look to another theory.

5 DER: A First Approximation

According to traditional proponents of DER, consequentialism goes awry (in part) by neglecting the morally relevant distinction between intending harm (either as an end or as a means) and merely forseeably causing it.[14] This line of thinking suggests an explanation as to why so many of us recoil at the idea of pulling the lever in Dr. Evil's Trolley Problem, though not in the standard case. In both examples, pulling the lever forseeably harms the man with the stuck foot, but only in the Dr. Evil version does this harm plausibly qualify as intended, since only in this latter case does the man's death constitute a necessary causal condition for the achievement of the sought after goods (i.e. the saving of the innocent five).

Another way to put this point is that in Dr. Evil's Trolley Problem, **Consequentialism** invites us to treat an innocent person's death as an instrumental reason for action. By viewing the goods that follow from this death as straightforward reasons to inflict it, the consequentialist impermissibly instrumentalizes evil for the sake of good. To avoid endorsing such tainted intentions, we need a moral decision rule that assesses the value of actions without assigning positive weight to any ill-gotten goods they may secure by means of bringing about bads. Intuitively, such a rule should value an action according to the value of its total consequences minus the value of any of those consequences that are achieved only by means of evil. An agent who followed such a decision rule, unlike the consequentialist, could not be charged with intending any harms or instrumentalizing evil for the sake of good since no consequences of such harms and evils would ever enter into their practical deliberations concerning what to

[14] The classic statement of DER within Catholic moral theology was given by Gury 1874. For some contemporary defenses of particular versions of DER (quite different both from one another and from the theory suggested here), see e.g. Boyle 1980, Quinn 1989, Cavanaugh 2006, Wedgwood 2011, Pruss 2013 and Masek 2018.

do in a motivating fashion.

To render this idea into a precise alternative to **Consequentialism**, we need to be able to demarcate good from bad consequences. There are many ways one might try to go about doing this, but one appealing suggestion is that the analysis should be essentially comparative in the sense that an act a's effect upon an event-variable X counts as a harm or an evil just in case there is some other action a' that could be performed instead that would result in a better member of X being made true. By analyzing the goodness and badness of particular consequences in this way, we can separate out an act's consequences into good ones and bad ones without reference to any resources beyond those already contained in our decision models.

Precisifying this, and assuming a fixed decision model $\langle A, E, \rightarrow, F, v\rangle$, say that variable $X \in V$ is an *aa_i-bad*, where $a, a_i \in A$, just in case $v_X(X(a_i)) > v_X(X(a))$ and $v_X(X(a_i)) \geq v_X(X(a_j))$ for all $a_j \in A$, where $X(a)$ is the member of X that would result, under the causal structure of F, if a were chosen. The aa_i-bads are simply those variables that are optimized by the choice of a_i and strictly worsened (relative to this optimal state) by the contrary choice of a. The set of all aa_i-bads may be denoted by V_{aa_i}. Letting I be an index set for A, we may define the *a-bads* as $B_a := \cup_{i \in I} V_{aa_i}$, i.e. the set of all variables that a worsens relative to some other action that might have been chosen. Now define $X(B_a)$, where $X \in E$, to be the worst member of X (relative to v) that can be obtained, given the causal structure encoded in F, by surgically replacing, in any arrangement, the members of each $Z \in an(X) \cap B_a$ by either $Z(a)$ or $Z(a_i)$ for each a_i such that $Z \in V_{aa_i}$.[15] Intuitively, $X(B_a)$ is intended to correspond to the value of X we would obtain by performing action a while selectively intervening to prevent any of the good (but not bad) consequences that flow from a via a-bads. Finally, we may define $W_p(a)$ as the maximally consistent conjunction formable from members of $\cup V$ that consists of a conjuncted with all propositions of the form $X(B_a)$, where $X \in E$. $W_p(a)$ is meant to be the *purified* world in which a realizes all of its legitimate consequences but secures no ill-gotten goods.

These formalisms allow us to take a first stab at stating a rival to consequentialism that is broadly motivated by DER:

> **Double Effect$_1$:** Given a fixed decision problem, the morally permissible actions are those that maximize $DE_1(a) := v(W_p(a))$.

This principle is no doubt rather opaque in the absence of examples, so let us gain a better grip on it by first verifying that it disagrees with **Consequentialism** in Dr. Evil's Trolley Problem. As with *Con*, to compare $DE_1(a)$ and $DE_1(a_0)$, we need to compare the relative values of two opposing worlds. However, in this case, the worlds are no longer $W(a) = ah\overline{b}$ and $W(a_0) = a_0\overline{h}b$ (i.e. the worlds that actually result from pulling or not pulling the lever) but rather $W_p(a) = ahb$ and $W_p(a_0) = a_0\overline{h}b$ (i.e. the worlds that result from pulling or not pulling the lever once we have deleted any ill-gotten gains). Given our earlier specification of v, $v(W_p(a_0)) = 1 > 0 = v(W_p(a))$. Thus, our formalization of DER rightly recommends against cooperating with Dr. Evil's nefarious plans.

So far, so good. Unfortunately, however, though I believe correct in broad outline, the theory as stated is still quite implausible. It will take a fair bit of chisolming to get it in full working order, but hopefully the product we arrive at in the end will justify the further complexities we must now introduce.

[15] This is a bit sloppy. What if there is no such unique worst member of X? Separability says it doesn't matter. Just interpret $X(B_a)$ as the *set of* worst members in that case, and define $W_p(a)$ as the class of maximally consistent conjunctions consisting of a conjuncted with *members of* $X(B_a)$. By separability, each member of $W_p(a)$ has the same value according to v (since $w, w' \in W_p(a)$ means that w can be obtained from w' by substitution of indifferent propositions), so we could meaningfully extend v to evaluate $W_p(a)$ as sharing that value as well. Similar remarks hold for the revisions of these definitions stated below. I assume there is always a unique worst member in the main text only to avoid the additional complexity involved in this way of stating matters.

