A Causal Modeler's Guide to Double Effect Reasoning

Gerard J. Rothfus
University of North Carolina, Chapel Hill

June 2024

(Forthcoming in *Philosophy and Phenomenological Research*)

Abstract

Trolley problems and like cases are often thought to show the inadequacy of purely consequentialist moral theories. In particular, they are often taken to reveal that consequentialists unduly neglect the moral significance of the *causal structure* of decision problems. To precisify such critiques and one sort of deontological morality they motivate, I develop a formal modeling framework within which trolley problems can be represented as suitably supplemented structural causal models and various consequentialist and double effect-inspired moral theories can be viewed as disagreeing over the inputs of a common decision rule.

1 Introduction

There is a tragic predicament nearly every philosopher has faced at some point in her career, though hopefully only in her mind.[1]

> **Bystander:** A runaway trolley barrels down a track toward five unaware transit workers. You have time to pull a nearby lever that will switch the trolley onto a sidetrack, averting disaster for these transit workers. Unhappily, an oblivious third party has gotten his foot stuck attempting to cross the second track. Pulling the lever

[1]For the most famous versions of the sort of trolley cases discussed here, see (Foot, 1967) and (Thomson, 1976, 1985). I take liberty in presenting the cases as best suits present purposes, however, and don't claim faithfulness to their original tellings.

means life for the transit workers 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 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 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 **Bystander** make trouble for consequentialism.

> **Footbridge:** A runaway trolley barrels down a track toward five unaware transit workers. You have time to pull a nearby lever that will open a trap door on a footbridge passing over the track. Standing on the trap door is a man of sufficient mass to bring the trolley to a halt following impact, averting disaster for the transit workers. Unhappily, the force of the trolley's impact on the dropped man's body would kill him. Pulling the lever means life for the transit workers but death for the dropped man. Failing to pull the lever reverses these fortunes.

The benefits notwithstanding, many of us inclined to pull the lever in **Bystander** find the prospect of pulling it in **Footbridge** deeply objectionable. Consequentialism, however, fails to account for this asymmetry in our moral judgments, since the prospective goods and bads that favor/disfavor pulling the lever in **Bystander** appear perfectly symmetric to those that favor/disfavor pulling it in **Footbridge**.

A common take among deontologists holds that consequentialism goes awry here because it fails to attend to how the prospective goods and bads wrought by an agent's actions are *causally structured*. While in **Bystander** the man on the sidetrack's being crushed is a foreseeable downstream effect of turning the trolley, it is in a no way a causal precondition for the five's being saved. Matters are otherwise in **Footbridge**, where the dropped man's being crushed is the very causal means by which the trolley is stopped and the five are saved. According to proponents of the traditional *Doctrine of Double Effect (DDE)* and similar principles, differences in causal structure like this can be morally significant and may explain (at least in part) why we are inclined to treat **Bystander** and **Footbridge** as morally disanalogous.

While the vast literature exploring these ideas has been lively and thought provoking, chronicling dozens of further tales of trolley-induced peril, surprisingly little effort has gone into developing formal frameworks within which the relevant bevy of cases and the principles that purportedly distinguish them might be usefully modeled 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 general algorithmic implementability and applicability to decision making under uncertainty. Fortunately, for those deontologists inclined to see DDE-type reasoning as key to navigating trolleyology, the tools of causal modeling offer an overlooked resource for redressing these deficiencies.

By viewing moral decision problems in terms of *structural causal models*, we can precisify the logic behind DDE and related deontological constraints, rendering them more readily applicable to the task of solving moral decision problems, including ones involving empirical uncertainty. In what follows, my goal is to vindicate this suggestion by introducing a simple modeling framework within which moral decision problems are viewed as structural causal models supplemented by normative parameters (§2). Within this set up, we can capture not only consequentialist moral reasoning (§3) but also the kind of deontological thinking encoded in DDE (§4). Ultimately, we will be able to view various consequentialist and DDE-inspired moral theories as all applying the same general decision rule of *purified utility maximization* under disparate parameter settings. From this vantage point, it becomes evident that deontological moral theories of the relevant sorts, no less than consequentialist ones, can be naturally extended to accommodate empirical uncertainty on the part of moral agents in a fully Bayesian fashion (§5). Prior to closing, I stress that what I offer here is merely a modeling framework and that its moral verdicts are always relative to particular models, which can be more or less apt representations of real-life decision problems. Hence, there is need to articulate some general guidelines covering proper model construction in this setting (§6). Finally, I conclude by gesturing at various natural extensions of the modeling framework and avenues for further research (§7).[2]

[2]For our present purposes, which concern not the direct defense of DDE in any of its many formulations but simply the construction a suitable framework for representing the kind of reasoning it enjoins, I largely set aside many well-known objections to DDE, including, *inter alia*, (Thomson, 1985)'s complaint that DDE gets the wrong verdict in her well-known *Loop* case, (Bennett, 1995)'s contention that DDE is vitiated by the infamous *closeness problem*, and (Greene, 2013)'s sophisticated debunking arguments against DDE. These have all been discussed by proponents and opponents

2 Decision Problems and Causal Models

Philosophers keen to do so have proposed a number of intricate modeling frameworks within which to formally represent moral decision making.[3] Unfortunately for our purposes, the frameworks so far ventured have generally ignored the causal structure of decision problems.[4] If we hope to model the logic behind causally-loaded moral principles like DDE, we need to attend more explicitly to the causal features of moral decision problems in our representations of them. The model of moral decision problems that I will suggest as suitable for present purposes takes such problems as consisting of five components: (i) a set of possible *actions*, A, (ii) a set of *event-variables*, E, (iii) a family of *causal dependence functions*, $\mathcal{F}$, (iv) a *value function*, v, and (v) an *indignity measure*, δ.

The first three of these components are descriptive in nature and effectively comprise a special kind of *structural causal model* in which one variable is treated as a privileged act-variable.[5] I will refer to a triple consisting of just these three aspects of a moral decision problem as a *descriptive decision problem*. The value function and indignity measure, by contrast, are normative parameters of the model, specifying morally significant features of the decision problem regarding which competing moral theories may disagree. In this paper, I understand a *moral theory* to consist of two parts: (i) a way of specifying a decision problem's normative parameters, which we might think of as an *evaluation rule* mapping descriptive decision problems to full moral ones, and (ii) a *decision rule* that maps moral decision problems to their morally permissible option sets.[6] One upshot of the modeling framework we develop here will be that it is possible to view a variety of interesting consequentialist and deontological moral theo-

of DDE extensively elsewhere, and I have little directly to add to such debates beyond noting that, by articulating a general theoretical framework for organizing moral philosophers' disparate reflections on DDE, our efforts here may prove indirectly relevant to these and other controversies surrounding the doctrine.

[3] E.g., (Oddie & Milne, 1991), (Colyvan et al., 2010), (Dietrich & List, 2017).

[4] One near exception is (Halpern & Kleiman-Weiner, 2018), who attempt an explication of morally-loaded concepts like blameworthiness, intention, and responsibility via causal models of the same general sort employed here, without however sharing our specific aspiration of codifying deontological principles via novel moral decision procedures.

[5] For surveys on structural causal models, see (Pearl, 2009), (Halpern, 2016), and (Hitchcock, 2023). Our distinction between *act* and *event* variables, which is absent from standard structural causal models, brings our models part of the way toward the more complex framework of *influence diagrams*, richly explored by (Jensen & Nielsen, 2007).

[6] A moral theory's decision rule might also make finer discriminations than this by, for example, further classifying some permissible acts as better or worse than others. But our concern here will be with moral theories only insofar as they distinguish between the permissible and the impermissible.

ries as differing with respect to their evaluation rules while sharing a common decision rule. This section will review the descriptive components of a moral decision problem, leaving discussion of the normative components for §3-4.

Begin with *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. For ease of formal modeling, I identify an action with the proposition that it is chosen.[7] 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 modeling 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.

Next, turn to *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 **Bystander**, we may need to represent which path the trolley takes, whether the man with the trapped foot dies, etc. We can do so via *event-variables*, each of which corresponds to a possible feature of the world that may or may not obtain. Event-variables may be modeled, like A, as partitions of propositions. However, unlike A, I assume that event variables are binary in nature and contain simply a proposition and its negation as members.[8] So, where h is the proposition that the man stuck on the sidetrack is hit, one event variable in **Bystander** may be: $H = \{h, \bar{h}\}$.[9] 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

[7] I follow the lead of (Jeffrey, 1965/1983) in this regard.

