Newcomb's Problem and Beyond

Choice, Determinism, and Cooperation

Gary L. Drescher
Center for Cognitive Studies
Tufts University
gld@alum.mit.edu

Abstract

A goal-pursuing agent must somehow ascertain when an action would serve as a means to achieving a goal. Various criteria (causal, evidential, counterfactual) have been proposed (e.g. Joyce 1999). Examining Newcomb's Problem (Nozick 1969) and more-mundane thought experiments, I argue for an acausal but non-evidentialist counterfactual criterion (but without invoking the "possible worlds" of e.g. Lewis 1973) for means-end recognition: an agent acts for the sake of what the outcome *would* then be, not necessarily for what the action *causes*.

Newcomb's Problem posits an imaginary situation in which a large reward was (irrevocably) set up for you if and only if a reliable prediction anticipated that you would now make a choice which (apart from the large reward) is slightly unfavorable to you. A paradox arises as to whether to make the choice that (almost certainly) implies that you reap the reward. Newcomb's Problem bears on the compatibility of choice with determinism. Further, Lewis (1979) and others argue that the Prisoner's Dilemma (e.g. Shubik 1982), a thought experiment that bears on the rationality of altruistic cooperation, reduces to Newcomb's Problem.

I propose a radical variant of Newcomb's Problem in which the already-inalterable reward outcome is also already *visible* to you, yet (I claim) you are still correct to act in pursuit of that outcome. I argue that the radical variant is key to reconciling choice with determinism, and that it suggests a foundation for cooperative behavior that goes substantially beyond what a resolution of the Prisoner's Dilemma alone would provide.

Keywords: cooperation, counterfactuals, decision theory, Newcomb's Problem, payoff dominance, Prisoner's Dilemma

1. Inalterability does not imply futility

If you have a goal that would be achieved if and only if you took a particular action, then (other things being equal) it makes sense for you to take that action for the sake of the goal. If there is something your action cannot alter (cannot make different from what it already is), then it is futile for you to act for the sake of its being one way or another. These two innocuous-sounding intuitions normally coexist peacefully. But the thought experiment known as Newcomb's Problem (Nozick 1969) places them in contradiction of one another, exposing a deep underlying conflict.

In Newcomb's Problem, a mischievous benefactor offers you a transparent box containing $1,000. Awhile ago, the benefactor predicted, very reliably, whether you would choose to accept or refuse the transparent box. There is an adjacent opaque box which is yours no matter what. In it, your benefactor has already placed $1,000,000 for you if the prediction showed you would refuse the transparent box; otherwise, the opaque box was left empty. The opaque box has been sealed, and its content cannot subsequently change—whichever choice you make, the box content ($0 or $1M) stays the same as whatever it already was before your choice. You are accurately and convincingly informed of these circumstances. If and only if you were to take just the opaque box, you would expect to find $1M in it, so that's apparently the right choice. But taking *both* boxes gets you an extra $1,000 and cannot alter whether there's $1M in the already-sealed opaque box, so *that's* apparently the right choice instead.

Now the two prescriptive intuitions conflict: *take an action for the sake of what would be the case if and only if you so acted*, vs. *don't bother to take an action for the sake of what your action cannot alter.* This conflict bears on the oft-perceived incompatibility between choice and determinism, and also bears on a similar intuitive conflict about the rationality of cooperative action in Prisoner's Dilemma situations:

- For any goal in a deterministic universe, the already-irrevocable past state of the universe (like the already-irrevocable opaque-box content) already guarantees the achievement or non-achievement of the goal, seemingly rendering any action futile.
- Similarly, suppose you and someone else must decide independently whether to cooperate with one another in a Prisoner's Dilemma situation where both of you would be better off if you both cooperate, but only the *other's* cooperation (not your own) *causes* any benefit to you. From the standpoint of your self-interest, it is seemingly futile for you to cooperate, since the other's decision whether to cooperate is inalterable by your own decision.

I argue here that (quantum indeterminacy notwithstanding) our universe is deterministic and predictable *enough—especially* as regards choice—that the non-futility of our choices can be vindicated if and only if choice would make sense even given full determinism; hence, it is important to reconcile choice with determinism, regardless of the details of actual physics. This reconciliation requires rebutting the fatalist intuition that inalterability implies futility. The case for cooperative action (even when your own cooperation cannot alter others' causally independent choices) further tests that rebuttal, as well as being important in its own right.

I argue in favor of choice given determinism, of cooperative action even without a causal link to others' cooperation, and of declining the $1,000 in Newcomb's Problem. Diverging in part from the familiar arguments for reconciling choice with determinism, I confront the inalterability-implies-futility intuition head-on by proposing a variant of Newcomb's Problem in which *both* boxes are transparent. Yet, I maintain, declining the $1,000 (in pursuit of the $1M) is *still* the right choice. The transparent-boxes problem has been offered as a *reductio ad absurdum* of declining the $1,000 in the original, opaque-box scenario (Gibbard and Harper 1977); after all, making the box transparent only makes obvious what was true all along, namely that the box already had a definite, inalterable content. I accept the *reductio* but challenge the *absurdum*: acting in pursuit of the $1M in the original, opaque-box scenario is rational only if it remains so when both boxes are transparent—and I claim it does.

But, as discussed below, even in mundane situations (with no fantastic predictors or the like) in a deterministic world, an already-inalterable goal state is often already known as well. I thus argue that if choice makes sense in a deterministic world—if, therefore, it makes sense to act for the sake of a goal whose achievement or non-achievement is already inalterable—then it makes sense even if the already-inalterable state is also plainly visible. A hidden inalterable state is merely less obvious, obscuring but not resolving the conflict with the fatalist intuition that inalterability implies futility. In addition, defending the one-box-only choice when both boxes are transparent supports a more robust version of the Prisoner's Dilemma in which cooperative action is defensible even when you already know whether the other's choice was to cooperate.

In what follows, section 2 asks under what circumstances it makes sense to take a given action for the sake of a given goal. Section 3, following e.g. Dennett (1984) and Minsky (1968, 1986), discusses acting for goals despite

determinism. Section 4 argues that it can make sense to act even for a goal that your action does not cause. Section 5 argues that nonetheless, sensible goal-pursuit requires more than a correlation between action and goal, and section 6 suggests what (if not causation) the extra ingredient might be. Section 7 analyzes Newcomb's Problem in light of the preceding discussion; section 8 considers the case where both boxes are transparent. Section 0 looks at the Prisoner's Dilemma as a variant of Newcomb's problem. Section 10 summarizes.

2. Means-end relations

Let us say that there is a *means-end* relation between a contemplated action and a goal just in case the desirability of the goal rationally contributes motivation for taking the action—that is, just in case, other things being equal (i.e., in the absence of conflicting consequences of higher priority), it makes sense to take the action for the sake of the goal.

By what criteria can we recognize the existence of a particular means-end relation? Prominent suggestions include (in order of increasing strictness):

- *Evidential* or *correlational* criterion: there is a means-end link from action to goal just in case the goal is more likely to be found to obtain when the action is found to be taken than when the action is found not to be taken (i.e., the action's occurrence is correlated with, and thus gives evidence of, the goal's occurrence).
- *Subjunctive* or *counterfactual*[1] criterion: there is a means-end link from action to goal just in case the goal *would* obtain if the action were to be taken, but not otherwise (or would more likely obtain if the action were to be taken than if otherwise).
- *Causal* criterion: there is a means-end link from action to goal just in case the action causes (or tends to cause) the goal to obtain.
- *Fatalist* criterion: there is never a means-end link from action to goal; all actions are futile. (No one takes fatalism seriously, but many believe it would indeed follow if the universe were deterministic; hence, they reject determinism.)

Often, the first three criteria coincide. Say I take the action of crossing the street in order to achieve the goal of getting to the other side. (Let's construe that action as initiating a series of muscle contractions, not as the actual passage across the street, so the goal's achievement doesn't follow tautologically from the action's occurrence.) Knowing that I will cross informs me (fairly reliably) that I will get to the other side, whereas knowing I will not cross informs me otherwise, fulfilling the evidential criterion. If I were to walk across the street, I would (very likely) get to the other side, but (very likely) not otherwise, fulfilling the subjunctive criterion. And finally, my walking across the street causes me to get to the other side, fulfilling the causal criterion. By any of those three criteria, there is a means-end link from the action of crossing to the goal of getting to the other side. Given that means-end link, and other things being equal, my desire to be on the other side rationally motivates my crossing.

In Newcomb's Problem, though, the criteria diverge: taking just the opaque box, forfeiting the $1,000, is strong evidence that you obtain $1M in the opaque box; whereas taking both boxes is strong evidence that the opaque box is empty. But taking the transparent box or not has no causal influence on the content of the already-sealed opaque box. The evidential criterion says there is a means-end link from the action of taking just the opaque box, to the goal of obtaining $1M in the opaque box; but the causal criterion says otherwise.

The subjunctive criterion's verdict, meanwhile, seems ambiguous, in part because of the broad range of intuitions as to what *would* means. Indeed, in the loosest construal, subjunctive links mimic evidential links: e.g., if you were to take just the opaque box, there would be—i.e. *would have to be* (given the circumstances)—$1M in the opaque box.[2] In a stricter construal, subjunctive links are just causal links: what would be the case if you were to take just the opaque box is simply whatever that action causes (and nothing more).[3]

---

[1] Subjunctive inference differs from material implication in that with the latter, a false antecedent implies *any* consequent. Subjunctive reasoning permits you to entertain a contrary-to-fact antecedent (hence *counterfactual* reasoning) and imply only what is relevant, in some elusive sense. So, for example, *If I had dropped my glass just now, it would have shattered* (assuming I did not in fact drop it) could be a true subjunctive statement, but *If I had dropped my glass just now, then the earth would be flat* is false, even though *just-dropped-glass* does materially imply *flat-earth* if I did not actually drop the glass.

[2] Horgan (1981), for example, argues that counterfactual propositions, properly construed for decision-making purposes, correspond to evidential links.

[3] Pearl (2000) and Joyce (1999), for example, identify what follows counterfactually with what follows causally. Gibbard and Harper (1977), discussing Newcomb's Problem, refer to the evidential approach as *V-maximization* (maximizing the expected utility given an action, compared to the expected utility given some other action), in contrast with the counterfactual *U-maximization* approach, which they advocate. V-

I argue below that there is a distinct, intermediate sense of *would*—narrower than correlation but broader than causation—that provides the correct subjunctive criterion for means-end relations—that is, the sense of *would* such that, to the extent that a goal would more likely obtain if a given action were taken than if not, the desirability of the goal rationally contributes to the motivation for taking the action. Call this the *choice-supporting* sense of *would*.