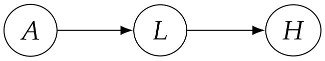
6 Refinement I: Overriding Goods

The first problem with **Double Effect**$_1$ is that it rules out too much. In some cases, unlike in the Dr. Evil Trolley Problem, goods that result from bads don't seem illegitimate or tainted in such a way that we ought to preclude them from counting as reasons for action.

> **The Lost Foot Trolley Problem:** You are now the man with your foot stuck in the trolley tracks. A trolley is barrelling down the tracks and you have no means of escape other than to pull a conveniently placed lever that will cause a mechanical saw to cut off your own foot, allowing you to leap to (relative) safety. No other lives are at stake.

We could craft a simple model of this problem as something like $\langle A, E, \rightarrow, F, v \rangle$, where:

- $A = \{a, a_0\}$, a being the act of pulling the lever and a_0 being the null act of doing nothing.
- $E = \{L, H\}$, with $L = \{l, \bar{l}\}$ being the event-variable corresponding to whether or not you lose your foot and $H = \{h, \bar{h}\}$ being the event-variable corresponding to whether or not you are struck by the trolley and killed.
- $\rightarrow$ is given by the causal graph:



- $F = \{f_L, f_H\}$, where $f_L(a) = l, f_L(a_0) = \bar{l}$ and $f_H(l) = \bar{h}, f_H(\bar{l}) = h$.
- Values given by something like: $v(alh) = 0, v(al\bar{h}) = 5, v(a\bar{l}h) = 1, v(a\bar{l}\bar{h}) = 6$. (Replacing a with a_0 resulting in the same values.)

Formally, it is easy to see that this decision problem is just a relabeling of Dr. Evil's Trolley Problem, and thus **Double Effect**$_1$ will yield the similar answer, namely that, since $DE_1(a) > DE_1(a_0)$, cutting off your foot in the Lost Foot Trolley Problem is impermissible. However, I take it this is intuitively the wrong result. In the Dr. Evil case, the loss of the trapped man's life could not legitimately be instrumentalized to further the greater good of saving five lives and this had led us to propose a general moral theory in which harms can *never* be treated as instrumental reasons for action. But the Lost Foot case teaches us that this requirement is too strong. Sometimes a harm can be *overridden* by an appropriately related good of sufficient importance such that it is after all permissible to count the good achieved via the harm as a reason for action.

To account for this overriding phenomenon, I propose that we introduce into our model of decision problems a sixth element: a binary relation $>$ on the set of propositions formable from members of $\cup V$ (together with the standard Boolean connectives) that I will call an *overriding relation*. The interpretation of $>$ is that $x > y$ holds just in case the events encoded by proposition x constitute an overriding good with respect to the events encoded by proposition y. The Dr. Evil and Lost Foot Trolley Problems can be pulled apart once we recognize that they involve different overriding relations. In the Lost Foot Trolley Problem, $\bar{h} > l$, whereas, in the Dr. Evil Trolley Problem, the $>$ relation is empty. That is, while the good of saving your life overrides the harm of losing your foot, the good of saving five people does not override the harm of killing another one.

With $>$ now introduced into our decision models, we can reformulate DER to get the right answers in both the Dr. Evil and Lost Foot Trolley Problems. The general form of the principle as stated in §5 can still be viewed as broadly correct; we just need to adjust our definition of $W_p(a)$, or the purified a-world, to account for the phenomenon of overriding goods. Say that a variable

$X \in V$ is a *non-overridden aa_i-bad* just in case X is an aa_i-bad and there is no collection of causal descendants $X_1, ...X_n \in V$ of X such that $\wedge_{i=1}^{n} X_i(a) \succ X(a)$. Let the set of all such non-overridden aa_i-bads be denoted by $V_{aa_i}^{\succ}$. We can then define the non-overridden a-bads as $B_a^{\succ} := \cup_{i \in I} V_{aa_i}^{\succ}$, with I again being an index over A. Now define, much as before, $X(B_a^{\succ})$, where $X \in E$, to be the worst member of X (relative to v) that can be obtained, given the causal structure encoded in F, by surgically replacing, in any arrangement, the members of each $Z \in an(X) \cap B_a^{\succ}$ by either $Z(a)$ or $Z(a_i)$ for each a_i such that $Z \in V_{aa_i}^{\succ}$. Intuitively, $X(B_a^{\succ})$ is intended to correspond to the value of X we would obtain by performing action a while selectively intervening to prevent any of the good (but not bad) consequences that flow from a via a-bads that are not overridden by overriding goods. Finally, we may define $W_p^{\succ}(a)$ as the maximally consistent conjunction formable from members of $\cup V$ that consists of a conjuncted with all propositions of the form $X(B_a^{\succ})$, where $X \in E$.

Our improved theory, adjusted to take account of overriding goods, can now be stated as:

> **Double Effect$_2$:** Given a fixed decision problem, the morally permissible actions are those that maximize $DE_2(a) := v(W_p^{\succ}(a))$.

Applied to the Dr. Evil Trolley Problem, **Double Effect$_2$** is identical in effect to **Double Effect$_1$** since the overriding relation is empty. However, applied to the Lost Foot Trolley Problem, **Double Effect$_2$** succeeds where **Double Effect$_1$** faltered since the harm of l is overridden by the benefit of $\bar{h}$: $v(W_p^{\succ}(a)) = v(al\bar{h}) = 5 > 1 = v(a_0\bar{l}h) = v(W_p^{\succ}(a_0))$. Thus, **Double Effect$_2$** joins **Consequentialism** in licensing cutting off your foot in the Lost Foot Trolley Problem, but joins **Double Effect$_1$** in maintaining a prohibition against killing the man with the stuck foot in the Dr. Evil Trolley Problem.

In a fully fleshed out theory, more, of course, needs to be said regarding the overriding relation. For the moment, it is simply an additional free parameter in our model that allows **Double Effect$_2$** to save the moral phenomema. An analogy with the value function v may be helpful here. Just as the proper construction of models to represent real decision problems requires prior axiological judgments to determine the value function v, so too am I now suggesting that such model construction also relies upon prior axiological judgments regarding the overriding relation $\succ$. It is beyond the scope of this paper to provide an axiological theory of the sort that could help us fix either of these parameters. My goal here is rather to explore the logical structure a moral theory should take once the values of v and $\succ$ are set.