[8] Restricting attention to finite, discrete models, this assumption will tremendously abet our efforts to formalize DDE-style reasoning without any terrible loss in generality, since a finite model featuring non-binary variables can often be rewritten as one involving only binary variables. Still, I can't claim confidently that the restriction is entirely without cost and so generalizing our framework to allow for models involving non-binary event variables is a worthy topic for future reflection; see §8.

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

(i.e., non-instrumentally) for and against each available action in the decision problem at hand. Again employing **Bystander** as an example, assuming that the lives of the transit workers and the man on the sidetrack are the relevant goods at stake in your decision, each of these persons' living/dying must be represented as possibilities via event-variables like H.[10] Throughout, I assume E to be finite. For convenience, I write V as a shorthand for $\{A\} \cup E$ and refer to the members of $\cup V$ as *atomic propositions* or, more simply, *atoms*. Moreover, I write $\mathcal{S}(V)$ for the set of *states generated by* V, that is, the set of conjunctions formable from atomic propositions. For convenience, I will write as if logically equivalent propositions are identical, assuming them to be completely intersubstitutable. An important class of states that deserves note is the set of *worlds* present in a model. A world is simply a maximally consistent conjunction of atoms, i.e., any state $z \in \mathcal{S}(V)$ such that, for every $x \in \cup V$, either $z \models x$ or $z \models \overline{x}$. The set of all worlds in a given decision problem is denoted W. Lastly, 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 and so corresponds to a distinct world.

Finally, we arrive at the distinctively causal component of our decision models: *causal dependence functions*. 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, in **Bystander**, whether the man on the sidetrack gets hit, H, causally depends upon your 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). We can encode information about how exactly one variable $X \in E$ causally depends upon others via a causal dependence function, $f_X : N_X \to X$, where $N_X \subseteq \mathcal{S}(V)$ is the set of maximally consistent conjunctions formable from the members of $\cup(V \backslash \{X\})$, i.e., atomic propositions that lie outside of X. That is, given any specification of how the variables other than X turn out, f_X indicates how X will turn out. A set, $\{f_X\}_{X \in E}$, of such causal dependence functions, we denote

[10] Note: there is no requirement here that event-variables or their members uniquely correspond to particular, finely-individuated reasons. What is required is simply that every good at stake be represented in the value of some variable. It is thus fine if a single variable value represents multiple goods (e.g., the survival of all five of the transit workers), provided the achievements of these goods suitably correlate in the problem.

as $\mathcal{F}$.

The family $\mathcal{F}$ allows us to define a relation, $\rightarrow$, of (direct) causal dependence amongst the members of V. Say that $X \rightarrow Y$ just in case f_Y yields non-constant output across some pair of inputs that differ only with respect to which member of X they entail. Informally, this means that Y causally depends upon X just in case it is ever possible that, holding fixed the true members of all other variables, which member of Y turns out true is contingent, according to f_Y, upon which member of X turns out true. To rule out cyclical causation and the like, we assume that $\mathcal{F}$ is structured so as to generate a relation of causal dependence that gives rise to a *directed acyclic graph* or, more simply, a *causal graph* in which the members of V serve as vertices, no two of which are connected in a loop. 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 $an(X)$.

With $\mathcal{F}$ in the background, variables can be partitioned into the *exogenous* and the *endogenous*. A variable X is exogenous just in case $par(X) = \emptyset$; otherwise, it is endogenous. So, exogenous variables are those that lack causal ancestors in the model, while endogenous variables are those that enjoy them. Note that, as seems appropriate in a model of free decision making, A is always treated as exogenous, since we left f_A undefined. When X is an exogenous event variable, f_X, though defined, is constant and can be interpreted as simply indicating the antecedently known and presently unalterable value of the variable X. With $\mathcal{F}$ in place, we have a specification of the values taken by all exogenous non-act variables in the model as well as a clear characterization of how precisely these variables, together with A, serve to fix the values of all causally downstream endogenous variables.[11]

Before moving on to normative matters, let's note how we might capture both **Bystander** and **Footbridge** within this framework. We might plausibly model the first of these problems via something like $\langle A, E, \mathcal{F} \rangle$, where:

- $A = \{a, \overline{a}\}$, a being the act of pulling the lever and $\overline{a}$ being the act of doing nothing.
- $E = \{H, S\}$, with $H = \{h, \overline{h}\}$ being the event-variable corresponding to

[11] We will generalize the present model to encompass cases where an agent may be subjectively uncertain of the causal structure of the problem she faces in §6.

whether or not the man on the sidetrack is hit and crushed by the trolley and $S = \{s, \overline{s}\}$ being the event-variable corresponding to whether or not the five transit workers are saved.

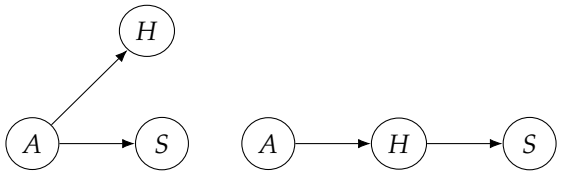
- $\mathcal{F} = \{f_H, f_S\}$, where $f_H(a) = h, f_H(\overline{a}) = \overline{h}$ and $f_S(a) = s, f_S(\overline{a}) = \overline{s}$. Here, abbreviation allows us to express f_H and f_S solely as functions of A, since neither H nor S causally depends upon the other.

Restricting ourselves to the same variables employed above, we might model **Footbridge** as a similar triple, $\langle A, E, \mathcal{F} \rangle$, such that:

- $A = \{a, \overline{a}\}$, a being the act of pulling the lever and $\overline{a}$ being the act of doing nothing.
- $E = \{H, S\}$, with $H = \{h, \overline{h}\}$ being the event-variable corresponding to whether or not the dropped man is hit and crushed by the trolley, and $S = \{s, \overline{s}\}$ being the event-variable corresponding to whether or not the five transit workers are saved.
- $F = \{f_H, f_S\}$, where $f_H(a) = h, f_H(\overline{a}) = \overline{h}$, and $f_S(h) = s, f_S(\overline{h}) = \overline{s}$. Here again, abbreviation spares us the need of specifying these functions more elaborately, since H only causally depends upon A and S only upon H.

The causal graphs determined by these causal models are depicted in Figures 1(a) and 1(b). Note that I have tried to model these cases in the simplest way possible while still bringing out their (potentially) morally relevant difference in causal structure: in **Bystander**, the harm to the man on the sidetrack is causally independent of the saving of the passengers; not so in **Footbridge**. There exists, of course, a vast array of equally accurate ways of modeling these decision problems, some of which may conceal this difference. For example, we could collapse H and S into a single event variable in both cases, rendering their causal graphs identical. Which of such equally accurate, yet more and less specific, models count as appropriate for morally evaluating an agent's options will depend upon the form the correct moral theory takes and what sort of features of a decision problem its evaluation and decision rules are sensitive to. In what follows, I will always treat moral principles as yielding (im)permissibility verdicts only relative to particular models of moral decision problems. The matter of determining whether a particular model constitutes an adequate representation of a real or hypothetical decision problem is one I take up in §6, where I will suggest simple heuristics for model construction

depending upon which moral theory one wishes to employ.



(a) **Bystander** (b) **Footbridge**

Figure 1

3 Consequentialism

Descriptive decision problems, characterized in terms of structural causal models, are free of any specification of the kind of normative features that guide both intuitive and principled moral evaluation of an agent's possible actions. For example, while we have assumed that the various goods and bads that tell for and against particular courses of action (e.g., the saving of five workers, the crushing of a man by a trolley, etc.) have been represented via the set of event variables, E, nothing in our models so far indicates the relative strength of competing goods and bads, nor even the direction in which any determination of event-variables bears upon the advisability of available actions. Further, while our causal models allow us to capture the idea that the crushing of the dropped man in **Footbridge** could serve as a causal instrument for the saving of the trolley's passengers in a way that the crushing of the man on the sidetrack in **Bystander** could not, nothing in our model indicates whether crushing a person with a trolley is the sort of thing that ever should be instrumentalized in such a way.