Below, I briefly recapitulate (and agree with) the usual argument that the evidential criterion is too lax—that it sometimes prescribes means-end relations where none exist. To make the case that the causal criterion is too strict, I argue that its appeal stems from the same fatalist fallacy that makes all choice seem futile given determinism—namely, the notion that inalterability implies futility. After outlining the case for the compatibility of choice and determinism, I present a (comparatively) clear-cut example of an acausal means-end link. And I argue that the principles extracted from rebutting the fatalist fallacy, and from considering a clear-cut acausal means-end link, lead to a version of the subjunctive means-end criterion that supports the assertion of an acausal means-end link in Newcomb's Problem and the Prisoner's Dilemma.

Before starting the game, let us pause to clarify the rules. Given a set of goals, which (putative) means-end criteria *should* you use as the basis for choosing actions, in order to best achieve the goals that are thus linked to? The question is circular: it asks what means-end-recognizing policy's use is a means to best achieving your goals. The answer cannot be deduced without already having some basis for recognizing means-end relations.

Fortunately, this circularity need not be paralyzing. It is reminiscent of the status of inductive reasoning, which, as Hume noticed, cannot be deductively supported; we may observe that induction has worked well in the past, but to therefore expect it to work in the future is circular. Nevertheless, it is easy to see, in broad terms, how intelligent organisms would have evolved to use inductive reasoning, based in part on some hardwired kernel of that reasoning; and similarly with means-end reasoning. Presumably, then, we perceive means-end links—at least in some routine situations—based in part on some built-in criteria for their recognition. We need not (and cannot) deduce those criteria from first principles, nor need we construct an explicit formalization of those criteria—but if we do the latter, we can expect to extend our innate kernel of competence to deal properly with more subtle problems,[4] just as having an explicit formal theory of induction extends our ability to make accurate predictions.

The game, then, is to use our means-end intuitions in clear-cut, routine situations to try to elicit explicitly the principles that drive those intuitions; Dennett (1980) calls such situations *intuition pumps*. Those explicit principles can then be applied to trickier situations (like Newcomb's Problem and the Prisoner's Dilemma) where the intuitive verdict itself is vague or ambiguous (analogous, say, to performing an explicit calculation of the statistical significance of a given sample, rather than just trusting what our intuition says about it).

Crucially, the principles we extract from the clear-cut situations are not merely *descriptive* of our means-end intuitions in those situations. If all goes well, the principles turn out to be *meta-circularly consistent*: they plug back into our means-end intuitions such that using those principles (rather than some others, or none) will indeed strike us as a good idea, as a means to achieving our goals. Thus, the principles will be *prescriptive* as well as descriptive. They will tell us what means-end criteria we *should* use, where the should-ness is ultimately grounded in what our means-end intuitions prescribe in clear-cut exemplary situations, and what they prescribe when we contemplate the use of the means-end criteria themselves.

Accordingly, the structure of the following argument is to elicit means-end intuitions in (compartively) mundane, uncontroversial situations—the choice machines of section 3, the hand-raising scenario in section 4, and the street-crossing scenario in section 5—and sketch a means-end-recognizing mechanism to account for those intuitions. Having thus motivated the mechanism, I then recruit it to analyze Newcomb's Problem and the Prisoner's Dilemma.

3. Deterministic choice

Consider a deterministic, artificial universe defined, say, by a computer program. The universe is superficially like our own, with three-dimensional space and various physical objects, including agents with sensory inputs, motor actions, and internal representations of various aspects of the state of the world. I make the case that meaningful

---

maximization computes conditional utility with respect to the conditional probability of an outcome given an action; U-maximization instead uses the probability that an outcome *would* occur if an action were to occur. Gibbard and Harper's U-maximization invokes considerations of physical law and causal independence, leading to a causal sense of *would*.

[4] For instance, you probably know someone who would undoubtedly act to dodge an imminent collision, but who, on the other hand, declines to wear a seatbelt on the grounds that *When your time is up, there's nothing you can do about it.* When consequences are sufficiently immediate and obvious, no one takes fatalist resignation seriously enough to behave accordingly. But when the consequences can only be appreciated with more-abstract reflection (e.g. about a small probability of a large consequence), one's abstract conception of means-end relations becomes more influential, and a fatalist theory may indeed impede an appropriate action that a better theory would instead promote.

choice by such agents is possible, and that any real-world indeterminacy is therefore unnecessary (and also insufficient) to account for genuine choice.

Suppose each agent's control system contains subjunctive assertions of the form *If conditions X apply* (the assertion's *context*)*, taking action Y would result in conditions Z* (the assertion's *result*)*,* with a specified degree of reliability (even with determinism, an action's result can differ depending on conditions not mentioned in the context), in the choice-supporting sense of *would*, as discussed in the previous section. Ignore for now the question of how the agent obtained these assertions; it just has them.

The context, action, and result conditions are each expressed as a predicate applied to the agent's current situation. The system also assigns quantitative *utilities*, positive or negative, to some conditions. The system can recognize when, according to the extant subjunctive assertions, an action or series of actions would result in a condition of positive or negative utility. This recognition influences the system for or against taking those actions. The strength of the influence is proportional to the utility, and to the stated reliability of the assertions.

Despite the determinism of the agent and its universe, the following sorts of questions are answerable with regard to the agent:

> Why did it take that action? What goal was the action selected for? Was that goal achieved? Would it have been achieved if the machine had taken this other action instead? The system includes the assertion that if it did *X*, then *Y* would (probably) have happened; is that true? The system does not include the assertion that if it did *P*, *Q* would probably have happened; is that omitted assertion true? Would the system have taken some other action if it had included that assertion? Would it then have better achieved its goals?

Insofar as such questions are meaningful and answerable, the agent makes choices in at least the sense that the correctness of its actions with respect to its designated goals is analyzable. That is to say, there can be means-end connections between its actions and its goals: its taking an action for the sake of a goal can make sense. And this is so despite the fact that everything that will happen—including every action taken and every goal achieved or not—is inalterably determined once the system starts up. Accordingly, I propose to call such an agent a *choice machine.*

Seeing choice as a mechanical process that contemplates hypothetical actions and outcomes addresses some of fatalism's central challenges to deterministic choice:

- Why (in some cases) does it make sense for us to take an action for the sake of an already inalterably obtained goal? Answer: because (in those cases) if we didn't, the goal wouldn't obtain. This answer is perfectly compatible with the fact that the goal *had to* obtain; there is no contradiction, because the action had to obtain too.
- But why, then, does it make sense to contemplate alternative actions, and their results, and select from among them, if it is already determined which sole action is chosen? Similarly: because if not for that contemplation and selection, the preferred action wouldn't (necessarily) have been taken. This is compatible with the fact that it *had to* be taken; there is no contradiction, because the contemplation and selection—i.e., the choice process—had to occur too.

In addition to making choices that are subject to teleological analyses, a choice machine comports with two intuitions which are plausibly among those that may lead people to conclude that their choices are not predetermined:

- Ordinarily, when you do X, you believe you did so only because, on the whole, at that moment, given the circumstances, you wanted to. If instead you had on the whole preferred at the time to do Y, you would have done that instead. Your choice is not bound by external constraints, except insofar as they exert influence through your preferences themselves (for instance, by attaching a consequence that you will regard as a penalty to an otherwise preferable action, e.g. by putting a gun to your head).
- If the universe is deterministic, then in principle someone could predict your choice in advance. But if someone did so and then told you the prediction, nothing would stop you from deliberately doing the opposite of whatever was predicted, thus making the prediction false. Its falsehood seemingly contradicts the statement that the prediction could be made in principle, thus in turn contradicting the premise that the universe is deterministic.

The first of these intuitions is as true of an artificial choice machine as of a human being: the machine's preferences and machinery do indeed control its actions, and if its preferences had dictated a different action, then its action would have been correspondingly different. Of course, its preferences and knowledge are in turn caused by past events beyond its control; it cannot just choose its current preferences as it can its current actions. But neither can we.

The second intuition also applies to artificial choice machines as well as to human beings. The determinism of the choice machine's universe indeed implies the predictability of its choices; and its world could easily be so constructed that an entity embodied in that world could carry out such a prediction, say by taking a snapshot of a sufficient portion of the state of the world (including the choice machine), and then applying the deterministic laws of the world to calculate in advance how that portion of the world will evolve.

It is also true that the machine's preferences could be such that if the mechanism were told which action it is about to choose on a given occasion, it would be sure to choose a different action instead. But this ability to thwart a prediction does not contradict the determinism of the machine's world (as discussed e.g. in MacKay 1960 and Dennett 1984). If an entity embodied in the machine's world is going to predict the machine's choice and then tell the choice machine the prediction before the choice is made, then the predictor cannot necessarily carry out the prediction in the first place. For in order to carry out the requisite simulation, the predictor must among other things simulate the choice machine's being told the prediction. To do that, the simulation must specify which prediction will be conveyed; but it doesn't know which prediction will be conveyed until the simulation is complete.

The predictor might, of course, carry out two separate simulations (assuming a binary choice of available actions), each presupposing a different prediction. (Such a tactic figures in the discussion below of Newcomb's Problem with both boxes transparent.) If either simulation predicts an action that accords with the prediction conveyed to the choice machine in that simulation, then that prediction can be made accurately. But if both simulations show the conveyed prediction being thwarted—as will be the case if the choice machine's preferences are such that it always acts to contradict the conveyed prediction—then the prediction cannot be accurately made and conveyed. Despite the determinism of the choice machine's universe, including the choice machine itself, the machine, like us, can easily choose to thwart any prediction conveyed to it.

Thus choice, in the sense just delineated, is a mechanical process compatible with determinism. The objection *The agent didn't really make a choice, because the outcome was already predetermined* is as much a non sequitur as the objection *The motor didn't really exert force, because the outcome was already predetermined*. (Or as Dennett 1984 puts it, wouldn't a predetermined thunderstorm still be a real thunderstorm?) Both choice-making and motor-spinning are particular kinds of mechanical processes. In neither case does the predetermination of the outcome imply that the process didn't really take place.[5]

To the extent that the past determines the future in the real universe, genuine choice is no more precluded than in a fully deterministic world. Some degree of real-world indeterminacy is also compatible with choice:

- Indeterminacy of some external events is no different, for most ordinary purposes, from an agent's limited-knowledge perspective, which makes an outcome seemingly random to the extent that the outcome depends on definite but unknown conditions. But excessive external indeterminacy or unpredictability would make choice futile: if just about anything could happen (or could happen as far as you know) regardless of what you were to do, there'd be no point in making any particular choice instead of some other.
- Likewise, some indeterminacy within the choice process itself is harmless, say as a tie-breaker among choices of comparable utility. But too much internal indeterminacy would again make choice futile: if, say, seeing an oncoming car, and wanting not to be struck, did not determine reliably that you'd choose to step out of the way, then choice would be of little use.

Moderate indeterminacy is compatible with choice, but not necessary to it. Profligate indeterminacy—contrary to a popular intuition—is subversive of choice, rather than supportive of it.