That said, it may still be appropriate to offer a few conjectures about the sort of cases in which I believe one good is likely to override or not override another. The first thing to note is that $x \succ y$ clearly amounts to more than just $v(x) > v(y)$ (assuming we somehow extend v to evaluate arbitrary propositions), since the good of five lives, we may grant, has greater non-moral value than the good of one and yet, as we have seen in the Dr. Evil example, the good of saving the five does not override the harm of losing the one, in the sense of rendering it acceptable to treat the loss of the one as an instrumental reason in acting to save the five. So the overriding relation should not be thought reducible to a more fleshed out value function; its presence in our decision models is not redundant. Nor should it be thought, I think, that $x \succ y$ amounts simply to the value gap between x and y (i.e. $v(x) - v(y)$) being sufficiently large (even assuming that value is measurable on a scale where such differences make sense). If we want to allow, like Anscombe appeared to,[16] that certain evils may be such that no good can override them, then, without assuming problematic axiological principles (e.g. that v is bounded), such a proposal is a dead end.

If the overriding relation is not directly a matter of relative value, however, then it does call for some other explication that could indicate its presence or absence. Two observations

[16] See, e.g., Anscombe 1961.

might be ventured here. First, in general, many of the most plausible instances in which it seems that $x > y$ appear to be ones in which x benefits the very agent for whom y qualified as a harm. E.g. in the Foot Loss Trolley Problem, the very same agent who was harmed by the loss of their foot (i.e. you) was also benefited by the resultant overriding good of having their life spared. So perhaps this fact should feature somehow in a full explication of the overriding relation. However, this observation shouldn't be carried too far since we may note, as a second observation, that not all instances of the overriding relation appear to be of this sort. In particular, it also seems plausible that in some instances where the kind of harm encoded by y is of a significantly less serious sort than the benefit encoded by x, x can override y even if the harms/benefits accrue to different agents. For example, if causing one person a short headache could lead to another's life being saved, it may well be that we should view the latter individual's good of life as overriding the former's concern to avoid minor discomfort. However we ultimately spell out a theory of the overriding relation, I have no more to say about the matter here and shall take it as intuitively fixed in the problems we consider.

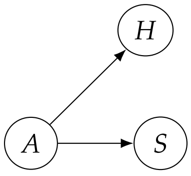
7 Refinement II: 'Closeness'

The move from **Double Effect**$_1$ to **Double Effect**$_2$ is well motivated: we need to account for the phenomenon of overriding goods. However, this move may also seem to call into question DER's capacity to deliver many of the verdicts that we had originally hoped it would yield. In particular, one may question whether **Double Effect**$_2$ is even capable of separating the classic pair of trolley problems that opened this essay. Recall:

> **The Standard Trolley Problem:** A runaway trolley carrying five passengers barrels down an unfinished track toward a precipitous cliff. You have time to pull a nearby lever to switch the trolley onto an alternate track, averting disaster for its passengers. Unhappily, an oblivious third party has gotten his foot stuck attempting to cross the second track. Pulling the lever means life for the trolley's passengers but death for the man with the trapped foot. Failing to pull the lever reverses these fortunes.
>
> **The Fat Man Trolley Problem:** A runaway trolley carrying five passengers barrels down an unfinished track toward a precipitous cliff. You have time to push on to the track a nearby man of sufficient mass to stop the trolley, averting disaster for its passengers. Of course, if you were to push the man, the force of the trolley's impact on his body would be lethal for him.

We might plausibly model the first decision scenario, i.e. the Standard Trolley Problem, via something like $\langle A, E, \rightarrow, F, v, >\rangle$, where:

- $A = \{a, a_0\}$, a being the act of pulling the lever and a_0 being the null act of doing nothing.
- $E = \{H, S\}$, with $H = \{h, \overline{h}\}$ being the event-variable corresponding to whether or not the stuck man is hit by the trolley and killed and $S = \{s, \overline{s}\}$ being the event-variable corresponding to whether or not the five trolley passengers are saved.
- $\rightarrow$ is given by the causal graph:

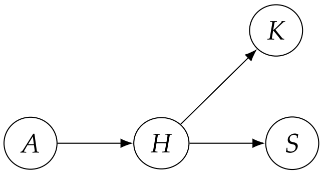


- $F = \{f_H, f_S\}$, where $f_H(a) = h$, $f_H(a_0) = \overline{h}$ and $f_S(a) = s$, $f_S(a_0) = \overline{s}$.

- Values might be given by lives saved, resulting in $v(ahs) = 5, v(ah\bar{s}) = 0, v(a\bar{h}s) = 6, v(a\bar{h}\bar{s}) = 1$. (Replacing a with a_0 resulting in the same values.)
- $\succ = \emptyset$.

At first glance, it may appear that a simple relabeling of the Standard Trolley Problem could yield a suitable model for the Fat Man Trolley Problem as well. However, in the fat man version, unlike in the standard one, part (though not all) of the harm effected by pushing the man is causally prior to the good of saving the trolley's passengers, and this fact would not be captured in any relabelling of *STP*. Hence, a more adequate model of the Fat Man Trolley Problem may be along the lines of: $\langle A, E, \rightarrow, F, v\rangle$, where:

- $A = \{a, a_0\}$, a being the act of pushing the man and a_0 being the null act of doing nothing.
- $E = \{H, K, S\}$, with $H = \{h, \bar{h}\}$ being the event-variable corresponding to whether or not the man is hit by the trolley, $K = \{k, \bar{k}\}$ being the event-variable corresponding to whether or not he is consequently killed, and $S = \{s, \bar{s}\}$ being the event-variable corresponding to whether or not the five trolley passengers are saved.
- $\rightarrow$ is given by the causal graph:



- $F = \{f_H, f_k, f_S\}$, where $f_H(a) = h$, $f_H(a_0) = \bar{h}$, $f_K(h) = k$, $f_k(\bar{h}) = \bar{k}$, and $f_S(h) = s$, $f_S(\bar{h}) = \bar{s}$.
- Values might be given (roughly) by lives saved, adjusted to take account of the small independent benefit of the pushed man not being hit: $v(ahks) = 5, v(ahk\bar{s}) = 0, v(a\bar{h}ks) = 5.1, v(a\bar{h}k\bar{s}) = 0.1, v(ah\bar{k}s) = 5.9, v(ah\bar{k}\bar{s}) = 0.9, v(a\overline{hk}s) = 6, v(a\overline{hk}\bar{s}) = 1$. (Replacing a with a_0 resulting in the same values, assuming, perhaps unreasonably, that the push itself does not result in any independent harm.)
- $s \succ h$.