To capture these specifically moral features of decision problems, we need to introduce corresponding normative parameters into our models. The most obvious such parameter we will incorporate into our model of decision problems is a *value function*, v, intended to measure the strength of the reasons provided by the good and bad consequences wrought by available acts. We can take the objects on which v is defined to be all states formable via V, i.e., $v : \mathcal{S}(V) \rightarrow \mathbb{R}$. That is, v indicates the desirability of particular sets of variables taking on particular sets of values. I will make two strong assumptions regarding the structure and measurement scale of v. Firstly, I will assume that v is additive

in the sense that $v(x_1...x_n) = \sum_{i=1}^{n} v(x_i)$, for any conjunction $x_1...x_n \in \mathcal{S}(V)$. This entails that we can see the desirability of disjoint states as separately assessable and amalgamatable for purposes of determining the value of composite states, including worlds. Second, I will assume that v is measured to at least the following degree of uniqueness: any morally equivalent value function v' must involve no more than (i) multiplication of v by a positive constant and/or (ii) manipulation of v by addition of any set of constants to its atomic values (along with corresponding adjustment of the values of non-atomic states in line with the additivity requirement), provided that the values of logically incompatible atoms (i.e., atoms included in the same variable) are altered by the same constant.[12] Granting this, we are able to assume that v is normalized in such a way that for any event variable, $X = \{x, \overline{x}\} \in E$, $v(x) \geq v(\overline{x})$ entails that $v(\overline{x}) = 0$, i.e., the value of each event variable's worst member is set to zero.

These assumptions are needed for the statement of the DDE-inspired decision rule introduced in §4 and its application to decision under uncertainty in §5.[13] To the extent that it may not be reasonable to model every choice scenario via a descriptive decision problem coupled with such an additive value function, our modelling framework's domain of applicability is resultantly narrowed. That said, our assumptions are, 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 roughly separable and measurable in the requisite way. Whatever the merits of these technical assumptions, however, they exhaust our framework's axiological commitments. We may otherwise remain largely neutral in our understanding of moral value. For example, while some consequentialist moral theories will want to understand v as agent-neutral, we need not require this; the values encoded by v may be understood in either an agent-neutral or agent-

[12] Together, our requirements amount to assuming that v can be determined by summing the relevant members of a set of variable-indexed value functions, $\{v_X : X \to \mathbb{R}\}_{X \in V}$, each of which is unique up to choice of individual zero and common unit.

[13] Jettisoning our assumed normalization convention, though possible, would necessitate complicating the decision rule introduced below by stating it in terms of sums of suitably discounted value *differences* rather than simply sums of discounted values. Given the comparative simplicity of the latter framing, let us be content to normalize. My thanks to a reviewer for pressing me to clarify this.

relative way.[14],[15]

Consequentialists maintain that v is the only normative parameter we need include when modeling moral decision problems. According to this way of thinking, which actions are morally permissible in a given decision problem is entirely determined by considerations related to the value of the consequences the various available actions bring about. To state an appropriate decision rule for consequentialist moral theories, note that, given a generic descriptive decision problem, $\langle A, E, \mathcal{F} \rangle$, $\mathcal{F}$ uniquely determines which world each action in A will lead to. For any given $a \in A$, we'll call the world that a makes actual a's *resultant world*, denoted $w_a \in W$, and then further define the *utility* of an action $a \in A$, written $u(a)$, to be the value of a's resultant world, i.e., $u(a) := v(w_a) = \sum_{X \in V} v(X(a))$, where $X(a)$ is the true member of X under the supposition that a is performed.[16] Consequentialist moral theories, while disagreeing amongst themselves concerning the substance of v, are characterized by a commitment to the general decision rule of:

> **Utility Maximization:** Given a fixed moral decision problem, the morally permissible actions are those that maximize utility.

Once values are fixed, 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 overall value.

Its simplicity notwithstanding, many deontologists will complain that this is demanding counsel. Any available action that falls short of another with respect to utility is deemed morally impermissible by **Utility Maximization**, typically leaving the space of permissible actions within a moral decision prob-

[14] For example, if one thinks a parent may reasonably weight the good of their child significantly higher than another agent might weight the good of the same child, our framework allows that such agents may, even when confronted with descriptively similar decision scenarios, face distinct moral decision problems involving different value functions.

[15] Setting aside concerns about the structure and measurement scale of v, there is one aspect of modeling relative value of worlds by means of any single value function that may be objectionable to some. In particular, doing so 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 modeled 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 theories 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 so here only to avoid paying the notational cost.

[16] Note that $w_a = \&_{X \in V} X(a)$.

lem extremely narrow. The rule thus allows no room for supererogation, for that wide variation in moral worth that common sense recognizes within the category of the morally permissible, from the just barely decent to the valiantly heroic. The cases in which this feature of **Utility Maximization** seems most problematic typically involve individual agents faced with decisions in which morally optimal action carries with it non-trivial cost to the acting agent (for example, problems in which you can sacrifice your own life in order to save multiple others). It is in such cases that we feel most inclined to draw the distinction between the optimal and the permissible that **Utility Maximization** refuses to make. However, many of the examples that motivate DDE and related principles are not of this sort. For example, neither **Bystander** nor **Footbridge** involves any cost to the acting agent, aside from the minimal expenditure of effort required to move one's arm and pull a lever. In such problems, the optimizing character of **Utility Maximization** looks much less objectionable. If one act is really morally superior to another, and neither is especially costly to you, why shouldn't morality demand that you not take the inferior act?

Moreover, the contexts in which explicit application of formal moral decision procedures like the ones explored here appears most appropriate are also contexts in which typical deontological concerns about the moral demandingness of consequentialism seem inapt (e.g., within fields like public policy and machine ethics). Hence, we should not allow concerns about the optimizing character of consequentialist moral theories in general to obscure that there may nonetheless be a large domain of interesting moral decision problems within which a maximizing decision rule is perfectly proper. Given this, I am content to allow the DDE-inspired decision rule introduced below to appropriate the maximizing character of **Utility Maximization**, simply noting the consequent limitation in our modeling framework's domain of applicability. While I believe we could restate this rule so as to make room for supererogation and thus generalize its applicability, we would gain little in doing so vis-à-vis our present purposes since the critiques of consequentialism that animate the DDE literature are largely orthogonal to moral demandingness objections to maximizing consequentialism, and we would pay a non-negligible cost in terms of ease of exposition. Thus, in what follows, we will blithely follow the lead of consequentialists in collapsing the morally permissible into the morally

optimal.[17]

Before we proceed to consider our modeling framework's final normative parameter and the above mentioned alternative to **Utility Maximization**, it is worth briefly highlighting how this rule can be employed to yield straightforward consequentialist verdicts in **Bystander** and **Footbridge**, once we supplement our earlier descriptive models of these scenarios with intuitive value functions. If we suppose that in both cases values are given by the number of lives saved, we arrive at two formally identical value functions each specifying that $v(a) = v(\overline{a}) = 0, v(\overline{h}) = 1$, and $v(s) = 5$. These in place, we can contrast the relative utility of pulling the lever (a) and not pulling it ($\overline{a}$):

$$u(a) = v(ahs) = 5 > 1 = v(\overline{a}\overline{h}\overline{s}) = u(\overline{a})$$

Redirecting the trolley to hit the man on the sidetrack in **Bystander** and dropping the man in **Footbridge** each equally result in a better world than the alternative actions of not pulling the levers, so consequentialists would have us pull the lever for the greater good in both cases. Many, as noted above, find this application of consequentialist reasoning seriously problematic, since it fails to distinguish between the intuitively permissible case of pulling the lever in **Bystander** and the intuitively impermissible case of pulling it in **Footbridge**. Moral philosophers of this bent must look to non-consequentialist theories.[18]

[17] That said, those concerned to extend the applicability of the decision rule presented below to problems in which the acting agent's own good is at stake in ways that cast doubt upon the suitability of maximizing decision procedures can do no better than turn to (Lazar, 2017), whose deontological decision theory offers an extremely illuminating template for how such an extension might go.

[18] Some authors, including (Dougherty, 2013) and (Setiya, 2018), have raised the possibility that the value of worlds may be affected by how the good and bad events they imply are causally ordered. For example, a world in which you kill one as a means of saving five may be worse than a world in which you kill one and save five but the two events are not causally dependent upon one another, which may open the door to a reconciliation of common intuitions about trolley cases with a fully consequentialist moral logic (albeit one plausibly involving non-additive value representations). Approaches of this sort, situated in the context of the large literature on *consequentializing* non-consequentialist moral theories, are, though interesting in their own right, quite removed from the spirit and aspirations of this paper, which aims to introduce a modeling framework that naturally accommodates double effect reasoning and related forms of deontological thinking without necessarily challenging typical judgments about the value of worlds (according to which, for example, the worlds wrought by pulling the levers in Footbridge and Bystander are roughly equivalent value-wise). I am grateful to a referee for pressing me to clarify the connection between this essay's aims and the consequentializing project.