Accordingly, in the examples to follow, I presume a predictable universe, either by idealizing our own universe as such (both by presuming deterministic physics, and by treating some miniscule uncertainties—uncertain due to limited knowledge—as having zero rather than miniscule probability) or by considering an artificial, deterministic universe populated with choice machines. The lessons extracted apply to the real world insofar as real-world choice is not dependent on whatever moderate indeterminacy does exist.

4. Acausal means-end links: choosing past states

Determinism does not imply fatalism, because inalterability does not imply futility. As just discussed, it can make sense for an agent to (be designed to) choose an action for the sake of what would be the case if the action were

[5] The term *free will* scarcely appears herein. As typically used, that term connotes something like *meaningful choice, as opposed to determinism.* Since I am defending meaningful choice, even given determinism, I avoid the f-w word and refer simply to the choice process.

taken (in the appropriate, choice-supporting sense of *would*), even though all events (including the achievement or non-achievement of any goals) are already inalterably determined.

Must an action *cause* a goal's achievement in order for there to be a means-end relation? A compelling reason to think so is that in the absence of a causal link, the goal's achievement is inalterable by the action. But as just discussed, any goal's achievement is inalterable *anyway*, given determinism; still, it can make sense to act for the sake of a goal. Such is sometimes the case, I claim, even for a goal that is not caused by an action.

Consider a simple example, presuming that our universe is deterministic (or consider a similar example set in an actually deterministic alternative universe). Define the predicate *P* to be true of the total state of the universe at a given moment if and only if the successor state one billion years thence—i.e., the state defined by applying the laws of physics to the given state to predict the new state one billion years later—shows my right hand raised. Suppose, on a whim, I would like the state of the universe one billion years ago to have been such that the predicate *P* is true of that state. I need only raise my right hand now and voilà, it was so.

Of course, I did not change what the distant-past state of the universe had been; the past is what it is and can never be changed. Furthermore, I have no causal influence over the past. Nonetheless, physical law (if deterministic) necessitates that if I do raise my right hand, *P* is in fact true of the state of the universe a billion years ago; or if I do lower it, *P* is false of that past state. Suppose, despite my wanting *P* to have been true a billion years ago, I forego raising my hand due to a belief that it would be futile to act for the sake of something past and inalterable. In that case, as always with fatalist resignation, I would be needlessly forfeiting an opportunity for my goal to be achieved. By physical necessity, the set of universe-states in which I do raise my hand is coextensive with the set of universe-states for which *P* was in fact true of the state a billion years prior. I thus have *exactly* as much choice about that particular aspect of the past, despite its inalterability, as I have about whether to raise my hand now—despite the inalterability of that too in a deterministic universe. And, as argued in the previous section, that much choice is choice enough.

There can also be an acausal means-end link in the presence of some uncertainty. Suppose, for instance, that the world's physical laws are *nearly* deterministic, but with some infinitesimal chance that the laws by which a distant-past state ordinarily leads to my raising my hand (or to the absence of that event) have an exception in a particular instance. There is still an (acausal) means-end link from my raising my hand, to the past state such that I would (almost certainly) raise my hand, albeit attenuated by the miniscule chance that a different past state—one that is almost certainly not a hand-raising precursor—led to hand-raising on this occasion.

A clarification is in order about the meaning of *cause*. One might maintain that a means-end relation is inherently causal by the very definition of cause: if you choose whether something is the case, then tautologically you thereby cause it to be the case or not. But this issue is merely a matter of terminology. I wish to distinguish two concepts: the concept of the propagation of information from a partial state of the universe to other states according to physical laws; and the concept of a means-end relation between an action and a goal state, such that if the state is desired the mechanism should take the action, other things being equal. I use *cause* to designate the first of these concepts; and regardless of terminological conventions, the substantive question here is whether the first of these is necessary for the second, or whether instead a subjunctive link can constitute a means-end link in the absence of a physical-propagation link from action to goal—i.e., in the absence of a causal link, as I use that term.

The past-predicate example argues that there are indeed acausal means-end links; I argue below that other such links occur in Newcomb's Problem and the Prisoner's Dilemma. Exhibiting an example, however trivial, of an acausal means-end link suffices to rebut the contention that means-end links are necessarily causal, and clears the way for more-interesting examples.

5. Street-crossing scenario: avoiding evidentialist excess

The means-end link to predicate *P* above is inconsistent with a causal criterion for means-end links, but not inconsistent with an evidential criterion: the probability of *P* being true of the distant past given that I raise my hand (i.e., 1) is indeed higher than the probability of its being true given that I do not raise my hand (i.e., 0).

But using an evidential criterion is problematic in a deterministic and predictable enough universe. To illustrate, consider the following scenario. I stand on a street corner, wanting to be on the other side of the street. I have a clear view of any oncoming traffic, and have looked carefully. I strongly prefer not to be struck by traffic, so I would not cross the street at a moment when, having just looked carefully with a clear view, I see dense, dangerous, onrushing traffic. Let us make the idealizing assumption that the probability of my crossing under such circumstances, given my preference not to be struck, is zero; and assume I know it. Thus, we ignore the infinitesimal possibility that,

say, some cosmic rays will disrupt my neurons such that I knowingly cross dangerously now despite my preference not to and despite my competence, having looked carefully with a clear view, to act accordingly.

To evaluate whether an action would be a means to achieving a particular state, an evidentialist compares the conditional probability of the state given that the action is taken with the conditional probability of the state given that another action is taken instead. But if I already know I will not cross in the presence of clearly-seen dangerous traffic—if the probability of my doing so is zero—then any probability conditioned on my crossing under those circumstances is simply undefined (because the defining expression is a quotient with a zero denominator). Accordingly, an evidentialist might hold out for the cosmic rays and refuse to accede to the zero-probability idealization.

But let us recast the scenario in an artificial world where a choice machine (with street-crossing knowledge and preferences similar to mine) is waiting to cross the street, but sees dangerous traffic. Suppose the probability of its crossing right now, given its preferences and sensory inputs, is actually zero (the world's physics just doesn't allow for the failure of the choice machinery), and is correctly represented as such by the choice machine itself. Despite that certainty, the machine must still choose, on some basis, whether or not to cross the street; as discussed in section 3, the inalterability of the forthcoming choice process—in this case, a process with an *already-known* outcome—does not imply that no choice is being made. On the contrary, the choice process—comparing what would happen if this or that action were taken, and initiating the action for which what would happen is preferred—operates as always, and crucially so. For it is that very process—the choice machine's anticipation of what the consequences would be, and its initiating an action based on that anticipation—that *makes* it impossible for the machine to cross now, in the presence of dangerous traffic.

The artificial-world thought experiment shows that a nonzero probability of crossing in the presence of dangerous traffic is not necessary for there to be (by some intuitively clear but yet-unexplicated criterion) the usual means-end link, in the presence of dangerous traffic, from the action of crossing, to the (negative-utility) outcome of injury (or from the action of not-crossing, to the goal of non-injury). Thus, even if my crossing dangerously—despite what I see and what I prefer—does have a slightly nonzero probability in the real world, that infinitesimal possibility—along with the cosmic rays etc. that bear it—is irrelevant to why I should not cross the street when the traffic is dangerous. The zero-probability idealization thus does not forfeit any necessary explanation for why I should not cross.

Despite the lack of a defined probability conditioned on the action of crossing in the *present* situation if the present probability of crossing is zero, there is a more or less evidentialist method by which we can still evaluate what would occur if I were to cross the street now: we can compare some actual situations where I cross with other actual situations where I do not. Such comparisons are central to the means-end criterion I advocate; but the approach raises an immediate problem, as follows.

Define the indexical assertions:

> S := (Street-crossing) I am standing at a street corner with a clear view of any traffic.
>
> D := (Disposition) My desire and competence are such that I will cross only if there is no dangerous traffic.
>
> T_1 := (Traffic oncoming) Dense, dangerous traffic is just about to speed past the intersection.
>
> T_2 := (Traffic passing) Dense, dangerous does speed past the intersection a second later [i.e. just as crossing begins, if that action is now taken].
>
> C := (Cross) I cross the street now [i.e., I initiate a particular series of muscle contractions].
>
> O := (Other side) I reach the other side of the street shortly.
>
> H := (Hit) I am hit by traffic shortly.

Each of these assertions is a predicate that applies to my present situation. Making the deterministic idealizations above (or recasting the scenario in a suitably deterministic artificial world), we can then express probabilities such as

$$\mathrm{P}(O{\sim}H \mid S{\sim}T_1C) = 1$$

$$\mathrm{P}(O{\sim}H \mid S{\sim}T_1{\sim}C) = 0\ ,$$

construing the probabilities as frequencies among actual instances. These probabilities express how the likelihood of reaching the other side without collision varies as a function of whether or not I cross, given the street-crossing situation and the absence of dangerous oncoming traffic: if I cross, I reach the other side safely, otherwise not (because otherwise I do not reach the other side at all). And in this example, that contrast also corresponds to the means-end link from crossing (in the absence of dangerous traffic) to reaching the other side without collision (de-

fining crossing as in section 1: initiating a particular sequence of motor actions which, unless I am struck by traffic, suffices to convey me to the other side). Other conditional probabilities express passive predictions as to what will happen next, such as

$$P(T_2 \mid ST_1) = 1$$

$$P(\sim T_2 \mid S\sim T_1) = 1$$

(i.e., the oncoming traffic neither vanishes, nor materializes out of nowhere, as it reaches the intersection).

Here, then, is a problem for an evidential means-end criterion of this sort. Suppose we also have

$$P(T_2 \mid SDC) = 0$$

$$P(T_2 \mid SD\sim C) = 0.3 \quad \text{(say)}.$$

That is, among situations that satisfy *SDC*, T_2 is never satisfied—dangerous traffic is never seen to speed past just as I cross (because of my desire and competence not to cross dangerously). But among situations that satisfy *SD~C*, T_2 is often satisfied; such traffic often does speed by when I do not cross. These probabilities express how the likelihood of the passage of dangerous traffic varies as a function of whether or not I cross, given the street-crossing situation and my desire and competence to cross safely.

Assume that I know that *S*, *D*, and T_1 are true right now. By the last two conditional probabilities above, my taking the action of crossing the street—given my disposition as to safety—is perfect evidence of the absence of dangerous traffic as I cross. And indeed, a third party with no view of the road itself would be justified to bet on the absence of dangerous traffic, just knowing my disposition and that I chose to cross. Crucially, though, that evidential relation does not suffice to establish a means-end link from the action of street-crossing to the goal that there be no dangerous traffic. Even if I prefer the absence of dangerous traffic (say, so I can continue across the street safely), it would be foolhardy and bizarre of me to cross the street now—as the dangerous traffic approaches—for the sake of achieving the traffic's absence—i.e., to misconstrue the evidential link as a means-end link.[6] The conditional probability $P(T_2|SDC)=0$ predicts what one finds as to T_2 whenever one finds *SDC*; it does not necessarily predict what one *would* find now as to T_2 if one were to bring about *C* given that *SD* does obtain.