Applied to these models, **Double Effect**$_2$ agrees with **Consequentialism** in rightly recommending pulling the lever in the Standard Trolley Problem but wrongly recommending pushing the man in the Fat Man Trolley Problem. The reason for the latter verdict is that I have assumed here that the good of saving ten lives overrides the harm of the fat man being hit by a trolley, when that harm is treated as independent from his resultant loss of life in perpetuity.[17] We run up here against the dreaded *closeness problem*, the rock upon which many have supposed DER to ultimately founder.[18] Granted, the good of saving ten lives does not override the harm of ending one, but, according to **Double Effect**$_2$, such overriding is unnecessary for the good in question to justify pushing the fat man because the good is, speaking strictly, brought about not by the grave harm of the fat man's irreversible loss of life, but only by the lesser harm of the fat

[17] This particular supposition may well be false. Perhaps saving ten lives does not override the serious harm of being hit by a trolley, even when such a harm is treated as a separate ill from the death it causes. If one believes this, then **Double Effect**$_2$ gets the Fat Man Trolley Problem just right. However, even those of this opinion will still likely find it ultimately necessary to adopt the amendment of the current section to handle other cases.

[18] The closeness objection was originally considered by Foot 1967, but perhaps most famously pressed by Bennett 1995. See also Davis 1984, as well as Fischer, Ravizza, and Copp 1993 in the context of raising a closeness objection to Warren Quinn's version of DER. For a more recent discussion, see Nelkin and Rickless 2015. A common move among recent proponents of DER in response to the closeness problem has involved retreating to thin construals of DER that make no pretense of solving the trolley problem, see e.g. Liao 2016 and Masek 2018. Against this trend, I will defend a solution here that still allows us to see DER as relevant to the trolley problem.

man's body being impacted by a trolley in a harsh way for a short period of time. (To see that the evils of being hit by a trolley and total loss of life are indeed logically distinct and that only the former is causally relevant to bringing about the good of saving the trolley passengers in the Fat Man Trolley Problem, just imagine a circumstance in which the fat man is miraculously resuscitated after his grizzly encounter with the trolley.)

For a good to be admissibly entertained as a reason in favor of a particular action, it is thus not sufficient that the good simply override any of its undesirable causal predecessors. Sometimes the undesirable causal predecessors are so intimately linked (so "close") to further negative consequences that, in order to qualify as an admissible instrumental reason for action, the good must override these further ills as well. The challenge posed by the closeness problem is to delineate precisely which harms count as sufficiently close to one another to merit joint treatment in this regard. I believe our modelling framework suggests an answer to this problem. In particular, I want to suggest that we can understand closeness in terms of the connections bridged between variables by $\rightarrow$.

To make this precise, we will again need to help ourselves to some novel terminology and then rebuild an appropriate notion of purified worlds, much as before. First, say that an aa_i-bad $X \in V$ is *continuous* with an aa_i-bad $Y \in V$ just in case there exists a path of aa_i-bads in V connecting X and Y via $\rightarrow$. We denote the set of aa_i-bads continuous with X as '$C_{aa_i}(X)$'. Now say that a variable $X \in V$ is an *undefeated aa_i-bad* just in case X is an aa_i-bad and there is no collection of causal descendants $X_1, ...X_n \in V$ of X such that $\wedge_{i=1}^{n} X_i(a) \succ \wedge_{j=1}^{m} Y_j(a)$, where $\{Y_1, ..., Y_m\} = C_{aa_i}(X)$. Let the set of all such undefeated aa_i-bads be denoted by $V^*_{aa_i}$. We can then define the undefeated a-bads as $B^*_a := \cup_{i \in I} V^*_{aa_i}$, with I again being an index over A. Now define, much as before, $X(B^*_a)$, where $X \in E$, to be the worst member of X (relative to v) that can be obtained, given the causal structure encoded in F, by surgically replacing, in any arrangement, the members of each $Z \in an(X) \cap B^*_a$ by either $Z(a)$ or $Z(a_i)$ for each a_i such that $Z \in V^*_{aa_i}$. Intuitively, $X(B^*_a)$ is intended to correspond to the value of X we would obtain by performing action a while selectively intervening to prevent any of the good (but not bad) consequences that flow from a via a-bads that are undefeated by overriding goods (i.e. a-bads that are part of a non-overriden collection of continuous bads). Finally, we may define $W^*_p(a)$ as the maximally consistent conjunction formable from members of $\cup V$ that consists of a conjuncted with all propositions of the form $X(B^*_a)$, where $X \in E$.

Naturally, this leads to a new statement of DER, which I propose we can (finally) recognize as a reasonable alternative to **Consequentialism** in cases of decision under certainty:

> **Double Effect*:** Given a fixed decision problem, the morally permissible actions are those that maximize $DE(a) := v(W^*_p(a))$.