4 Double Effect

According to proponents of the *Doctrine of Double Effect (DDE)*, consequentialist moral theories go awry (in part) by neglecting the morally relevant distinction between intending harm (either as an end or as a means) and merely foreseeably causing it.[19] This line of thinking suggests an explanation as to why so many of us recoil at the idea of pulling the lever in **Footbridge**, though not in **Bystander**. In both examples, pulling the lever foreseeably harms an innocent person, but only in **Footbridge** does this harm plausibly qualify as intended, since only in this case does the victim's getting hit by a trolley constitute a necessary causal condition for the achievement of the sought after goods (i.e., the saving of the five transit workers).

Another way to put this point is that in **Footbridge**, the rule of **Utility Maximization** invites us to treat an innocent person's being fatally hit by a trolley as an instrumental reason for action. By viewing the goods that follow from this horrific event as straightforward reasons to bring it about, 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 evils. 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 adherent of **Utility Maximization**, could not be charged with intending any evil or instrumentalizing it for the sake of good since no consequences of such evils would ever enter into their practical deliberations concerning what to do in a motivating fashion.

In order to precisfy such a decision rule and formally capture moral theories of the DDE kind within our framework, we need to introduce a further normative parameter into our models, one indicating whether and to what extent a given event may be permissibly instrumentalized for the sake of bringing about another. For this purpose, I suggest that we supplement our initial normative parameter, v, with what I will call an *indignity measure*, denoted

[19] The classic statement of DDE within Catholic moral theology was given by (Gury, 1874). For some contemporary defenses of particular versions of DDE and relevantly similar doctrines like the *Means Principle* and the *Pauline Principle*, see, e.g., (Donagan, 1977), (Boyle, 1980), (Nagel, 1986), (Quinn, 1989b), (FitzPatrick, 2003), (Cavanaugh, 2006), (Mikhail, 2011), (Wedgwood, 2011), (Pruss, 2013), (Nelkin & Rickless, 2014), (Tadros, 2015), (Alexander, 2016), (Bronner & Goldstein, 2018), (Masek, 2018), and (Stuchlik, 2022).

$\delta : \mathcal{S}(V) \times \cup V \rightarrow [0, 1]$. An indignity measure takes as input a state-atom pair and is intended to return, intuitively, a measure of the extent to which any gains accruing from the truth of the given atom must be discounted, for purposes of moral deliberation, if they were realized only by means of the given state. For example, recalling our models of **Bystander** and **Footbridge**, $\delta(h, s) = 0$ indicates that the good of saving five transit workers (s) must be entirely discounted, for purposes of moral deliberation, if brought about only by means of someone else being hit and crushed by a trolley (h), while $\delta(h, s) = 1$ aligns with the consequentialist judgment that no such discounting is called for. Intermediate values correspond to partial discounting of tainted rewards and thus make room for non-absolutist variants of DDE according to which bringing about harm intentionally is, though more difficult to justify than merely forseeably causing it, not categorically forbidden.[20]

With v and δ given as normative parameters, the problem now becomes formulating a general decision rule that captures the logic of DDE. To aid us in this project, we will first introduce the notion of an atomic state $X(a)$'s *purified value in world* w_a: $v_p(X(a), w_a)$. Once such purified values are in hand, we will be able to define the *purified utility* of an act a, written $u_p(a)$, as the sum of the purified values of the atomic states a's performance results in, analogous to how the utility of an act was identified above with the sum of the values of the atomic states its performance results in. Informally, the purified value of atomic state $X(a)$ in world w_a is the state's value, discounted by some measure of the extent to which it was achieved by morally illegitimate means in w_a. Cashing this out precisely, using v and δ, requires wading through a number of subtleties, however.

Intuitively, we want $v_p(X(a), w_a)$ to equal something like $\delta(z, X(a))[v(X(a))]$, where z is some state that constitutes a complete causal means to $X(a)$ in world w_a. For example, in our model of **Bystander**, pulling the lever, a, constitutes such a complete causal means to the end of saving the five transit workers, s, and hence the purified value of this latter state can be viewed as $v_p(s, w_a) = \delta(a, s)v(s)$. If $\delta(a, s)$ equals one here (i.e., if pulling a train switch is a perfectly legitimate

[20] Absolutist proponents of DDE, including (Anscombe, 1961) and (Boyle, 1980), maintain that there are certain evils so grave that one ought *never* intend them, no matter how great the prospective benefits of doing so. In our terms, they hold that for some state x (perhaps one involving the death of an innocent person), $\delta(x, y) = 0$ even for atoms y of arbitrarily high value. On the other hand, Quinn (1989b) presents DDE rather modestly as the principle that "the pursuit of a good tends to be less acceptable where a resulting harm is intended as a means than where it is merely foreseen" (p. 335), which seems to invite consideration of less extreme indignity measures.

thing to instrumentalize for the sake of saving five lives), we secure the verdict that the purified value of saving the five in **Bystander** simply coincides with this event's unpurified value. However, in our model of **Footbridge**, the complete causal means by which the transit workers are saved, if you pull the lever, includes not only your pulling of the lever but also the dropped man's being hit, *ah*, and thus the purified value of saving the workers becomes $v_p(s, w_a) = \delta(ah, s)v(s)$. If $\delta(ah, s)$ equals zero here (i.e., if an event that encompasses dropping a man in front of a lethally fast moving trolley is the sort of thing that ought not be in any way willfully instrumentalized even to save five lives), then the purified value of saving the five lives in **Footbridge** will be reckoned as zero as well.

While hopefully somewhat intuitive, the notion of a complete causal means is as yet undefined. Our immediate task is thus to spell out precisely the conditions under which one state constitutes a complete causal means relative to another in a way that vindicates our choice of a as the first argument of $\delta(\cdot, s)$ when computing $v_p(s, w_a)$ in **Bystander** and *ah* as our choice when doing so in **Footbridge**. As a preliminary step, define a descriptive decision problem, $\langle A, E, \mathcal{F} \rangle$'s, *choice independent state*, denoted $c_A \in \mathcal{S}(V)$, as the logically strongest state entailed by w_a for every $a \in A$. That is, a choice independent state specifies the value of all variables that cannot be influenced by an agent's actions. Letting a descriptive decision problem, $\langle A, E, \mathcal{F} \rangle$, and a variable $X \in V$ be given, we can now propose four conditions that define whether $z \in \mathcal{S}(V)$ constitutes a *causal means to* $X(a)$ *in world* w_a:

1. **Actuality**: $w_a \models z$.
2. **Control**: For every $z' \in \mathcal{S}(V)$, if $c_A \models z'$, then $z \not\models z'$.
3. **Sufficiency**: $f_X(z') = X(a)$, for every $z' \in \mathcal{S}(V)$ such that $z' \models z$ and $z' \models c_A$.
4. **Relevance**: For every $Y \in V$ and $y \in Y$, if $z \models y$, then $Y \in an(X)$.

The motivation behind these requirements is straightforward. **Actuality** simply requires that for an event to be a causal means it must actually obtain. **Control** rules out treating as a causal means, in the sense that concerns us, any event that is determined to obtain, regardless of which act the agent chooses to perform. Such events cannot be ones that an agent might problematically will to bring about, since they lie outside the scope of her control altogether and hence don't

seem to supply good grounds for any downstream discounting.[21] **Sufficiency** guarantees that a causal means is, together with any choice independent facts, causally sufficient to secure its end. Finally, **Relevance** is intended to forestall cluttering a causal means to $X(a)$ with specifications of variable values that have no bearing upon the obtaining of $X(a)$. A causal means to an event must not entail anything about what states obtain in variables that don't lie causally upstream of the event. It is this assumption that disqualifies *ah* from counting as a causal means to *s* in our model of **Bystander**.

To extend these conditions into a definition of a complete causal means, we say that state z is a *complete causal means to* $X(a)$ *in world* w_a just in case it is a causal means toward $X(a)$ in w_a that additionally satisfies:

5. **Completeness**: For every $x \in \cup E$, if $z \models x$, then $z \models z'$ for some state $z' \in \mathcal{S}(V)$ such that z' is a causal means toward x in w_a.

A causal means to $X(a)$ that satisfies **Completeness** specifies a full story about how, assuming c_A in the background, the agent's choice of a causally guarantees the truth of $X(a)$. This is an important condition because when we discount goods for purposes of computing purified values, we want to discount according to the full causal means by which a leads to $X(a)$ and not just by a proper, and perhaps less morally problematic, part of such a means. If, for example, we retell the story behind **Footbridge** so that the dropped man only needs to crash (at lethal speed) into a button in order to trigger the trolley's brakes, we would not want to discount the goods so achieved merely by the extent to which pushing a button or triggering brakes are legitimate means to employ in saving lives, but by the extent to which the whole causal apparatus, including dropping a man to his foreseeable death, qualifies as such a means.