Of course, the prediction as to T_2-if-*C* expressed by $P(T_2|SDC)=0$ is contradicted by that expressed by $P(T_2|ST_1C)=1$; $P(T_2|ST_1C)=1$ can be inferred from $P(T_2|ST_1)=1$ above if $P(ST_1C)\neq 0$. Given disposition *D*, crossing is perfect evidence for the absence of passing dangerous traffic T_2; but given oncoming dangerous traffic T_1, crossing is no evidence at all for that absence (T_1 is said to *screen off* T_2 from *C*, meaning that $P(T_2|ST_1)=P(T_2|ST_1C)$—i.e., if we condition on T_1, further conditioning on *C* does not change the probability of T_2). Since we know both *D* and T_1, we face an ambiguity as to which of those predicates to condition on. We would like to use both; in general, we should condition on information as specific as is available and relevant. But as noted above, a probability conditioned on SDT_1C is undefined, under the assumption that $P(SDT_1C)=0$.

Regarding what *would* happen now if *C*, $P(T_2|ST_1C)=1$ obviously expresses a more intuitively plausible prediction here than does $P(T_2|SDC)=0$—that is, we know the oncoming traffic wouldn't vanish if I were to cross now—but on what basis? True, one might observe that given the dangerous traffic T_1 (which is in fact present), if I do cross now, it cannot be true that I have disposition *D*; hence, one might argue that when ascertaining what *C* is evidence for, we should not condition on *D*, because *D* itself depends on *C*. But symmetrically, one might observe that given my disposition *D* (which is in fact present), if I do cross now, it cannot be true that dangerous traffic T_1 is present; hence, T_1 apparently depends on *C*, so we seemingly should not condition on T_1. It is only because we (somehow) already know intuitively what *C* is a means to that we know which predicate, *D* or T_1, to condition on—not vice versa.

One might propose an ad hoc rule not to condition on an agent's competence or desires when computing the probability of an outcome given an agent's action in order to appraise a putative means-end link.[7] But suppose someone

[6] A similar point is often made, as by Nozick (1969), in terms of an imaginary scenario in which we discover that smoking does not cause cancer; rather, a gene that predisposes people to smoke also independently predisposes them to cancer. Then, it would be irrational to avoid smoking in order to avoid cancer; there is a correlation, but no causal link and no means-end link, from not smoking, to avoiding cancer. The smoker's scenario is similar to the street-crossing scenario: the gene (like T_1) is a common causal influence upon both a choice (smoking, or *~C*) and upon a second effect which is thereby correlated with the choice (cancer, or T_2); but the choice does not cause the second effect, nor does making the alternative choice serve as a means to avoiding the second effect, despite giving evidence of the second effect's absence.

[7] Christopher Taylor (personal communication) suggested such a rule. Others (see note 9) propose, on the contrary, that we *must* condition on the agent's dispositions and decisions (which in this case yields the wrong answer). Devising rules for means-end recognition is tricky in part because one is tempted to postulate a rule that gives the right answer in the scenario under consideration, without noticing that it gives the wrong answer in other, equally fundamental examples.

has just made a detailed, accurate copy of the agent's cognitive state. We could then condition on that copy instead, circumventing the ad hoc rule. One might try to plug the loophole by also excluding states that are strongly correlated with the agent's disposition. But then we could not even condition on T_1, since my disposition to cross now correlates strongly with the absence of oncoming dangerous traffic.

Here again, an evidentialist might chafe at the zero-probability idealization, and insist on invoking miniscule but nonzero probabilities in order to rebut the false means-end connection from C to $\sim T_2$. The evidentialist is then able to define the conditional probability

$$\mathrm{P}(T_2 \mid SDT_1C) = 1 \ ,$$

conditioning on both T_1 and D. This is the probability that the oncoming traffic does not vanish, given the hypothesis that despite my actual desire and competence to the contrary, I do cross now in the presence of the dangerous traffic. Intuitively, we know that that conditional probability is 1 (or nearly so) if the conjunction SDT_1C is not quite impossible. But in the absence of any actual instances of the all-but-impossible event of my crossing in the path of clearly seen dangerous traffic despite my contrary desire and competence D, on what basis can the evidentialist assign a conditional probability to the continued presence of traffic in that hypothetical situation? The same dilemma arises as with the zero-probability idealization: how do we know whether the oncoming traffic proceeds now as it is always in fact observed to do (i.e., by not suddenly vanishing), or whether instead my act of safely-disposed crossing proceeds now as it is always in fact observed to (i.e., safely and successfully, with no dangerous traffic passing as I cross)? Of course it is intuitively obvious which happens, but how is that intuition implemented?[8] On what basis can we conclude that $\mathrm{P}(T_2|SDT_1C)$ equals $\mathrm{P}(T_2|ST_1C)$ rather than equaling $\mathrm{P}(T_2|SDC)$? (An ad hoc rule against conditioning on the agent's disposition is unfeasible for the same reasons as noted above.)

One might just invoke a raw *subjective* (not to be confused with *subjunctive*) conditional probability (as to what transpires given the almost-impossible and never-actually-occurring hypothetical event SDT_1C), and then assign means-end links according to the subjective probability. That tactic, however, does not explain how to ascertain the means-end link here at all, but rather just passes the buck to whatever homunculus generates the subjective conditional-probability intuition. The putative explanation thus circularly presupposes that something, somehow has already solved the problem. The explanation I seek addresses *how* the intuition here, corresponding to the subjective conditional probability $\mathrm{P}(T_2|SDT_1C)=1$, could reasonably be arrived at by our cognitive machinery, given the contradictory evidence (as to T_2-if-C) offered by actual *SDC* situations (where T_2 is never true) and actual ST_1 situations (where T_2 is always true), and the absence of actual SDT_1C situations (even if they are not quite impossible).

What's worse, the probability conditioned on SDT_1C, even if known somehow, might still create an evidential relationship that is not a means-end relationship. Suppose we write a computer program to simulate an artificial world where $\mathrm{P}(C|SDT_1)=0$, and run that program on a real-world computer (redefining the predicates to refer to events as simulated by the program). The probability of SDT_1C occurring in the running program is nonzero, due to the chance of a hardware error (which can be made arbitrarily unlikely, but not impossible). Suppose the probabilities used by the street-crossing agent accurately reflect (among other things) the possibility of such an error. Depending on the implementation details, it could be that the most likely hardware failure that erroneously yields SDT_1C also erroneously yields $\sim T_2$ (even though, when the hardware works properly, the computer's simulation of SDT_1 is actually causative of its subsequent simulation of $\sim C$ and of T_2). Then, $\mathrm{P}(T_2|SDT_1C)$ could be close to zero, while $\mathrm{P}(T_2|SDT_1{\sim}C)$ is still close to unity

Under those assumptions, the evidentialist contrast shows $\sim T_2$ to be far more likely if C than if $\sim C$, *even given* SDT_1. But clearly, the street-crossing agent's correct decision in the running computer program has almost *nothing to do* with the arbitrarily rare possibility of a hardware error that lets $\mathrm{P}(T_2|SDT_1C)$ be defined, but instead has everything to do with what *would* be the case now—given SDT_1 and, almost certainly, correctly functioning hardware—if, contrary to fact, C were to occur now.[9] Yet if we try to conditionalize explicitly on $\sim F$ too (where F refers

[8] The overwhelming intuitive obviousness of the right answer (as to which outcome would really happen) can obscure the fact that there is even a problem here to be solved. Envisioning the design of a machine that can figure out that answer helps bring the problem into focus.

[9] Horgan (1981) defends evidentialism against Nozick's smoker's analysis (note 6) by appeal to screening-off by one's knowledge of one's inclination prior to acting: conditioned on a smoker's knowledge of her desire to smoke (which desire is what we imagine mediates the gene's influence on smoking), the act of smoking itself contributes no *further* evidence as to the gene or the propensity for cancer. Similarly, Eells (1982) and Jeffrey (1983) argue that in smoker-like scenarios, knowledge of the putative result is screened off from knowledge of the action by knowledge of one's decision just prior to acting (assuming at least an infinitesimal probability that the decision and action are discrepant, to allow the requisite conditional probabilities to be defined). But in the present example, T_1 already serves much the same screening-off function as does the foreknown decision. And if the analysis were recast, conditioning also on the already-known decision not to cross ($\sim C_d$), and thus comparing $\mathrm{P}(T_2|SDT_1{\sim}C_dC)$ to $\mathrm{P}(T_2|SDT_1{\sim}C_d{\sim}C)$, the above problems would remain: 1) the analysis now requires a nonzero probability of $\sim C_dC$ and of SDT_1C, but a thought experiment set in an artificial world where those probabilities are zero shows that a coherent choice is still possi-

to hardware failure), we once again confront an undefined, zero-denominator probability, $P(T_2|SDT_1C{\sim}F)$—even though we are now analyzing an entirely feasible real-world situation (involving the behavior of a real-world computer) without making idealizing assumptions of determinism or zero probability.

A causalist, of course, is not at risk for the present dilemma as to which evidential relation (the one conditioned on T_1, or the one conditioned on D) to use in assessing whether there is a means-end relation. Because there is no causal link from my crossing the street, to the absence of dangerous traffic—there is a causal link between the two, but it points the other way—the causalist acknowleges no means-end connection. But the causal criterion is too strict, subjecting the causalist to fatalist resignation concerning some readily achievable goals, as in section 4's past-predicate example (where the causal link also points in the opposite direction of the means-end link). An agent that perceives a means-end link in the past-predicate scenario thereby achieves its past-predicate goals (and other, less whimsical goals, discussed below) better than does an agent using only a causal means-end criterion; but an agent that (mis)perceives a means-end link from crossing to no-dangerous-traffic does not thereby achieve its no-dangerous-traffic goal.[10]

If a causal link is unnecessary for there to be a means-end link, and an evidential link is insufficient, how then is a means-end link to be recognized? My proposal is that an evidential link—defined as above in terms of contrasting actual situations—works well as a presumed means-end link, but the presumption can be rebutted by certain considerations, as discussed in the following section. Informally, the rebutting intuition is easily stated in the street-crossing scenario: yes, there's certain to be no dangerous traffic whenever I do cross the street; but that's *only because* that inverse correlation between crossing and traffic is never tested (by crossing the street) in the circumstance where the correlation *would* be found not to hold (namely, in the presence of dangerous traffic). In the next section, I attempt to present more formally the foregoing intuition as to how an evidential link can be superseded, countering the default presumption that it corresponds to a means-end link.

6. Subjunctive means-end recognition

The choice machine in section 3 uses subjunctive assertions—means-end links—of unspecified origin. But consider a *deliberative* choice machine—one that assesses for itself the validity of proposed means-end connections. Due to the circularity noted in section 2, some kernel of the machine's means-end recognition must be built in; otherwise, even if the machine could reason well enough that it figured out that using a given means-end criterion would be advantageous, it would not thereby be influenced to use that criterion! This section sketches aspects of the design of a deliberative choice mechanism; the design effectively defines my proposed subjunctive criterion for means-end links, intermediate between evidential and causal criteria.