Double Effect* preserves the virtues of **Double Effect**$_1$ and **Double Effect**$_2$ while correcting their flaws. It is straightforward to verify that pulling the lever in the the Foot Loss Trolley Problem and abstaining in the Dr. Evil Trolley Problem are uniquely recommended by **Double Effect***. In the Standard Trolley Problem, it is equally simple to show that all three formulations of DER agree with **Consequentialism** in endorsing pulling the lever. The infamous Fat Man Trolley Problem, however, allows **Double Effect*** to demonstrate its worth over **Double Effect**$_2$. In this case, even granting that the good of saving five lives overrides the harm of being hit by a trolley, when the harm in question is recognized as continuous with the pushed man's long term loss of life, it can be seen as undefeated by the good consequences, and hence: $DE(a) = v(W^*_p(a)) = v(ahk\bar{s}) = 0 < 1 = v(a\overline{hk}\bar{s}) = v(W^*_p(a_0) = DE(a_0))$, leaving us with a sensible recommendation against pushing an innocent man in front of a runaway trolley.

There are many further cases from the literature on both trolley problems and DER that it would be illuminating to test **Double Effect*** on. For example, can we employ **Double Effect*** to justify the intuitive contrast between terror and strategic bombing in military ethics? Between euthanasia and death-hastening pain management in medical ethics? Etc.[19] Unfortunately, space constraints preclude a careful examination of these cases here. For my part, I think, when these cases are properly modelled, **Double Effect*** can indeed yield the right results. However, this cautious phrasing highlights the essential model relativity of **Double Effect***. The recommendations it offers are sensitive to how precisely we opt to model a real-life decision problem. How we go about carving up the problem into distinct variables can make a substantive difference to its verdicts. Some may worry that this model relativity risks rendering the theory largely vacuous, an objection I will address by offering some general heuristics for how to discriminate between more and less adequate models of a given decision problem in §9. First, however, there is a final extension of **Double Effect*** that needs stating.

8 Decision Under Uncertainty

Most philosophical discussion of both trolley problems and DER has operated under a policy of feigned certainty regarding the causal structure of contemplated decision problems. We assume, for example, that the agents facing cases like those considered so far know the exact causal structure of these problems as they face them. But life is rarely so kind. Typically, we are uncertain regarding the exact causal structure of the decision problems we face. In such cases, we are forced to entertain various causal hypotheses regarding what effect our actions may have in the world and somehow make a decision that takes into account our uncertainty concerning which of these hypotheses is true. One of the principal advantages I claim for **Double Effect*** as an approach to modelling DER is the ease with which it generalizes to an account of decision making under uncertainty.

For concreteness, consider a case like the following (no doubt somehow the handiwork of Dr. Evil):

> **The Uncertain Trolley Problem:** A runaway trolley carrying five passengers barrels down an unfinished track toward a precipitous cliff. You have time to pull a nearby lever which you know will do exactly one of two things: (i) switch the trolley onto an alternate track, averting disaster for its passengers, but killing an oblivious third party who has gotten his foot stuck attempting to cross the second track, or (ii) trigger a crane to push on to the main track a nearby man of sufficient mass to stop the trolley, averting disaster for its passengers, but killing the pushed man. You are equally confident of each of these possibilities.

The Uncertain Trolley Problem is clearly a mixture of the standard and fat man trolley cases. If you knew how the switch was designed, the problem would collapse into one of these two. But you don't. In the language of our decision models, you are uncertain what form $\rightarrow$ and F take on, so you have no way of representing the Uncertain Trolley Problem in the framework developed so far. For problems such as this, we need to extend our simple decision models to account for uncertainty regarding causal structure. There are several models of uncertainty one might employ to do so, but by far the most familiar is orthodox probability theory. Given its relatively dominant standing, this is the approach I will adopt here, though I have no objection to exploring other models of uncertainty as well.[20]

Fixing a set of variables V, say that a *causal model on* V is a pair $\langle \rightarrow, F \rangle$, where $\rightarrow$ is a causal ordering on V and F is a causal dependence function with respect to $\rightarrow$. Heretofore, we had assumed that every decision problem specifies a unique causal model. But let us now dispense

[19] On the trolley front, the infamous 'Loop Case' of Thomson 1985 and the 'Trolley Track Case' of Kamm 2007 would be instructive to consider from the perspective of **Double Effect***. In these cases, how we characterize the exact causal structure of the situation matters for what **Double Effect*** ultimately recommends. When causal structure is properly specified, however, I don't find any of the theory's recommendations in these cases objectionable.

[20] For an excellent survey of various models of uncertainty one might appropriate for our present purposes, see Halpern 2017.

with that assumption and model decision problems as sextuples of the form $\langle A, E, v, \succ, CM_V, P\rangle$, where A, E, v, and $\succ$ are interpreted as before, while CM_V is a finite collection of causal models on V and $P : CM_V \to [0, 1]$ is a probability mass function satisfying $\sum_{i \in I} P(x_i) = 1$, where I is an index set for CM_V. The interpretation is that CM_V is the set of all causal models on V deemed possible by the agent facing the decision problem at hand, and P measures her degree of belief (or level of confidence) that each member of CM_V corresponds to the actual (but unknown) causal structure of the problem she faces. (Note that we can still accommodate our previous examples of decision under certainty in this new framework by allowing that CM_V may be a singleton.)

With our decision models generalized in this way, we can capture the Uncertain Trolley Problem as: $\langle A, E, v, \succ, CM_V, P\rangle$, where:

- $A = \{a, a_0\}$, a being the act of pushing the man and a_0 being the null act of doing nothing.
- $E = \{H_1, H_2, K, S\}$, with $H_1 = \{h_1, \overline{h_1}\}$ being the event-variable corresponding to whether or not the man on the tracks is hit and killed by the trolley, $H_2 = \{h_2, \overline{h_2}\}$ being the event-variable corresponding to whether or not the potentially pushed man is hit by the trolley, $K = \{k, \overline{k}\}$ being the event-variable corresponding to whether or not the potentially pushed man is consequently killed, and $S = \{s, \overline{s}\}$ being the event-variable corresponding to whether or not the five trolley passengers are saved.
- Values may be identified (roughly) with lives saved, adjusted to take account of the small independent benefit of the pushed man not being hit: $v(ah_1\overline{h_2k}s) = 11, v(a\overline{h_1h_2k}s) = 12, v(ah_1h_2k\overline{s}) = 0$, etc. (Replacing a with a_0 resulting in the same values, assuming, perhaps unreasonably, that the push itself does not result in any independent harm.)
- $s \succ h$.
- $CM_V = \{\langle \to_1, F_1\rangle, \langle \to_2, F_2\rangle\}$, where $\langle \to_1, F_1\rangle$ corresponds to the causal model at play in the Standard Trolley Problem and $\langle \to_2, F_2\rangle$ corresponds to the causal model at play in the Fat Man Trolley Problem (each extended in the natural way to take account of all present variables).
- $P(\langle \to_1, F_1\rangle) = P(\langle \to_2, F_2\rangle) = 0.5$.