Note that, in our models of **Bystander** and **Footbridge**, *a* and *ah*, respectively, uniquely qualify as complete causal means to *s*. However, letting $S(X(a), w_a) \subseteq \mathcal{S}(V)$ be the set of complete causal means for given atomic state $X(a)$ in world w_a, it will not always the case that $S(X(a), w_a)$ is a singleton. In some decision problems, there may be several independent, complete causal means that are each individually sufficient for the attainment of $X(a)$. When this occurs, we face

[21] Admittedly, this might be questioned in some contexts. For example, if a victim is murdered in order that his organs may be subsequently stolen for use in otherwise laudable transplants, some may view not only the act of killing the victim as morally wrong but also, and independently, the subsequent act of using the illicitly obtained organs. Those that think we ought to discount goods obtained by means of evils, even where such evils are already causally fixed, may relax the **Control** assumption as they see fit.

a problem vis-à-vis selecting a candidate causal means to employ in discounting $X(a)$ for purposes of computing its purified value. As an illustration, consider:

> **Boxes:** A runaway trolley barrels down a track toward five unaware transit workers. However, you have time to pull a nearby lever that will open a trap door on a footbridge passing over the track. Standing on the trap door is a man of sufficient mass to bring the trolley to a halt following impact, averting disaster for the transit workers. Unhappily, the force of the trolley's impact on the dropped man's body would kill him. Next to the man on the footbridge, and also above the trap door, are several boxes of sufficient mass to bring the trolley to a complete halt on their own as well. Pulling the lever means life for the transit workers but death for the dropped man.[22]

We could naturally model this scenario as one of our moral decision problems, $\langle A, E, \mathcal{F}, v, \delta \rangle$, where:

- $A = \{a, \overline{a}\}$, a being the act of pulling the lever and $\overline{a}$ being the act of doing nothing.
- $E = \{H, B, S\}$, with $H = \{h, \overline{h}\}$ being the event-variable corresponding to whether or not the man on the footbridge is hit by the trolley and killed, $B = \{b, \overline{b}\}$ being the event-variable corresponding to whether or not the boxes fall into the trolley's path, and $S = \{s, \overline{s}\}$ being the event-variable corresponding to whether or not the five transit workers are saved.
- $\mathcal{F} = \{f_H, f_B, f_S\}$, where $f_H(a) = h$, $f_H(\overline{a}) = \overline{h}$, $f_B(a) = b$, $f_B(\overline{a}) = \overline{b}$, $f_S(h, b) = s$, $f_S(\overline{h}, b) = s$, $f_S(h, \overline{b}) = s$, $f_S(\overline{h}, \overline{b}) = \overline{s}$. Here, our abbreviation assumes f_H and f_B to be solely functions of A and f_S to be a function of H and B.
- Values are given by lives saved, resulting in $v(s) = 5$, $v(\overline{h}) = 1$, and $v(x) = 0$ for all other atoms $x \in \cup V$.
- We treat the dropped person and the five transit workers' deaths as indignities, that is, $\delta(z, x) = 0$ for all atoms x if $z \models h$ or $z \models \overline{s}$, otherwise $\delta(z, x) = 1$.

The causal graph corresponding to this model of **Boxes** is depicted in Figure 2. Note that in the world where you pull the lever, dropping both the man and the boxes from the footbridge, there are several complete causal means present

[22] This sort of case was suggested to me by Ryan Doody.

for s, the saving of the transit workers. In particular, all three of ah, ab, and ahb count as complete causal means of s in w_a. Dropping either the man, the boxes, or both would be sufficient to bring about the end of saving the transit workers.

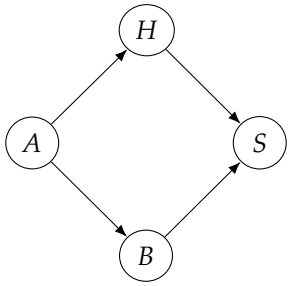


Figure 2: **Boxes**

Contrary to how matters stand in **Footbridge**, by pulling the lever so as to save the five workers' lives in **Boxes**, one need in no way will that the person dropped from the footbridge be hit by the trolley, given that the boxes are causally sufficient in themselves to achieve the end of stopping the trolley. Indeed, in this case, the dropped person's being hit provides no greater motivation for pulling the lever than the person's being hit did in **Bystander**; in both cases, the harms can be construed as incidental and outside of the acting agent's intentions. In cases like this then, where multiple causal factors prove independently sufficient for an atom $X(a)$, it seems most in keeping with the broad spirit of DDE that one should compute the purified value of $X(a)$ relative to a complete causal means z that maximizes $\delta(z, X(a))$. In **Boxes**, this means that one should discount the value of saving the transit workers' lives, $v(s)$, for purposes of assessing the purified utility of pulling the lever, a, only by $\delta(ab, s) = 1$ (i.e., not at all) rather than by $\delta(ah, s) = \delta(ahb, s) = 0$.

This motivates a general definition of an atomic state $X(a)$'s purified value in world w_a as: $v_p(X(a), w_a) = max_{z \in S(X(a), w_a)}\{\delta(z, X(a))\}[v(X(a))]$, if $S(X(a), w_a) \neq \emptyset$, while $v_p(X(a), w_a) = v(X(a))$ otherwise. An act a's purified utility then becomes: $u_p(a) = \sum_{X \in V} v_p(X(a), w_a)$. With purified utilities thus defined, we can finally state a decision rule capable of accommodating DDE-type moral theories:

> **Purified Utility Maximization:** Given a fixed moral decision problem, the morally permissible actions are those that maximize purified utility.

It should be clear that this rule, while agreeing with **Utility Maximization** in **Bystander**, disagrees with it in **Footbridge** (assuming that we borrow a specification of δ from our above characterization of **Boxes**). Consider **Bystander** first. To assess the permissibility of pulling the lever (a) and of abstaining ($\overline{a}$), we compare these two acts' purified utility values. Note, however, that in **Bystander** purified values coincide with standard ones, since none of the decision problem's variables are causally posterior to its potentially tainted ones (i.e., H and S). Hence, in this case, maximization of purified utility appears identical to maximization of utility *simpliciter*, resulting in the verdict that a (pulling the lever) is uniquely morally permissible:

$$u_p(a) = \sum_{X \in V} v_p(X(a)) = 5 > 1 = \sum_{X \in V} v_p(X(\overline{a})) = u_p(\overline{a})$$

However, application of **Purified Utility Maximization** to **Footbridge** illustrates the rule's potential to depart from **Utility Maximization**. The relevant purified values used to assess purified utility in this problem no longer simply coincide with standard values, since ah is the only complete causal means to s in w_a but $\delta(ah, s) = 0 \neq 1$. Hence, we can obtain the result that *not* pulling the lever ($\overline{a}$) is uniquely moral:

$$u_p(a) = \sum_{X \in V} v_p(X(a)) = 0 < 1 = \sum_{X \in V} v_p(X(\overline{a})) = u_p(\overline{a})$$

Note that we can think of **Purified Utility Maximization** as a generalization of **Utility Maximization**, since it is possible to recover all the verdicts of the latter by employing the former, provided it is always the case that $\delta(z, x) = 1$, i.e., the indignity measure is a constant function that always yields one. It is thus possible to understand consequentialist moral theories as employing the same general decision rule as DDE-type theories (namely, **Purified Utility Maximization**), while differing with DDE-type theories over whether any events ever truly count as indignities in the technical sense employed here, i.e., as the sort of thing one has non-consequentialist reasons not to intentionally bring about or instrumentalize for further ends.