I begin by outlining an essentially evidentialist relation. Then, I introduce a further condition to try to limit the relations to choice-supporting subjunctive ones, i.e. means-end links.

Consider predicates such as the indexical assertions (i.e., assertions about a pointed-to situation) defined in section 5 (e.g., S is *I am standing at a street corner*...) defined now from the point of view of a deliberative choice machine. Let us use the notation

$$C_1 \ldots C_i \,:\, A_1 \ldots A_j \_ R_1 \ldots R_k \;\; (r)$$

to express the assertion: given context predicates $C_1 \ldots C_i$, that are satisfied in the current situation (i.e. that are true when applied to the current situation), if action conditions $A_1 \ldots A_j$ are also satisfied in that situation, there is probability r (the schema's *reliability*) that conditions $R_1 \ldots R_k$ are satisfied; i.e.,

$$P(R_1 \ldots R_k \mid C_1 \ldots C_i \, A_1 \ldots A_j) = r \; .$$

---

ble; 2) even with those probabilities slightly nonzero, the agent, never having actually encountered $\sim C_d C$ or SDT_1C, has no way to ascertain whether $P(T_2|SDT_1{\sim}C_dC)$ equals $P(T_2|SDT_1{\sim}C_d)$=1 or instead equals $P(T_2|SDC)$=0, unless the agent has already somehow solved the very problem under discussion; 3) in the case of a real-world computer simulating the zero-probability world, the screening-off might still give the wrong answer with regard to a means-end link—depending on the hardware-failure details, $P(T_2|SDT_1{\sim}C_dC)$ might actually be close to zero, with $P(T_2|SDT_1{\sim}C_d{\sim}C)$ still close to unity.

[10] Here I invoke meta-circular consistency as discussed in section 2: the past-predicate means-end construal is a means to the goal of achieving (some) past-predicates, but the no-traffic means-end (mis)construal is not an effective means to achieving the no-traffic goal. Meta-circular consistency is not definitive, because it is circular: to apply it, we need to know by what criteria (causal, subjunctive, evidential, or whatever) we can say that using a given means-end construal policy is a means to achieving one's goals. But by *any* of those three criteria, it is ineffective to perceive a means-end link from crossing to no-dangerous-traffic, so meta-circular consistency does give us some purchase on the problem.

As in the previous section, let us regard the probabilities simply as frequencies among actual situations; the choice machine can empirically verify or adjust a schema's specified reliability, assuming the machine encounters a large enough sample of relevant situations.[11] Following Drescher (1991) (inspired by Piaget 1952), we can call these conditional-probability assertions *schemas*. When the schema notation omits a designated reliability, assume the reliability is 1.

I do not address here how a choice machine might propose particular schemas for consideration in the first place (out of the exponentially many combinatorial possibilities), how it might define the predicates in terms of which its schemas are expressed, or how it might ascertain whether a given predicate is currently satisfied (but see Drescher 1991 for some suggestions). What concerns me here instead is: *given* that an agent somehow amasses knowledge about how the world *is*—evaluating the current truth of various predicates, and constructing schemas that tabulate some correlations among the predicated states' actual occurrences—how can the agent get *from there* to knowing how the world *would* be (in the choice-supporting sense) if this or that action were taken? Given a set of predicates and schemas, how does the agent's machinery then recognize means-end links? Accordingly, in this discussion, I just postulate the presence of whatever (accurately maintained) predicates and schemas are needed to illustrate the abilities and vulnerabilities of the proposed means-end-recognizing machinery.

A given schema is said to be currently *applicable* when its context conditions are all satisfied in the current situation. An applicable schema is said to be *activated* when its action conditions obtain as well. An applicable schema is currently *empirically overridden* if another currently-applicable schema asserts a different probability for the same result given the same action, conditioned on context predicates that designate a strictly more specific condition than the given schema's context (i.e., the first schema is always applicable too when the overriding schema is). The currently overriding schema's probability is then asserted in place of the overridden schema's probability. Thus if the choice machine has, say, schemas

$C_1 : A_1 \_ R_1$ (0.90)

$C_1C_2 : A_1 \_ R_1$ (0.14) (or equivalently, $C_1C_2 : A_1 \_ \sim R_1$ (0.86)),

and C_1, C_2 are both satisfied now, the first schema is currently empirically overridden by the second schema, which asserts just a 0.14 probability of R_1 if A_1. But if the choice machine did not have the second schema, the first schema would now assert a 0.90 probability of R_1 if A_1. Thus, the choice machinery effectively makes a default presumption that the probability expressed by a schema is conditionally independent of aspects of the world not designated in the schema's context. But that presumption can be overridden by another schema (with its own empirical support) that expresses a more-specific conditionalization.

Assume that if the choice machine has schema C:A_R (x), it also has a complementary schema C:~A_R (y). That is, the machinery keeps track of the result's probability both with and without the specified action (given the context).[12] Arbitrary conditions (not just personal motor events) can appear in the action part of a schema (and in the context or result). By the *composition* of schemas (discussed just below), the choice machinery can sometimes use other schemas to bring about the satisfaction of a given schema's designated action-conditions.

Each predicate has a *utility* attributed to it (positive, negative, or zero), allowing some predicates to serve as goals. If C:A_R (x) and C:~A_R (y) are currently applicable and not overridden, and R has utility u, the pair of schemas currently attribute to action A the attenuated utility $u(x\text{-}y)$ with respect to result R—a conventional expected-utility calculation.[13]

For each designated action, the machinery keeps track, from moment to moment, of the utilities currently attributed to that action by the then-applicable, non-overridden schemas that use the action. The action to which is currently assigned the greatest net attenuated utility is selected for activation. Thus, the mechanism is influenced to take an action from which there is a link—via a currently applicable pair of complementary schemas—to a state of positive utility (while avoiding actions that link to a state of negative utility).

By virtue of the foregoing provisions, the choice machinery treats schemas as expressing means-end links: if an applicable, non-overridden schema links from an action to a state that is more positively valued than the states linked to by the negation of that action, the schema influences the choice machine to take that action. A further provision allows schemas to *compose* together, such that if C and D are now true, the schemas

[11] Alternatively, a choice machine might ascertain some schema probabilities simply by being told what they are, or by other techniques that are indirectly grounded in actual observations. But for present purposes, a presumption of direct empirical sampling will suffice.

[12] In Drescher (1991), a given schema keeps track of both probabilities, combining what I here call two complementary schemas. The difference is just a change of terminology.

[13] The expected utility must also be scaled by the probability of the schema's applicability. In the examples here, I make the simplifying idealization that it is certain that a schema's context conditions are satisfied (or are not) at each given moment.

$C : A \_ B\ (x)$, $C : {\sim}A \_ B\ (y)$, $D : B \_ R\ (v)$, $D : {\sim}B \_ R\ (w)$ $(x>y, v>w)$,

assuming they are not currently overridden, combine to imply that action A is a means to achieving B, and the action of achieving B in turn achieves R.[14] Hence, given CD now, A's utility with respect to R now is $u(x\text{-}y)(v\text{-}w)$ (making the usual conditional-independence presumptions).

The machinery postulated so far is still within, or close to, the evidentialist paradigm—the machinery effectively presumes that a kind of evidential link (albeit comparing frequencies among different actual situations, rather than comparing subjective conditional probabilities in the current situation) is a means-end link. As such, the machinery stipulated so far is vulnerable to the street-crossing problem discussed in the previous section, given the idealizations proposed there.

To demonstrate that vulnerability, say the choice machine has the schemas

$*SD : C \_ {\sim}T_2$ (safely-disposed crossing)

$*SD : {\sim}C \_ T_2\ (0.3)$

with the various predicates defined as in section 5. The schemas assert that if the choice machine crosses the street, there is then no dangerous traffic; if it does not cross, there may be dangerous traffic. The asterisks are to denote that—intuitively, and by the criteria outlined below—the schemas are misleading if construed as expressing a means-end link from crossing, to the traffic condition, rather than just a correlation between the two. That is, although the conditional probabilities expressed by these schemas are correct as conditional probabilities—which address what *is* the case when the specified conditions are actually met—it is mistaken to presume that these are also the probabilities that there *would* be dangerous traffic if I *were* to cross now, or if I were not to cross (in the choice-supporting sense of *would*). Although C reliably gives evidence of ${\sim}T_2$ (given SD), C is not a means to achieving ${\sim}T_2$. But the choice machinery, as described so far, does presume a means-end link whenever schemas assert an evidential link.

Suppose S, D, and T_1 are true of the current situation. Thus, the above schemas' contexts are satisfied in the current situation, and (let us suppose) the choice machine has found extensive, exceptionless empirical support for these schemas. How could the choice machine know that the schemas nevertheless do not express a means-end link from action to result? That is, how could it know that crossing the street now—with oncoming dangerous traffic—would not achieve the absence of dangerous traffic at the moment of crossing?

Suppose further that the choice machine has the schemas

$ST_1 : C \_ T_2$ (conserved traffic)

$S{\sim}T_1 : C \_ {\sim}T_2$ (conserved non-traffic)

which assert that the dangerous traffic neither vanishes nor materializes when I cross. The conserved-traffic schema, in particular, contradicts the prediction made by the safely-disposed-crossing schema when both schemas are applicable (i.e., when SDT_1 is true). But the conserved-traffic schema does not meet the criteria for overriding the safely-disposed-crossing schema; it is not more-specifically conditionalized. Just as in section 5 (where we considered the evidential conflict between how my safely-disposed street-crossing always in fact proceeds—safely and without dangerous traffic—and how oncoming dangerous traffic always in fact proceeds—without suddenly vanishing), the choice machine needs some additional principle to resolve the conflict, to determine that the conserved-traffic schema is the one to trust—even though both conflicting schemas enjoy exceptionless empirical confirmation.[15]

We would like to be able to override the safely-disposed-crossing schema $*SD{:}C\_{\sim}T_2$ with one that says

$**SDT_1 : C \_ T_2$.

[14] Composing schemas together allows the choice machine to anticipate the outcome of an action in a perhaps previously unencountered situation (e.g. CD), even though each schema individually tabulates statistics over actually-encountered situations. Other such combinatorial techniques (e.g. in Drescher 1991) are also useful, but are not needed for the present analysis.

[15] In order to suppose that $ST_1{:}C \_ T_2$ has direct empirical support, we must imagine (unrealistically) that there are many sample occurrences where D is false—where I cross in front of dangerous traffic, without my usual safe disposition. Alternatively, Drescher (1991) suggests additional machinery by which a schema-based system could construct such a schema via more-general observations (say, observations concerning the street-crossing behavior of many other people, where some unsafe-crossing samples might realistically be found; or else, even more generally, observations concerning the physics of objects colliding at certain velocities). I use the unrealistic assumption of an available ${\sim}D$ sample as a stand-in to obviate a discussion here of the additional machinery in Drescher (1991). Using more-general observations would pose the same conflicting-evidence problem as here, requiring a similar resolution.