Double Effect*, for all its merits, falls silent on the Uncertain Trolley Problem, which lacks the right form for our theory of decision under certainty to yield any direct advice. However, given that DE gives us a way of assigning values to actions, conditional upon knowledge of the right causal hypotheses, we can naturally generalize **Double Effect*** to handle decisions under uncertainty, in an analogous way to how consequentialists might generalize their theory. To do so, I will employ $W_p^*(a, x_i)$ to designate the familiar purified world $W_p^*(a)$ on the assumption that x_i is the true member of CM_V. We can then state:

> **Generalized Double Effect*:** Given a fixed decision problem involving uncertainty, the morally permissible actions are those that maximize $E_P[DE(a)] = \sum_i P(x_i)v(W_p^*(a, x_i))$.

According to **Generalized Double Effect***, an agent faced with a problem of decision under uncertainty ought to value her options according to the expectation of the value of their associated purified worlds, where the expectation is taken relative to the probability measure capturing her uncertainty. This is just like an approach a consequentialist might adopt in handling decisions under uncertainty, except that, as in decision under certainty, the values of relevance to DER are the values of purified worlds whereas for the consequentialist they would be the values of resultant worlds. Applied to the Uncertain Trolley Problem, **Generalized Double**

Effect* yields the result that you ought to pull the lever since:

$$\begin{aligned} E_P[DE(a)] &= P(\langle \rightarrow_1, F_1 \rangle) v(W^*_p(a, \langle \rightarrow_1, F_1 \rangle)) + P(\langle \rightarrow_2, F_2 \rangle) v(W^*_p(a, \langle \rightarrow_2, F_2 \rangle)) \\ &= 0.5 \times v(ah_1\overline{h_2k}s) + 0.5 \times v(a\overline{h_1}h_2k\overline{s}) \\ &= 0.5 \times (11 + 1) \\ &= 6 \end{aligned}$$

$$\begin{aligned} E_P[DE(a_0)] &= P(\langle \rightarrow_1, F_1 \rangle) v(W^*_p(a_0, \langle \rightarrow_1, F_1 \rangle)) + P(\langle \rightarrow_2, F_2 \rangle) v(W^*_p(a_0, \langle \rightarrow_2, F_2 \rangle)) \\ &= 0.5 \times v(a_0\overline{h_1h_2k}\overline{s}) + 0.5 \times v(a_0\overline{h_1h_2k}\overline{s}) \\ &= 0.5 \times (2 + 2) \\ &= 2 \end{aligned}$$

Hence, $E_P[DE(a)] > E_P[DE(a_0)]$.

Generalized Double Effect* stays true to the motivating spirit of DER by not instrumentalizing evil for the sake of good. While an agent who opts to pull the lever in the Uncertain Trolley Problem risks pushing the fat man on to the tracks, neither this possible evil nor any of its potential good consequences is treated by the decision rule as any reason to pull the lever. In fact, a gamble in which the lever would either function as in the Standard Trolley Problem or else simply fail to operate would be strictly preferred, according to **Generalized Double Effect***, to the gamble it actually recommends in the Uncertain Trolley Problem, revealing that its recommendation, unlike that of **Consequentialism**, is in no way driven by the prospect of any ill-gotten goods.

9 Model Relativity

Given a formal representation of a decision problem of the specified sort, **(Generalized) Double Effect*** straightforwardly (in fact, computably) fixes the set of morally permissible actions. But models are not reality. An agent's own mental understanding of the decision problem she faces will invariably be far richer than any quintuples or sextuples a formal modeller can cook up. And it is clearly the former rather than the latter that is determinative of the moral quality of the agent's possible actions. The rather artificial models of this essay will be useful then in pinning down the morally permissible options in hypothetical decision problems only to the extent that they succeed in capturing (albeit in simplified form) the morally relevant features of the actual problems they seek to represent.

There are two respects in which a decision model may fail to capture the structure of a real decision scenario. First, it may simply get things wrong and represent features of the scenario as other than they are. Such errors can concern the nature of variables as well as their evaluative significance. A representation of the Standard Trolley Problem that reckoned stopping the trolley with telepathy as an available action or that treated five lives as strictly less valuable than one would involve errors of this kind. Of course, the possibility of making such mistakes doesn't generate much of an objection either to the modelling enterprise or to our theory of **Double Effect***. It is hardly surprising or troubling that inaccurate descriptions of a problem should lead to unreliable verdicts concerning its solution. The blame for any poor conclusions reached on account of such errors clearly lies with the modeller rather than with the modelling framework.

Potentially more troubling for our framework itself is the second respect in which a decision model may distort the moral character of the problem it aims to represent: incompleteness.

As noted, any decision model built in the framework proposed here, even if a scrupulously accurate reflection of a hypothetical decision problem, is bound to leave out various features of the problem as it would be intuitively grasped by a thoughtful human agent. Such incompleteness would be little cause for concern were it not for the fact that different incomplete, though equally accurate, representations of a single scenario can lead a moral theory to issue divergent permissibility verdicts. We saw this above when we noted that the Fat Man Trolley Problem could be accurately, though unsatisfactorily, modelled by a simple reinterpretation of the model we employed to characterize the Standard Trolley Problem, in which case **Double Effect*** would offer the obscene advice of pushing the man. It evidently matters then which representation of a given problem we use as input when asking our theory to separate out our permissible and impermissible options.