We have designed our modeling framework so as to be flexible enough to accommodate a wide array of theories about the nature of δ, including those of both an absolutist and non-absolutist variety.[23] In any given class of moral

[23] Attentive readers may raise the following objection to modeling the attitudes of firmly absolutist proponents of double effect reasoning via purified utility maximization. If two actions will lead to the same indignity and the same consequences independent of the indignity, but differ in that the first will bring more good out of the indignity than the latter will, a purified utility maximizer with an absolutist indignity measure will be indifferent between these acts since they

decision problems, actually determining δ, like fixing v, requires substantive moral theorizing and competing DDE-type theories will go about the task in varying ways. Postulating any general principles governing the estimation of δ thus lies beyond the scope of our modest and ecumenical modeling efforts here. Nonetheless, we might at least mention one assumption that might prove convenient to impose upon the structure of δ in many cases and that substantially parallels the assumption of the additivity of v, namely, *multiplicativity*: $\delta(z, x) = \prod_{\{y \in \cup V : z \models y\}} \delta(y, x)$. Assuming that δ is multiplicative ensures that reasons against instrumentalizing distinct states compound in natural ways. If, for example, one atomic state z is morally problematic with respect to bringing about an atom x to the extent that it calls for a discounting of $v(x)$ by, say, $\delta(z, x) = 0.5$, while another logically distinct atomic state z' is problematic to the same extent, i.e., $\delta(z', x) = 0.5$, then the conjunction zz', if instrumentalized for the sake x, requires that $v(x)$ be discounted via multiplication by $\delta(zz', x) = 0.5 \times 0.5 = 0.25$.[24] While this assumption is often natural enough to employ, however, we should emphasize that our modeling framework and its attendant rule of **Purified Utility Maximization** have no essential need for the requirement that δ be multiplicative in general.

5 Decision Under Uncertainty

Most philosophical discussion of both trolley problems and deontological principles like DDE has operated under a policy of feigned certainty regarding the causal structure of contemplated decision problems.[25] We assume, for example,

will completely discount the different goods that fall downstream of the indignity which these otherwise equivalent acts cause. While especially stern absolutists may not find this result troubling (perhaps any assignment of practical weight to an indignity's positive effects comes too close to instrumentalizing it), others likely will. For those of this persuasion, a relatively simple fix is available in the form of expanding the codomain of our indignity measures to include infinitessimal values à la (Robinson, 1996). If $\delta(z, x) = \epsilon$, where $r > \epsilon > 0$ for all positive reals r, then, assuming that z involves some non-infinitessimal cost, a purified utility maximizer will still be utterly unwilling to bring about z for the sake of x (so long as x is not of infinite value), in keeping with the spirit of absolutism. However, such a purified utility maximizer will, unlike the sterner absolutist who sets δ to zero, be willing nonetheless to optimize the consequences of z on the assumption that its actualization can be independently justified on grounds that do not involve instrumentalizing it. I am grateful to an attentive reviewer for pressing this worry.

[24] Note that, for the consequentialist, invariably mapping δ to the multiplicative identity renders it trivially multiplicative. On the other hand, for an absolutist who mandates that $\delta(y, x) = 0$ for some atom y and every atom x, multiplicativity ensures also that $\delta(z, x) = 0$ for every state z that entails y. The δ proposed in the above model of **Boxes** is multiplicative to this effect.

[25] This feigned certainty has been a source of scorn heaped upon trolleyologists by, *inter alia*, (Fried, 2012), (Hansson, 2012), (Goodall, 2016), (Nyholm & Smids, 2016), (Nyholm, 2023).

that 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 the modeling framework developed here as an approach to analyzing DDE-style reasoning is the ease with which it allows us to seamlessly integrate a major strand of deontological ethics with a fully Bayesian approach to risk management.[26]

For concreteness, consider a case like the following:

> **Uncertain Means:** A runaway trolley barrels down a track toward five unaware transit workers. You have time to pull a nearby lever which you know will do exactly one of two things: (i) switch the trolley onto a sidetrack, averting disaster for these transit workers, but killing an oblivious third party who has gotten his foot stuck attempting to cross the sidetrack, or (ii) open a trap door on a footbridge passing over the track, dropping onto it a man of sufficient mass to bring the trolley to a halt following impact, averting disaster for the transit workers but killing the dropped man. You are equally confident in each of these possibilities.

Uncertain Means is clearly a mixture of **Bystander** and **Footbridge**. 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 $\mathcal{F}$ takes on, so you have no way of representing **Uncertain Means** 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 a number of models of uncertainty one might employ in

[26] A common complaint against deontological ethics in recent years has been that deontologists lack a reasonable account of how their principles ought to guide decision making under uncertainty. (Jackson & Smith, 2015) have dubbed this deontology's *implementation problem*. A fascinating literature has grown up in response to this problem; see, *inter alia*, (Aboodi et al., 2008), (Lazar, 2017, 2020), (Bjorndahl et al., 2017), (Tarsney, 2018), and (Tomlin, 2019). However, the strand of deontology at issue here (i.e., DDE and related doctrines like the Means Principle) has been surprisingly ignored, with (Alexander, 2016) and (Nelkin & Rickless, 2023) notable exceptions. The former of these authors, while not developing a full modeling framework or precise decision rule, comes the closest to the concerns of this essay and discerningly suggests the possibility of integrating DDE-style reasoning (or, in his case, the Means Principle) with expected value theory along lines similar in spirit to those taken here. (See pp. 262-3.)

doing so, but by far the most familiar is orthodox probability theory, which justifies its presumptive adoption.

Heretofore, we have assumed that every decision problem specifies a unique family of causal dependence functions, $\mathcal{F}$. But let us now dispense with this assumption and model *descriptive decision problems under uncertainty* as quadruples of the form $\langle A, E, \{\mathcal{F}_j\}_{j\in J}, P\rangle$, where A and E are interpreted as before, while $\{\mathcal{F}_j\}_{j\in J}$ is a finite collection of families of causal dependence functions defined with respect to V and indexed by a set J, and $P : \{\mathcal{F}_j\}_{j\in J} \rightarrow [0,1]$ is a probability mass function satisfying $\sum_{j\in J} P(\mathcal{F}_j) = 1$. The interpretation is that $\{\mathcal{F}_j\}_{j\in J}$ is the set of all causal structures on V deemed possible by the agent facing the decision problem at hand, and P measures the agent's degree of belief (or level of confidence) that each member of this set corresponds to the actual (but unknown) causal structure of the problem she faces.[27]

With our decision models generalized in this way, we can capture **Uncertain Means** as a *moral decision problem under uncertainty*, $\langle A, E, \{\mathcal{F}_j\}_{j\in J}, P, v, \delta\rangle$, where:

- $A = \{a, \overline{a}\}$, a being the act of pulling the lever and $\overline{a}$ being the act of doing nothing.
- $E = \{H_1, H_2, S\}$, with $H_1 = \{h_1, \overline{h_1}\}$ being the event-variable corresponding to whether or not the man on the sidetrack is fatally hit by the trolley, $H_2 = \{h_2, \overline{h_2}\}$ being the event-variable corresponding to whether or not the potentially dropped man is fatally hit by the trolley, and $S = \{s, \overline{s}\}$ being the event-variable corresponding to whether or not the five transit workers are saved.
- $\{\mathcal{F}_j\}_{j\in J} = \{\mathcal{F}_1, \mathcal{F}_2\}$, where $\mathcal{F}_1$ corresponds to the causal structure at play in **Bystander** and $\mathcal{F}_2$ corresponds to the causal structure at play in **Footbridge** (each extended in the natural way to take account of all present variables).
- $P(\mathcal{F}_1) = P(\mathcal{F}_2) = 0.5$.
- Values may be identified with lives saved: $v(s) = 5, v(\overline{h_1}) = 1$, $v(\overline{h_2}) = 1$, and $v(x) = 0$ for all other atoms $x \in \cup V$.[28]

[27] Note that we can still accommodate our previous examples of decision under certainty in this new framework by allowing that J may be a singleton.

[28] I assume here that the mere presence of epistemic uncertainty on the part of the deliberating agent does not alter the value of the goods achieved by pulling or not pulling the lever, whatever these actions' actual causal fruits turn out to be. If, however, one doubts this claim and judges,

- We treat deaths as indignities, that is, $\delta(z, x) = 0$ for all atoms x if $z \models h_1$, $z \models h_2$, or $z \models \bar{s}$, otherwise $\delta(z, x) = 1$.

Strictly speaking, **Purified Utility Maximization**, whatever its merits, falls silent on **Uncertain Means**, which lacks the right form for this rule to yield any direct advice. However, given that **Purified Utility Maximization** simply directs agents to optimize their behavior relative to a suitably specified numerical measure of value, conditional upon knowledge of the right causal hypotheses, we can naturally generalize this rule to encompass decisions under uncertainty in a manner entirely analogous to how consequentialists typically generalize **Utility Maximization**. To do so, define $u_p(a; \mathcal{F}_j)$ to be the purified utility of action a under causal hypothesis $\mathcal{F}_j$. We can then state:

> **Expected Purified Utility Maximization:** Given a fixed decision problem involving uncertainty, the morally permissible actions are those that maximize $E_P[u_p(a)] = \sum_j P(\mathcal{F}_j) u_p(a; \mathcal{F}_j)$.