That is, although the action of (initiating) crossing the street (with safe disposition) ordinarily implies the absence of dangerous traffic passing at that moment (T_2), if crossing occurs under this strictly more specific condition—i.e., in the presence of dangerous oncoming traffic (T_1)—then the traffic T_2 is still present. Unfortunately, the probability expressed by this schema is undefined (as denoted by the double asterisk), because SDT_1C has zero probability. For the same reason, the choice machine can obtain no empirical support for this schema (C never occurs when SDT_1); indeed, as discussed in section 5, that empirical support would be lacking even in the case of an infinitesimal probability of SDT_1C (and there is contradictory empirical evidence based on different subsets of the conditions SDT_1C, namely SDC—given which T_2 never actually occurs—and ST_1—given which T_2 always actually occurs), calling into question the basis of any "subjective probability" conditioned on SDT_1C.

There is, however, a plausible basis for trusting the conserved-traffic schema ST_1**:**C_T_2 over the safely-disposed-crossing schema *SD**:**$C\_{\sim}T_2$. A key observation is that the conserved-non-traffic schema $S{\sim}T_1$**:**$C\_{\sim}T_2$ is both *more general than* and *explanatory of* the safely-disposed-crossing schema, in the following sense:

- A given schema is *more general* than another if they share the same action conditions, and if the given schema has been activated (i.e., its action conditions have obtained when its context conditions obtain) in a strictly wider set of circumstances than the other schema has been activated in. Here, the conserved-non-traffic schema is more general than the safely-disposed-crossing schema because the latter is applicable only when D (hence activated only when D);[16] and although the former schema is applicable only when $\sim T_1$, nonetheless the latter schema too has never been (indeed cannot be) activated except when $\sim T_1$ (i.e., SDT_1C is an impossibility, under the proposed idealizations).
- A given schema is *explanatory of* another if it predicts the same result of the same action with the same reliability.

Intuitively, when a given schema is explained by a more general schema in the foregoing sense, an Occam's-razor presumption suggests that the given schema is just a consequence of the explanatory schema, owing its apparent validity to the explanatory schema, and so should not be insisted on as expressing an independent principle. The explained schema should just defer to the explanatory schema, and so should be considered inapplicable, even though its context is satisfied.[17] Call this provision an *explanatory deferral* of the explained schema. I propose that the explanatory deferral, too, be built into the choice machinery, revising the earlier definition of a schema's applicability.

As a consequence of the explanatory deferral, the explained schema *SD**:**$C\_{\sim}T_2$ does not contribute (as it otherwise would) to the calculation of C's utility with regard to $\sim T_2$ when SD. Instead, that explained schema defers to the explanatory schema $S{\sim}T_1$**:**$C\_{\sim}T_2$. But when T_1 is true, that explanatory schema too is inapplicable, and so the conflicting applicable schema ST_1**:**C_T_2 prevails instead. Thus, even though the explained schema *SD**:**$C\_{\sim}T_2$ is exceptionless (given our zero-probability idealization), the machinery does not construe it as expressing a means-end link.

The explanatory deferral tugs in a different direction than the empirical override, giving priority to more-general schemas rather than more-specific. Their rationales are complementary. When a schema has in fact been activated in a specific circumstance (enough times to obtain a significant sample), an expectation of the then-observed outcome takes precedence (in future occurrences of that circumstance) over a conflicting prediction as to what is expected to occur more generally; hence, the empirical override. The explanatory-deferral principle, in contrast, addresses schemas that are in *agreement* as to what is expected to occur. But the deferral shifts the context for that expectation to that of the more-general explanatory schema. Given a specific circumstance in which an explained schema has *not* been activated (e.g., I have never stepped in front of clearly-seen dangerous oncoming traffic while having the contrary desire and competence)—so there is no basis for an empirical override regarding that specific circumstance—the deferral of an otherwise-applicable schema to a currently-inapplicable explanatory schema allows a conflicting applicable schema to prevail, thus effectively adjudicating between the two reliable (even exceptionless) conflicting schemas.

As already noted, the safely-disposed-crossing schema *SD**:**$C\_{\sim}T_2$ is impeccable as an expression of conditional probability. On the idealizing assumptions above, 100% of actual SDC situations also exhibit $\sim T_2$; the action C of

[16] As per note 15 above, I am imagining that D has sometimes been false (though I stipulate it is known to be true *now*), giving the choice machine an empirical basis for ST_1**:**C_T_2.

[17] A more advanced version of this principle would allow explanation jointly by multiple more-general schemas, each of which has empirical support—for example, A_1**:**$B_1\_C_1$ and A_2**:**$B_2\_C_2$—to explain (via a presumption of conditional independence) a schema—for example, A_1A_2**:**$B_1B_2\_C_1C_2$—for which there may be no direct empirical support (B_1B_2 might never have been observed to occur when A_1A_2, for example). But the present simpler principle suffices for the examples here.

(initiating) crossing the street given *SD* is indeed perfect evidence for the absence of dangerous passing traffic. No adjustment or override of the conditional probability per se is warranted, nor of the expected utility defined by the product of that conditional probability and any utility ascribed to the schema's result condition $\sim T_2$. Nonetheless, it would be nonsensical, given T_1, to use that expected utility to assess the desirability of choosing action *C*—i.e., to treat the evidential relation between *C* and $\sim T_2$ (although perfectly valid as such) as a means-end relation.[18] The explanatory deferral suppresses that means-end construal, countering the default presumption that a schema's evidential link also corresponds to a means-end link, thus addressing the evidentialist dilemma discussed in section 5.[19]

The explanatory deferral thus enables the choice machine to distinguish, in some situations, between the conditional probability

P(result | context & action) = P(result & context & action) / P(context & action)

—the probability that the result conditions are the case given that the context and action conditions actually are—and the *subjunctive* (or *modal* or *counterfactual*) probability

P(result \ action | context) ,

where P(*A**B*|*C*) is the probability that *A would* be the case now if *B* were (in the choice-supporting sense of *would*), given that *C* actually is the case. (Similarly, e.g. Pearl 2000 distinguishes P(*A*|*BC*) from the subjunctive P(*A*|*do*(*B*),*C*), but Pearl's version only refers to what *B* would *cause*. Similarly too with Gibbard and Harper 1977.)

The machinery proposed here does not compute subjunctive probabilities as such. Rather, each schema keeps track only of an associated conditional probability. Whenever the schema's context is satisfied and the schema is not overridden, the associated conditional probability is presumed also to be the subjunctive probability, except when an explanatory deferral defeats that presumption (at which times the subjunctive probability is obtained instead from other, non-deferred schemas, if available). In lieu of proposing a mathematical expression for the value of a subjunctive probability, I am sketching a presumption-and-deferral mechanism for finding (some) subjunctive-probability values.

The central challenge of subjunctive reasoning is to find a principled way to determine what actually-true propositions to "hold fixed", and what propositions to "let vary" instead for consistency with a counterfactual antecedent. E.g., when contemplating the antecedent *C* given DT_1, do we ask what must be true assuming *D* stays the same as it actually is, or assuming instead that T_1 stays the same? The present proposal reduces that challenge to a problem of adjudicating among conflicting inductive projections (e.g. those given by the safely-disposed-crossing schema and the conserved-traffic-schema), a problem addressed by the explanatory deferral.[20] Some such adjudication is necessarily a fundamental cognitive ability, even apart from subjunctive reasoning.

---

[18] Allais (1979) calls attention to other ways that people's decision intuitions diverge from what the maximization of expected utility would dictate. But that divergence involves situations with substantial uncertainty; Allais shows that people often place a premium on the predictability of a good outcome, even at the cost of a lower expected utility. In contrast, the present distinction contradicts expected-utility based decisions even when uncertainty is negligible (or zero). The present decision approach still selects an action based on the product of a conditional probability and a utility value (as does an expected-utility calculation), but substituting a subjunctive probability for a conventional conditional probability in that calculation.

[19] Recall the running-program variation of this scenario, discussed in section 5. Depending on the hardware-failure mode, it could be that P(T_2|SDT_1C)=0.001, in which case there could be a schema

$*SDT_1 : C\_\sim T_2$ (0.999)

that empirically overrides both $*SD{:}C\_\sim T_2$ and $ST_1{:}C\_T_2$, but unhelpfully: the overriding schema still asserts a high probability for $\sim T_2$-if-*C* (and correctly so for predictive purposes, though not for means-end purposes). But suppose the machinery also has some predicates and schemas that pertain to the possibility of hardware failure, namely

F := (Failure) The simulation's hardware fails to execute the program properly.

$FST_1 : C\_\sim T_2$ (0.999) , $*\sim FSD : C\_\sim T_2$, $\sim FST_1 : C\_T_2$, $\sim FS\sim T_1 : C\_\sim T_2$.

The schema $FST_1{:}C\_\sim T_2$ (0.999) is more general than and explanatory of $*SDT_1{:}C\_\sim T_2$ (0.999) (note that the latter schema is never activated except when *F*), so the latter schema defers. But the explanatory schema is itself inapplicable when $\sim F$. Presuming $\sim F$ to be almost certain, the explanatory schema is currently inapplicable, allowing the conflicting applicable schema $\sim FST_1{:}C\_T_2$ to prevail. (Then, among the schemas with $\sim F$ in their contexts, the explanatory deferral works as in the previous discussion, allowing $\sim FST_1{:}C\_T_2$ to prevail over $*\sim FSD{:}C\_\sim T_2$ just as $ST_1{:}C\_T_2$ prevails over $*SD{:}C\_\sim T_2$.)

[20] A competing approach to analyzing subjunctive or counterfactual inference appeals to *possible worlds* (e.g. Lewis 1973). By that approach, what *would* be the case if I were to cross the street now (as dangerous traffic plainly approaches, and in fact I do not cross) is whatever *is* the case in imaginary possible worlds in which I *do* cross the street now, but which, given that difference from the actual world, are otherwise as similar as possible to the actual world. But everyday intuitive measures of similarity give wrong answers (see Fine 1975). For instance, a possible world in which there is a momentary lull in traffic right now (so that I now cross) would thereby differ from the actual world in a much more ordinary way than would a possible world where I cross now, despite the danger and my safe disposition, because of a bizarre failure of

Clearly, the explanatory-deferral principle applies as well to a wide class of everyday problems with the same structure as the street-crossing problem—including problems that would have been ubiquitous during our ancestors' evolution. I speculate that some such principle built into our choice machinery[21] is what makes it "intuitively obvious" to us that my crossing the street, although perfectly evidential of the absence of dangerous traffic, *would* not yield that condition right now (as the traffic approaches), and thus does not now serve as a means to that end.