Some (at least loose) guidance concerning model construction is thus critical for practical application of the theories discussed here. If it were feasible, we could keep the guidance simple: just model everything. Representing every logically distinct feature of the problem at hand via a maximally fine-grained 'grand world' model would certainly provide us with a suitable representation for reliable application of **Double Effect***, but doing so is clearly well beyond our ken. Fortunately though, we needn't produce such elaborate models to be reasonably confident that our representation is adequate enough for the application of our theory. Coarse-grained 'small world' models will do just fine provided they represent just enough of the problem at hand. I think we can helpfully boil down the 'just enough' in terms of three basic rules or heuristics of model construction.

First, as emphasized from the beginning, to provide an adequate representation of a given decision problem, a model must represent, via its event variables, every reason (in the sense of prospective goods and bads) that tells for or against each of the available actions. Or, at least, it needs to represent every significant such reason, i.e. every reason that might have a real chance of impacting the relative standing of the agent's options. We could call this the *Rule of Significant Reasons*. There is nothing special about **Double Effect*** in this regard. **Consequentialism** has equal need of the same rule. For neither theory, in a trolley problem of the sort we've considered, will it do to forget about the plight of the man with the stuck foot or the prospective life of the trolley's third passenger, etc. Somewhere in the decision model, these reasons must appear so that their attainment/frustration can be accounted for in measuring the value of the model's various possible worlds for purposes of evaluating actions.

The Rule of Significant Reasons is the only one **Consequentialism** has need of. Once all significant reasons are accounted for in the model, the consequentialist runs no further risk of having underspecified her decision situation for purposes of applying her decision rule. Not so for the proponents of DER. Our theory is sensitive to more than the value of the anticipated effects of an agent's actions, taking into account also the causal relations amongst the action and such effects. There are two ways in which a decision model might leave out causal information of significance to the application of **Double Effect***. In the first instance, a model might fail to depict the causal dependence of a good event upon the realization of a causally prior bad event. For example, if we modelled the Fat Man Trolley Problem in a manner analogous to the Standard Trolley Problem, we would have a model with just this feature. The model would fail to reflect that the saving of the five passengers is the causal byproduct of a significant harm to the man pushed in front of the trolley, and hence, applying **Double Effect*** to this inadequate model, we would miss that the benefit of saving the five needs to be excluded from consideration when weighing the merits of pushing the man. The second general rule of model construction is thus what we might call the *Rule of Prior Bads*: if a significant good could potentially be caused by a bad in the problem, then this must be reflected in the constructed model.

The second class of causal facts that needs to be reflected in a decision model in order for **Double Effect*** to be reliably applied is the class of facts about the causal (dis)continuity of represented bads. If two bad events are, in the agent's intuitive grand world representation of her decision situation, causally *discontinuous* bads, i.e. are not causally linked by a pathway consisting entirely of bads, then these events should not be represented as continuous bads within the small world formal model we construct of the situation either. Call this the *Rule of Continuous Bads*. If we violate this rule, we leave ourselves liable to wrongly failing to recognize an overridden bad as defeated by its overriding good on account of treating it as continuous with non-overridable bads with which it is actually discontinuous. This could preclude an agent from legitimately taking account of the relevant overriding good in her decision making.

It should be stressed that, useful as I take them to be, the rules of Significant Reasons, Prior Bads, and Continuous Bads offer no mechanical recipe for constructing decision models. They are intended as nothing more than heuristics or rules of thumb for the careful modeller to make use of as she goes about her art. For any real world decision problem a human agent may actually face, there will be an indefinite number of models that, from the perspective of **Double Effect***, constitute adequate formal representations of its morally relevant features. These models will each carve up the space of event variables differently, many being cluttered with needless details and distinctions. But all should, if the modeller has done her job right, yield the same output when run through **Double Effect***.

10 Further Work

I started this essay with the goal of developing a formal theory of DER, as well as an attendant modelling framework, in order to advance one sort of non-consequentialist response to the trolley problem and render DER algorithmically implementable and serviceable in contexts of decision under uncertainty. Hopefully, we have made substantial progress in this regard. Still, there is clearly much remaining work to be done. In this closing section, I would like to highlight some of the many important questions I have left open, as well as new directions one could take in expanding or revising the models deployed here.

In order to focus on the logical structure of the proposed decision rules, I have largely punted on difficult axiological questions. But in a fuller treatment, such questions merit far deeper reflection, given the significance of both the value function and the overriding relation for any application of **Double Effect***. The defensibility of separability, for example, deserves closer inspection. If separability holds up, might it be defensible to extend this principle even further and insist that v generally take on an additive form? I have assumed that it does in the trolley problems considered, but stopped short of insisting upon this as an assumption of the theory. Computationally, additive value functions have significant advantages over non-additive ones, but they commit us to an even more atomistic approach to understanding value than is already (perhaps troublingly) required by separability.[21] And, of course, delineating how the value function ought to be determined in the first place, regardless of the form we discover it to take on, involves answering many controversial questions about the nature of objective value. If we reject (as I think we should) a purely subjective approach to well-being that assesses goods and bads in terms of arbitrary preference satisfaction, then the difficulty of measuring value in a way suitable for the application of **Double Effect*** and its generalized version is perhaps in some ways heightened.[22] Far more serious inquiry into the measurement of value from an objective goods point of view is certainly called for. (And this is still to say nothing of determining the overriding relation, which I have suggested (§6) must also be settled somehow.)[23]

[21] An additive value function can be determined by specifying $|\cup V|$ many real numbers, far fewer than the $2^{|\cup V|}$ required in the current non-additive formulation of the theory.

[22] Though subjectivist accounts face well known difficulties of their own involving interpersonal comparison of utilities. For discussion, see Skyrms and Narens 2020.