According to **Expected Purified Utility Maximization**, an agent faced with a problem of decision under uncertainty ought to value her options according to the expectation of their possible purified utilities, where the expectation is taken relative to the probability measure capturing the agent's uncertainty regarding causal hypotheses. Applied to **Uncertain Means**, **Expected Purified Utility Maximization** yields the result that you ought to pull the lever since:

$$\begin{aligned} E_P[u_p(a)] &= P(\mathcal{F}_1) u_p(a; \mathcal{F}_1) + P(\mathcal{F}_2) u_p(a; \mathcal{F}_2) \\ &= 0.5 \times 6 + 0.5 \times 1 \\ &= 3.5 \end{aligned}$$

$$\begin{aligned} E_P[u_p(\bar{a})] &= P(\mathcal{F}_1) u_p(\bar{a}; \mathcal{F}_1) + P(\mathcal{F}_2) u_p(\bar{a}; \mathcal{F}_2) \\ &= 0.5 \times 2 + 0.5 \times 2 \\ &= 2 \end{aligned}$$

Hence, $E_P[u_p(a)] > E_P[u_p(\bar{a})]$.

---

say, the value of the five lives on the main track to be greater or lesser here than in **Bystander** and **Footbridge** one may adjust the value measure accordingly.

This application of **Expected Purified Utility Maximization** stays true to the motivating spirit of DDE by not instrumentalizing evil for the sake of good. While an agent who opts to pull the lever in **Uncertain Means** risks dropping the man on the footbridge onto the tracks, neither this possible indignity nor any of its potential good consequences is treated by the decision rule as any reason to pull the lever, in light of the specified indignity measure. In fact, a gamble in which the lever would either function as in **Bystander** or else simply fail to operate would be strictly preferred, according to **Expected Purified Utility Maximization**, to the gamble it actually recommends in **Uncertain Means**, revealing that its recommendation, unlike that of an expectational generalization of **Utility Maximization**, is in no way driven by the prospect of attaining any ill-gotten goods.

Though our focus here has been on modeling the moral decision making of *empirically* uncertain agents, it is worth noting that the framework developed here may also prove useful for modeling further forms of uncertainty relevant to moral agency. In particular, recent philosophers have grown increasingly interested in the phenomenon of *normative uncertainty*, or uncertainty about moral matters, which may not reduce to any form of empirical uncertainty. Allowing for genuinely normative uncertainty, we can envision agents who fail to be dogmatic adherents of particular moral principles, like utilitarianism or the doctrine of double effect, but rather exhibit doubt regarding which moral theory is correct. Prescribing decision rules for such agents has proved challenging, but the kind of ecumenical modeling framework developed here suggests a potentially helpful way forward, since we can allow for uncertainty regarding a decision problem's normative parameters in much the same manner as we have allowed for uncertainty regarding its descriptive features. If, for example, an agent spreads her credences over multiple possible indignity measures, we can compute the purified utility of her acts with respect to each such measure and then compute overall expected purified utilities by taking credence-weighted averages of these possible purified utility values. Though handling normative uncertainty in this way no doubt demands some degree of caution, the potential merits of the approach ought to be further explored, especially in light of the difficulty of the problem posed by normative uncertainty.[29]

[29] For an overview of some philosophical problems relating to moral uncertinaty, see (MacAskill et al., 2020). Many thanks to an anonymous reviewer for calling my attention to the potential relevance of the framework developed here for modeling uncertainty of this sort.

6 Model Relativity

Given a formal representation of a moral decision problem of the specified sort, **(Expected) Purified Utility Maximization** straightforwardly 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 modeler 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 specification of both descriptive and normative parameters. A representation of **Bystander** 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 modeling enterprise or to decision rule of **Purified Utility Maximization**. 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 modeler rather than with the modeling 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 hinted at this above when we noted that **Footbridge** and **Bystander** could

each be accurately, though unsatisfactorily, modeled by a simple two variable graph that collapsed H and S into a single variable, in which case **Purified Utility Maximization** would fail to pull apart these cases, regardless of how δ is characterized. It evidently matters then which representation of a given problem we use as input when asking our theory to separate out permissible from 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 intuitively reliable application of **Purified Utility Maximization**, but doing so is clearly well beyond our ken. Fortunately though, we needn't produce such elaborate models to be reasonably confident that a representation is adequate enough for the application of a given rule. Coarse-grained 'small world' models will do fine provided they represent just enough of the problem at hand. In the context of capturing DDE-inspired moral theories, I think we can helpfully boil down the 'just enough' to two basic rules or heuristics of model construction, one corresponding to each of our modeling framework's two normative parameters.

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 prospective good or bad that tells for or against each of the available actions. Or, at least, it needs to represent every significant such good or bad, i.e., every value that might have a real chance of impacting the relative standing of the agent's options. There is nothing special about DDE-type theories in this regard. Consequentialist ones have equal need of the same rule. For neither approach, 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.

This rule is the only one consequentialist moral theories have need of. Once all significant goods and bads are accounted for in the model, the consequen-

tialist runs no further risk of having underspecified her decision situation for purposes of applying her decision rule. Not so for the proponents of a deontological principle like DDE. Their theories are 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 various ways in which a decision model might leave out causal information of significance to the application of these approaches. Most significantly, a model might fail to depict the causal dependence of a good event upon the realization of a causally prior indignity in the right way. For example, if we modeled **Footbridge** in a manner analogous to **Bystander**, 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 an indignity, and hence, applying **Purified Utility Maximization** 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 pulling the lever. The second general rule of model construction thus directs: if a significant good could potentially be caused by an indignity in the actual decision problem, then this must be reflected in the constructed model.

It should be obvious that, useful as I take them to be, these rules offer no mechanical recipe for constructing decision models. They are intended as nothing more than heuristics for the careful modeler to make use of as she goes about her art. For any real world decision problem that a human (or artificial) agent may actually face, there will be an indefinite number of models that, from the perspective of **Purified Utility Maximization**, 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 modeler has done her job right (i.e., represented all significant goods and all causal dependencies between indignities and their effects), yield the same output when run through **Purified Utility Maximization**.

7 Further Work

We started this essay with the goal of developing a modeling framework within which double effect reasoning might be rendered algorithmically implementable and serviceable in contexts of decision under uncertainty. Hopefully, we have made some progress in this regard. Still, there is clearly much remain-

ing work to be done. In this closing section, I would like briefly to highlight some of the many directions one could take in expanding or revising the models deployed here.

While we have aimed for generality wherever possible in developing our framework, we have at several key points allowed considerations of tractability to limit these aspirations. For example, we have required that event variables be binary and that values be additive. These assumptions enabled us to define purified values and utilities with relative ease. In the absence of such admittedly limiting assumptions, however, adequately defining these notions becomes a much subtler affair. It is far from obvious that such a task is hopeless, however, and hence investigating the prospects for generalizing our model of moral decision problems so as to countenance non-binary event variables and non-additive values seems well worthwhile.[30]

Another critical restriction of our framework as so far developed is that its analysis of causation is perfectly deterministic. We have assumed that an event variable causally depends upon a set of parent variables only when the true values of these parent variables suffice to deterministically fix the child's own true value. If our world turns out to be fundamentally indeterministic, however, not all relations of causal dependence need take this form. That is, some variables might depend upon others indeterministically, with determination of the parent variables only fixing a chance distribution over the child's values. Under such a picture, an agent's actions might, even apart from any subjective uncertainty regarding causal structure, fail to guarantee the realization of any particular set of consequences. How should agent's factor such *objective risk* into their moral deliberation? In a manner analogous to our approach to handling subjective uncertainty or in some other way? The matter requires careful consideration.