Plausibly, causal regularities correspond to the most widely applicable schemas we can find, because we inhabit a universe where a small number of such regularities intercombine to specify all that occurs. The lower the level of abstraction, the more widely applicable are the regularities (e.g., principles about the behavior of gears and levers apply more widely than principles about the black-box behavior of a machine made up of some specific arrangement of gears and levers, because the components individually occur far more widely than in some particular combination; similarly for molecules, atoms, quarks, etc.). The explanatory deferral thus promotes a reductionist presumption that the (more widely applicable) regularities of the constituent parts of an object determine the expectation of the object's behavior in circumstances where that behavior has not been tested directly.

But it is not only causal regularities that the explanatory-deferral principle allows to be recognized as means-end links. Define the predicates

> *H* := (Hand) I lift up my hand now.
>
> *R* := (Raised) My hand is raised.
>
> *P* := (Past) The state of the universe 1 billion years ago is such that, according to physics, my hand is raised 1 billion years thence.

and suppose that the choice machinery includes the (empty-context) schemas

> **:** *H* _ *R*
>
> **:** ~*H* _ ~*R*
>
> **:** *H* _ *P*
>
> **:** ~*H* _ ~*P* .

Since *R* is true exactly when *P* is true (presuming determinism), empirical support for the first pair of schemas will be identical to support for the second pair. No circumstances will provide an override or deferral of the second pair without doing so for the first pair. The choice machinery will recognize a means-end link to *P* to the same extent as to *R* (and appropriately so, as argued in section 4) even though the former link is acausal.[22]

---

my street-crossing competence. Construing the more-ordinary difference as "smaller" leads to the wrong conclusion (at least with regard to the choice-supporting sense of *would*): that if I were to cross now, the dangerous traffic would be absent. Alternative similarity criteria proposed by Lewis and others hinge in some way on the physical extent of the differences just prior to the action. Altering the current traffic is a physically bulkier change than tweaking a few of my neurons to make me cross despite the danger; in that sense, the former difference from the actual world is indeed larger than the latter. But the relative physical bulk of the two changes is an inessential feature of the scenario; it is easy to contrive situations of the same structure where the traffic-equivalent is physically smaller than the choice machinery.

[21] If it were not built in, we might still reason (as above) that the principle is a good one to use. That reasoning would conflict with, but not necessarily prevail over, the influence of, say, the misleading street-crossing schema that tells us it is safe to cross no matter what (assuming that our choice machinery is generally schema-like in the first place).

[22] The explanatory deferral can be seen as a simplification of the preference for minimal latent structures in Pearl (2000); that preference too can resolve the evidential ambiguity as to what would happen if I were to cross despite the dangerous traffic. Pearl's inference of causal models from Bayes nets does not apply to problems in which full determinism is presumed (because some requisite conditional probabilities become undefined). If, however, we allow an infinitesimal probability that *P* does not lead to *H* and that *H* does not lead to *R*, then (even if such deviations are so improbable that they never actually occur in the lifetime of the universe) Pearl's approach is applicable, and it can show correctly that *P* causes *H* and *H* causes *R*, but *H* does not cause *P*. But then, an agent using only causal links as means-end links would needlessly forfeit the opportunity to have *P* be true if *P* is a goal (and similarly in Newcomb's Problem, as discussed below).

Pearl (p. 108) distinguishes an *act* ("a consequence of an agent's beliefs, disposition, and environmental inputs") from an *action* ("an option of choice in contemplated decision making, usually involving comparison of consequences"). Pearl models the latter as uncaused; if represented as a node in a causal structure, an action has no parent nodes (p. 71). An action is a "free choice" (p. 109) or "surgical intervention" (an externally originated change to the causal structure itself) with respect to which Pearl calculates the subjunctive probability of an outcome (rather than using the conditional probability of an outcome given an act) to assess a contemplated action's utility.

But the important distinction between choice-supporting subjunctive probability and evidentialist conditional probability does not require Pearl's further distinction between act and action. In fact, a choice is both of those, and can be so modeled for decision-making purposes. The difference is both philosophical—in effect, Pearl's formalism models free will rather than mechanical choice—and practical, in that Pearl's formalism fails to recognize what I argue are (in some situations) valid acausal means-end links.

The recognition of causal relationships as such is quite plausibly a far more sophisticated task than the recognition of subjunctive relations generally, as performed by something like the machinery sketched here. If so, there is no reason to expect that our built-in choice machinery is (or should be) designed to treat only causal links as means-end links. An exclusion of acausal links would be hard to implement (because specifically causal relations are hard to recognize as such) and would be of no benefit; indeed, the exclusion would only impair an agent's ability to pursue some of its goals—both whimsical goals like the hand-raising past-predicate, and (as argued below) more important goals in Newcomb's Problem (esoterically) and in Prisoner's Dilemma situations (routinely and importantly).[23]

I propose, then, that the choice-supporting sense of *would* just *is* (something like) the relation given by a schema-based correlation that is not currently overridden or deferred.[24] That is what (I claim) would-ness, in the choice-supporting sense, turns out to consist of; that is what a means-end relation turns out to consist of. As the street-crossing example illustrates, an agent that was designed (or that evolved) to use that subjunctive criterion of means-end relations would thereby fare well in achieving its goals, compared to an agent that used a purely evidential criterion—thus passing the meta-circular-consistency test with regard to that comparison. And as the hand-raising example illustrates, going to the extra trouble of distinguishing specifically causal relations, and insisting on their use alone as means-end links, would not improve an agent's goal pursuit, and indeed would sometimes hamper it, compared to using the subjunctive criterion; so the subjunctive criterion passes the meta-circular consistency test with regard to that comparison as well.

In sum, with a deterministic universe, no outcome ever *changes* (from what it was already set to be); all is inalterable, just sitting statically in space-time. With sufficient foreknowledge—often achievable, as in the (idealized or artificial-world) street-crossing scenario—there can be no defined probabilities conditioned on an alternative to the actual action in our specific situation, and so the usual evidentialist contrast is impossible (or even if—foregoing the zero-probability idealization—such conditional probabilities can be defined, we still need a non-arbitrary, non-circular basis for calculating those probabilities in the face of conflicting evidence from different partial matches to the specific situation).

Even though an action cannot change an outcome, we can draw a contrast between actual situations where the action occurs, and other actual situations where it does not. Correlations found in such contrasts support a (somewhat) evidentialist presumption of a means-end link. Sometimes, as in the past-predicate example, the means-end link is acausal, but still correct—an agent that fails to recognize and act on the means-end link foregoes achievement of its goal, a needless fatalist resignation. But often, as in the street-crossing example, a mere correlation—even an exceptionless one—is not a means-end link; an agent misconstruing it as such would not achieve its goal either. Inalterability does not imply futility, nor does acausality; but conversely, mere evidence does not ensure efficacy, and the explanatory-deferral principle says why, providing a way to selectively defeat the default evidentialist presumption.

There is yet another dilemma, not yet addressed by the explanatory-deferral principle, that challenges the reconciliation of choice and determinism, in both mundane and esoteric situations. This dilemma comes up in the discussion of the transparent-boxes variation of Newcomb's Problem in section 8. But the machinery presented so far suffices to address first the conventional, opaque-box version of Newcomb's Problem.

7. Newcomb's Problem

Suppose we implement Newcomb's Problem as follows. Awhile ago, a mischievous benefactor, preparing to present the two boxes to you (as in section 1), took a detailed snapshot of the nearby state of the universe, and then used that snapshot to run a simulation of the subsequent unfolding of events, up to and including your forthcoming choice. If the simulation showed you choosing both boxes, the benefactor put nothing in the opaque box; if it showed you choosing the opaque box alone, the benefactor put $1M in that box. As usual, let us make the idealization that the universe is deterministic enough, and predictable enough even in practice, to carry out the simulation with perfect reliability (or, recast the scenario in an artificial world where those idealizations actually hold). Let us also consider an alternative situation in which the simulation is fallible, such that, say,

[23] There are, however, specific aspects of causal links whose recognition is plausibly hardwired. For example, well-known experiments show young infants are (at some level) aware that an object's response to another object's motion requires physical contact at the time of the response (e.g. Flavell and Markman 1983). But such perceptually based special-case criteria are distinct from a general basis for ascertaing causal relations among phenomena that may be novel, abstract, or widely dispersed in space and time.

[24] The parenthetical qualification refers to three hedges. First, section 8 below proposes an additional principle that adjusts what gets construed as a means-end link. Second, as mentioned above in note 17, a choice machine would benefit from a more advanced version of explanatory deferral whereby several more-general schemas can combine to be explanatory of another schema.. Third, the present proposal is preliminary and tentative; even if it proves to be a step in the right direction, refinements are no doubt needed.

$$P(M \mid NP) = P(\sim M \mid N\sim P) = x < 1,$$

where the indexical assertions

N := (Newcomb) I am now presented with a Newcomb's Problem choice.

P := (Past) The past state of the universe at the time of the snapshot is such that (according to physics) I will take only the opaque box, forfeiting the $1,000 transparent box.

M := (Million) The opaque box in front of me contains $1M.

are defined from your point of view. For the sake of argument, assume money has linear utility, and assume that the expected monetary payoff is the only relevant goal here.

The box's opaqueness serves a critical technical function. If the box were transparent, then its content could affect the outcome of your choice process. As in the discussion in section 3, the simulation would therefore need to take into account the content of the box in order to accurately simulate your choice process; but the simulation doesn't know what the content of the box will be until the simulation has completed. However, because the box is opaque—and assuming more generally that until you make your choice, you are sufficiently insulated from any effect caused by the content of the box or by the process or outcome of the simulation—it is possible to conduct the simulation in such a way as to leave the content of the box unspecified, and still be able to simulate what goes on outside the box, and in particular what goes on in your choice process.

Making the case *for* taking only the opaque box—even with a fallible simulation—is relatively easy; the challenge is to rebut the arguments *against* the one-box choice. The argument for taking just the one box is that if you were to do so, you would (probably or certainly, depending on how reliable a simulator we postulate) obtain $1M in the opaque box; if you were to take both boxes, you would (probably or certainly) obtain nothing in the opaque box. Accordingly, using the same calculation that we would use if your choice *caused* the opaque-box content to change, we find that someone who would take both boxes has a much lower expected gain from the encounter than someone who would take just the opaque box—even if the simulation's reliability is only, say, 0.9, or even 0.6 or less. (The intuitive appeal of taking just the opaque box is boosted dramatically—especially for a perfect simulation—if you imagine a series of practice trials, using play money, in which you vary your choice at will and always then find the corresponding content in the opaque box.)