[23] One suggestion, though perhaps not terribly satisfying, would be to take what might be seen as a quasi-Aristotelian virtue ethics approach to fixing v and $>$. The idea would be to measure values according to a subjective method (i.e. via

Secondly, all of the decision problems discussed here have been what decision theorists call *static* or *one-shot* decision problems in which an agent has to make a single choice (e.g. whether or not to pull a lever) at a fixed point in time. But typically our real-life decision problems are *dynamic*. They evolve through time and involve multiple choice points, perhaps interspersed with various learning events. For example, perhaps pulling a lever in a certain way leads to a subsequent choice of pulling another lever. Our decision models should be extended to take account of such problems. In particular, we ought to allow the existence of multiple distinct action sets located at various vertices of a decision model's causal network. I do not envision any particular difficulties with carrying out such an extension, but the properties exhibited by **Double Effect*** in such a setting out to be examined.[24]

Finally, there are some aspects of our models that may perhaps be fruitfully altered or adjusted further. For example, I have essentially assumed in our models a fairly simplistic counterfactual analysis of harm: $X(a)$ is a bad just in case $X(a')$ is better, for available acts a and a'. In typical decision problems, I think this is good enough. But in relatively marginal cases (e.g. involving causal overdetermination) this might prove problematic. Incorporating more sophisticated causal analyses into the model may thus be a worthwhile project in this regard.[25] Alternatively, we should not dismiss the possibility of exploring different ways of carving up events into goods and bads altogether.[26] Another aspect of **Double Effect*** that I have not commented on directly but that has surely drawn alarm from some non-consequentialist readers is its maximizing character. The theory escapes many of the infamous permissibility verdicts of **Consequentialism** (e.g. its recommendation of pushing the man in the Fat Man Trolley Problem) but not its general moral demandingness and banishment of supererogation. If one finds this aspect of **Double Effect*** objectionable (as I am inclined to), then perhaps some sort of satisficing parameter must be introduced into our models. How to go about doing so in a way that doesn't disturb the good results **Double Effect*** has already achieved would require careful consideration, however. I make no claim that **Double Effect*** as defended here offers a maximally detailed map of morality's entire landscape; it will have reached the goals set by this essay if it merely succeeds as a reasonably serviceable (albeit idealized) map of an important swath of that terrain.

Works Cited

Anscombe, Elizabeth. *Intention*. Oxford: Basil Blackwell, 1957. Print.

—–."War and Murder". *Nuclear Weapons: A Catholic Response*. New York: Sheed and Ward, 1961. 45–62. Print.

Bennett, Jonathan. *The Act Itself*. New York: Oxford University Press, 1995. Print.

Boyle, Jospeh. "Toward Understanding the Principle of Double Effect". *Ethics* 90 (1980): 527–38. Print.

Cavanaugh, T.A. *Double-Effect Reasoning: Doing Good and Avoiding Evil*. Oxford: Clarendon Press, 2006. Print.

Davis, Nancy. "The Doctrine of Double Effect: Problems of Interpretation". *Pacific Philosophical Quarterly* 65 (1984): 107–23. Print.

methods akin to those commonly used by economists to gauge utility) but with a carefully chosen subject: the virtuous or exemplary agent. Of course, this assumes that we can recognize subjects as virtuous and that the preferences of virtuous agents are a reliable indicator of objective goodness (and, thus, would tend to agree). The latter assumption may be plausible enough, though perhaps the former is more questionable.

[24] An obvious question to ask here concerns the *dynamic consistency* of **Generalized Double Effect***, an attractive property I have explored previously in purely decision theoretic contexts, e.g. in Rothfus 2020.

[25] Here, some of the more careful models of Pearl 2009, Spohn 2012, or Halpern 2016 may be relevant.

[26] Might a more absolute (i.e. less comparative) account of good and bad be possible? I am skeptical, but the issue is worth considering more deeply.

Fischer, John Martin, Mark Ravizza, and David Copp. "Quinn on Double Effect: The Problem of"Closeness"". *Ethics* 103 (1993): 707–25. Print.

Foot, Philippa. "Abortion and the Doctrine of Double Effect". *Oxford Review* 5 (1967): 28–41. Print.

Greene, Joshua. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York: Penguin Press, 2013. Print.

Gury, J.P. *Compendium Theologiae Moralis*. Regensburg: Georgii Josephi Manz, 1874. Print.

Halpern, Joseph. *Actual Causality*. Cambridge, MA: MIT Press, 2016. Print.

——.*Reasoning About Uncertainty*. Cambridge, MA: MIT Press, 2017. Print.

Jeffrey, Richard. *The Logic of Decision*. Chicago: University of Chicago Press, 1965/1983. Print.

Kamm, Frances. *Intricate Ethics*. Oxford: Oxford University Press, 2007. Print.

Liao, S. Matthew. "The Closeness Problem and the Doctrine of Double Effect: A Way Forward". *Criminal Law and Philosophy* 10(4) (2016): 849–863. Print.

Masek, Lawrence. *Intention, Character, and Double Effect*. South Bend: Notre Dame Press, 2018. Print.

Nelkin, Dana and Samuel Rickless. "So Close, Yet So Far: Why Solutions to the Closeness Problem for the Doctrine of Double Effect Fall Short". *Noûs* 49(2) (2015): 376–409. Print.

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, 2009. Print.

Pruss, Alexander. "The accomplishment of plans: a new version of the principle of double effect". *Philosophical Studies* 165(1) (2013): 49–69. Print.

Quinn, Warren. "Actions, Intentions, and Consequences: The Doctrine of Double Effect". *Philosophy and Public Affairs* 18(4) (1989): 334–351. Print.

Rothfus, Gerard. "Dynamic Consistency in the Logic of Decision". *Philosophical Studies* 117(12) (2020): 3923–34. Print.

Skyrms, Brian and Louis Narens. *The Pursuit of Happiness: Philosophical and Psychological Foundations of Utility*. Oxford: Oxford University Press, 2020. Print.

Spohn, Wolfgang. *The Laws of Belief*. Oxford: Oxford University Press, 2012. Print.

Thomson, Judith Jarvis. "Killing, Letting Die, and the Trolley Problem". *The Monist* 59(2) (1976): 204–217. Print.

——."The Trolley Problem". *The Yale Law Journal* 94 (1985): 1395–1415. Print.

Wedgwood, Ralph. "Defending Double Effect". *Ratio* 24(4) (2011): 384–401. Print.