Further, 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

[30] In the course of carrying out such generalizations, it may prove helpful to compare the approach developed here with recent work on *path-specific objectives* carried out by, e.g., (Farquhar et al., 2022), who suggest a fascinating model of decision making in the context of AI safety research that bears more than faint affinity, at least in spirit if not in technical detail, to the one developed here for quite different ends.

learning events. For example, perhaps pulling a lever in a certain way leads to a subsequent choice of pulling another lever.[31] 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. The properties exhibited by **Purified Utility Maximization** in such a setting ought then to be examined.[32]

Finally, we have striven to capture in our framework only one general class of deontological moral principles, namely, those akin to DDE. But many DDE proponents don't take the doctrine to offer a fully adequate corrective to the errors of consequentialism. Some believe additional deontological principles are needed as well, like perhaps the *Doctrine of Doing and Allowing (DDA)*, according to which there is a moral asymmetry between *actively doing* harm to someone and *merely allowing* comparable harm to befall someone.[33] It would be natural to extend the framework developed here to accommodate moral reasoning built upon principles like DDA via the introduction of further normative parameters. Notably, it may prove useful in any such endeavors to lean upon the work of causal modelers who have sought to incorporate notions like *normality* and *default conditions* into their formal analyses of *actual causation*, and perhaps interpret DDA as a prohibition against *actually* causing certain sorts of evils.[34]

I make no claim that the modeling framework developed above and its associated decision rule of **Purified Utility Maximization** possess anything like sufficient representational power to enable us to map out morality's entire landscape; no formal framework is capable of doing as much. Rather, we will have reached the goals set by this essay if our models merely succeed as a reasonably serviceable tool for understanding and mapping out (albeit in an idealized way) what proponents of DDE-type moral theories have taken to be

[31] Relevant here would be the *tree trolley* cases discussed by (Kamm, 2007).

[32] An obvious question to ask here concerns the *dynamic consistency* of **Purified Utility Maximization**, an attractive property often explored in purely decision theoretic contexts, e.g., in (Rothfus, 2020).

[33] For accounts and defenses of DDA, see, *inter alia*, (Foot, 1967, 1984, 1985), (Quinn, 1989a), and (Woollard, 2013, 2015).

[34] For a sampling of such approaches, see (Halpern, 2008, 2016), (Halpern & Hitchcock, 2015), and (Gallow, 2023). These authors are sensitive to complex issues involving causal preemption and overdetermination as well, which a reviewer has highlighted may be relevant to modeling deontological reasoning about various further cases in which some are wont to give non-consequentialist advice (e.g., Bernard Williams' famous example in (Williams & Smart, 1973) of a man offered the choice of shooting a single person in place of someone else shooting this same person along with others).

an important swath of that terrain.

References

Aboodi, R., Borer, A., & Enoch, D. (2008). Deontology, individualism, and uncertainty: A reply to jackson and smith. *Journal of Philosophy*, *105(5)*, 259–272.

Alexander, L. (2016). The means principle. In K. Ferzan & S. Morse (Eds.), *Legal, moral, and metaphysical truths: The philosophy of michael s. moore* (pp. 251–264). Oxford University Press.

Anscombe, E. (1961). War and murder. In *Nuclear weapons: A catholic response* (pp. 45–62). Sheed; Ward.

Bennett, J. (1995). *The act itself*. Oxford University Press.

Bjorndahl, A., London, A., & Zollman, K. (2017). Kantian decision making under uncertainty: Dignity, price, and consistency. *Philosophers' Imprint*, *17*, 1–22.

Boyle, J. (1980). Toward understanding the principle of double effect. *Ethics*, *90*, 527–38.

Bronner, B., & Goldstein, S. (2018). A stronger doctrine of double effect. *Australasian Journal of Philosophy*, *96(4)*, 793–805.

Cavanaugh, T. (2006). *Double-effect reasoning: Doing good and avoiding evil*. Clarendon Press.

Colyvan, M., Cox, D., & Steele, K. (2010). Modelling the moral dimension of decisions. *Noûs*, *44(3)*, 503–529.

Dietrich, F., & List, C. (2017). What matters and how it matters: A choice-theoretic representation of moral theories. *Philosophical Review*, *126(4)*, 421–479.

Donagan, A. (1977). *The theory of morality*. The University of Chicago Press.

Dougherty, T. (2013). Agent-neutral deontology. *Philosophical Studies*, *163(2)*, 527–537.

Farquhar, S., Carey, R., & Everitt, T. (2022). Path-specific objectives for safer agent incentives. *AAAI*.

FitzPatrick, W. J. (2003). Acts, intentions, and moral permissibility: In defence of the doctrine of double effect'. *Analysis*, *63(4)*, 317–321.

Foot, P. (1967). Abortion and the doctrine of double effect. *Oxford Review*, *5*, 28–41.

Foot, P. (1984). Killing and letting die. In *Abortion: Moral and legal perspectives* (pp. 355–382). University of Amherst Press.

Foot, P. (1985). Morality, action, and outcome. In *Morality and objectivity: A tribute to j.l. mackie* (pp. 23–38). Routledge; Kegan Paul.

Fried, B. (2012). What *does* matter? the case for killing the trolley problem (or letting it die). *The Philosophical Quarterly*, *62*, 1–25.

Gallow, D. (2023). How to trace a causal process. *Philosophical Perspectives*, *36(1)*, 95–117.

Goodall, N. (2016). Away from the trolley problem and toward risk management. *Applied Artificial Intelligence*, *30*, 810–821.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.

Gury, J. (1874). *Compendium theologiae moralis*. Georgii Josephi Manz.

Halpern, J. (2008). Defaults and normality in causal structures. In *Principles of knowledge representation and reasoning: Proceedings of the eleventh international conference* (pp. 198–208). AAAI Press.

Halpern, J. (2016). *Actual causality*. MIT Press.

Halpern, J., & Hitchcock, C. (2015). Graded causation and defaults. *British Journal for Philosophy of Science*, *66(2)*, 413–57.

Halpern, J., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. *AAAI*.

Hansson, S. (2012). A panorama of the philosophy of risk. In *Handbook of risk theory* (pp. 27–54). Springer.

Hitchcock, C. (2023). Causal Models. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2023). Metaphysics Research Lab, Stanford University.

Jackson, F., & Smith, M. (2015). The implementation problem for deontology. In B. Maguire & E. Lord (Eds.), *Weighing reasons* (pp. 279–91). Oxford University Press.

Jeffrey, R. (1965/1983). *The logic of decision*. University of Chicago Press.

Jensen, F., & Nielsen, T. (2007). *Bayesian networks and decision graphs*. Springer.

Kamm, F. (2007). *Intricate ethics*. Oxford University Press.

Lazar, S. (2017). Deontological decision theory and agent-centered options. *Ethics*, *127*, 579–609.

Lazar, S. (2020). Duty and doubt. *Journal of Practical Ethics*, *8(1)*, 28–55.

MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral uncertainty*. Oxford University Press.

Masek, L. (2018). *Intention, character, and double effect*. Notre Dame Press.

Mikhail, J. (2011). *Elements of moral cognition: Linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge.

Nagel, T. (1986). *The view from nowhere*. Oxford University Press.

Nelkin, D., & Rickless, S. (2014). Three cheers for double effect. *Philosophy and Phenomenological Research*, *89*, 125–158.

Nelkin, D., & Rickless, S. (2023). Non-consequentialist principles under conditions of uncertainty: A framework. In *The trolley problem* (pp. 79–100). Cambridge University Press.

Nyholm, S. (2023). Ethical accident algorithms for autonomous vehicles and the trolley problem: Three philosophical disputes. In *The trolley problem* (pp. 211–230). Cambridge University Press.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, *19*, 1275–89.

Oddie, G., & Milne, P. (1991). Act and value: Expectation and the representability of moral theories. *Theoria*, *57*, 42–76.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press.

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

Quinn, W. (1989a). Actions, intentions, and consequences: The doctrine of doing and allowing. *Philosophical Review*, *98(3)*, 287–312.

Quinn, W. (1989b). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy and Public Affairs*, *18(4)*, 334–351.

Robinson, A. (1996). *Non-standard analysis*. Princeton University Press.

Rothfus, G. (2020). Dynamic consistency in the logic of decision. *Philosophical Studies*, *117(12)*, 3923–34.

Setiya, K. (2018). Must consequentialists kill? *Journal of Philosophy*, *115(2)*, 92–105.

Stuchlik, J. (2022). *Intention and wrongdoing: In defense of double effect*. Cambridge University Press.

Tadros, V. (2015). Wrongful intentions without closeness. *Philosophy and Public Affairs*, *43(1)*, 52–74.

Tarsney, C. (2018). Moral uncertainty for deontologists. *Ethical Theory and Moral Practice*, *21*, 505–520.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59(2)*, 204–217.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*, 1395–1415.

Tomlin, P. (2019). Subjective proportionality. *Ethics*, *129*, 254–283.

Wedgwood, R. (2011). Defending double effect. *Ratio*, *24(4)*, 384–401.

Williams, B., & Smart, J. (1973). *Utilitarianism: For and against*. Cambridge University Press.

Woollard, F. (2013). If this is my body...: A defence of the doctrine of doing and allowing. *Pacific Philosophical Quarterly*, *94*, 315–341.

Woollard, F. (2015). *Doing and allowing harm*. Oxford University Press.