Thus, one-box choosers do get a better payoff than two-box choosers (or, in the probabilistic case, they get a better payoff on average). Of course, this consideration is not definitive. A skeptic might maintain (as do Gibbard and Harper 1977, for example) that taking both boxes is the rational choice, and if one-box choosers fare better than two-box choosers, it is only because the situation has been rigged in advance to reward the former's (predicted) irrationality (much as if a written exam were perversely graded so as to reward mistaken answers). But this position is suspect unless (unlike in the present situation) the chooser is unaware of the rigging, and thus unable to take it into account when choosing (as the exam-taker could do if she knew the grading scheme). Otherwise, a committed fatalist in a deterministic universe could maintain that the entirety of space-time is effectively a grand Newcomb's box whose content—including the outcome of any choice—is already inalterably sealed-in. Those who insist (irrationally, says the fatalist) on making choices in pursuit of goals do fare much better, on average, than those who succumb to fatalist resignation; but (concludes the fatalist) that contrast just shows that the universe's rules are rigged in advance—not necessarily by any deliberate design—in a manner that rewards the irrational.

As discussed in section 2, means-end criteria cannot be deduced without some built-in starting point, so the fatalist cannot be definitively refuted. Still, as argued in the previous sections, if we put aside complete fatalist skepticism, acknowledging instead that there is meaningful choice, even given determinism, then the door is opened to a means-end link to an already-inalterable result. And section 4 presented an acausal link—from the action of raising my hand, to the satisfaction of the hand-raising past-predicate—that serves as a means-end link to that past state, to the same extent that there is also a means-end link to the (also already-inalterable, given determinism) state of my hand being elevated a moment from now. Hence, neither a result's inalterability, nor the acausality of a link to that result, is necessarily disqualifying. And indeed, Newcomb's problem harnesses a past-predicate link (to the snapshot-time state of the universe) quite similar to the hand-raising past predicate, followed by an ordinary causal link from the past state to the box content (via the benefactor's simulation etc.).

A more difficult challenge to the one-box choice stems from the question of what a means-end link does consist of, if it is not just a causal link. Some versions of the one-box argument merely appeal to what would *have to* be true of the opaque box content (or, for a fallible predictor, what would *likely* be true) if the one-box choice were made,

compared to if the two-box choice were made, given the way the opaque box content was set up. But similarly, in the street-crossing scenario, the absence of dangerous traffic would *have to* be the case if I were to cross now, given my safe disposition. As discussed in sections 5 and 6, that criterion is merely evidential, and does not suffice to establish a means-end link.

The previous section's discussion of the street-crossing scenario—and the subjunctive means-end criterion implemented by schemas with explanatory deferrals—addresses the challenge by replying that not just any acausal correlation establishes a means-end link. Let us define

B := I choose both boxes now.

N_{100} := The simulator is 100% reliable.

N_{99} := The simulator is 99% reliable.

in addition to N, P and M above, and suppose the choice machinery has the schemas

$N : B \_ \sim P, \quad N : \sim B \_ P,$

$NN_{100} : P \_ M, \quad NN_{100} : \sim P \_ \sim M, \quad NN_{99} : P \_ M\,(.99), \quad NN_{99} : \sim P \_ \sim M\,(.99).$

The first pair of schemas corresponds to a past-predicate link similar to the hand-raising example: iff you were to take just the opaque box, the snapshot-time past state would be such that (according to physics) you will so choose. The second or third pair of schemas (whichever pair is applicable, supposing that either N_{100} or N_{99} is true) corresponds to an ordinary causal link from the past-predicate state to the box content: the past state causes the snapshot to cause the simulation to cause the opaque box to contain \$1M. As argued in section 6, causal links are especially general (the more so the lower-level they are), hence resistant to explanatory deferral; and similarly for links to result conditions that are coextensive with causal result conditions (e.g. the corresponding past-predicates). In the absence of a deferral of the above schemas (or of more-specific overriding conditions), they compose together (recall the composition provision in section 6) to establish, by the proposed subjunctive criterion, a means-end link from the one-box choice, to having \$1M in the opaque box: if you were to choose just one box, there would (more likely) be \$1M in the opaque box, in the choice-supporting sense of *would*, despite the lack of a causal link from your choice to the box content. The link is not merely evidential; by the proposed criterion, the means-end link, although acausal, does not exist merely by virtue of a correlation, but also by virtue of the absence of (currently inapplicable) more-general explanatory schemas for the above past-predicate link or the ordinary causal link.

Newcomb's Problem and the street-crossing scenario have similar structure in terms of both the causal and the probabilistic dependencies among their respective states; in both scenarios, there is a common causal influence (P or $\sim T_1$) upon both a choice and a goal state, creating a correlation between the choice and goal. What distinguishes the scenarios, by the present account, is the comparative generality of the conflicting schemas in each scenario. The explanatory deferral, appealing to that distinction, permits an acausal evidential link to serve as a means-end link in Newcomb's Problem, while preventing a similar construal in the street-crossing scenario.

Another, even deeper challenge arises from a suggestion by Nozick (1969) to boost the *dominance* argument for the two-box choice—the argument that the box already contains either \$1M or nothing, and either way, you do better (by \$1,000) if you take both boxes than if not.[25] Suppose a video hookup permits a friend of yours to peer into the opaque box while you contemplate your choice. Your friend is remotely located and cannot in any way communicate with you until you have chosen. Your friend, who has been briefed about the situation, sees either \$1M, or nothing, in the opaque box. Either way, your friend concludes that your more lucrative choice is to take both boxes, thereby gaining \$1,000 in addition to the \$1M (or in addition to the \$0, as the case may be).

Your own expectation as to the box content is contingent on your expectation as to your choice, but your friend's perception of the box content is unconditional. Your friend not only has strictly more knowledge than you of the situation—she also has *all* the knowledge of the situation that is needed to answer the question at hand: which is the more lucrative choice for you to make? (I.e., which is the choice for which the payoff would be greater than if the other choice were made?) She might realize that, from the perspective of your more limited knowledge, the other

[25] Nozick (1969) endorsed the one-box choice only in the case of an infallible predictor, on the grounds that the dominance argument does not apply if only one possible choice is consistent with the (albeit yet-unknown to you) prediction. But with even an infinitesimal chance of simulator error, Nozick saw no way to counter the dominance argument, though he acknowledged his discomfort with that dependence on the difference between a zero probability and an arbitrarily small one. Later, Nozick (1981) revised his position to advocate using a weighted sum of the apparently intractably conflicting decision strategies.

choice might reasonably *seem* more lucrative; but she, from her perspective, knows (if she is reasoning correctly) which choice *is* the more lucrative.

The problem, though, is that you *do* know—if you recapitulate the above reasoning—that your friend, if she is reasoning correctly, must have concluded that taking both boxes would be your more lucrative choice. You need not know what she sees in the box to know that (if she is reasoning correctly) she has reached that conclusion—because you know she would reach it *regardless* of what she sees in the box. But if you know that she has all information needed to identify the more lucrative choice, and you know what her conclusion must be if she is reasoning correctly from that information, you thereby know what the correct conclusion is in fact—i.e., that taking both boxes is more lucrative than taking just the opaque box.

There is one way to escape the force of that argument. The one-box choice can be correct only if, contrary to the foregoing, your friend would *not* be reasoning correctly, given what she sees in the box, to conclude that taking both boxes would actually be your more lucrative choice. But in that case, we may as well let the opaque box be transparent instead, along with the $1,000 box. If (somehow) your friend would be correct to conclude that the one-box choice would be more lucrative for you, despite her seeing what is in the box, then you yourself would (somehow) be correct to reach the same conclusion, even if the box content were visible to you as well.

I accept this reduction of the opaque-box version of the problem to the transparent-boxes version. (Indeed, if, as I maintain, the one-box choice is correct in the opaque-box version of the problem, then—as Gibbard and Harper 1977 point out—in principle you can understand the problem well enough to be certain that you will so choose, and—in the case of a reliable simulator—you can thus know in advance that the box contains $1M, even though the box is opaque.[26]) I argue in the next section that—as counterintuitive as it seems at first—the one-box choice remains correct in the transparent-boxes case. I argue that inalterability does not imply futility even when the already-inalterable state is also already visible, and that the mistaken contrary intuition leads as well to the perceived incompatibility of choice and determinism.

8. Newcomb's Problem with transparent boxes

Making both boxes transparent in Newcomb's Problem poses an immediate technical difficulty—the infinite regress discussed in section 3. If a prediction of your choice will be communicated to you before you make the choice (in this case, communicated via the now-visible box content), then the prediction cannot necessarily be made in the first place. For in the course of the simulation, the predictor reaches the point where the prediction is conveyed to you, and is unable to proceed, because the prediction itself is still pending.

But this technical problem admits of a technical solution. The simulation can just presume that the prediction will be to take only one box; accordingly, the simulation shows $1M in a (now also transparent) large box. The simulation proceeds, and if it then shows you taking just the $1M box, the one-box prediction is made, and the $1M large box is presented to you in reality (along with $1,000 in a small transparent box). If the simulation instead shows you taking both boxes, then an empty large transparent box is presented to you (along with the $1,000 box). Thus, in this version of the problem, we give up on requiring an *empty* large box, if presented to you, to be predictive of the choice you then make (but see note 32).

Consider, first, the case of a perfectly reliable simulation. Suppose you find yourself presented with $1M in the large box. Is there any conceivable reason then for you to choose to take that box alone, forfeiting the extra $1,000? I claim there is the same reason as in the opaque-box case: iff you were to take both boxes, the simulation would have so predicted; and iff the simulation had so predicted, there would be no money in the large box.

Under the circumstances, the large box's content already tells you what you are about to do, which may seem to imply that no choice remains for you to make. On the contrary, though, as in the deterministic street-crossing sce-

[26] Along these lines, Eells (1982) and Jeffrey (1983) provide an evidentialist argument for taking *both* boxes in Newcomb's Problem (the opaque-box version) by appeal to screening-off via knowledge of one's decision just prior to acting (recall note 9 above). First, they assume that the agent's final decision to take just one box might (rarely) result in the action of taking both, or vice versa (thus averting the zero-denominator problem that would otherwise prevent the requisite conditional probabilities from being defined). Then, they argue, given a decision to take both boxes, the opaque box is already known (almost certainly) to be empty, and if it is empty, it must remain so whether or not the action accords with that decision. Or, given a decision to take just the large box, the opaque box is known (almost certainly) to hold $1M, and again does not change regardless of the action itself. Thus, knowledge of either decision (just prior to the action itself) screens off knowledge of the box content from knowledge of the action, just as though the box were transparent—given knowledge of the decision (and hence of the box content), the action itself provides no (further) evidence for the box content, so there is no evidentialist link between the action and the content. But as argued in section 5 and note 9, evidentialism (even with screening-off by knowledge of the decision prior to acting) gives the wrong answer even in some mundane situations. (Screening-off by foreknowledge of the decision also leads to the fatalist street-crossing problem discussed below.) If evidentialism is thus unfounded, its prescription in Newcomb's Problem is moot.

nario, where your choice is also a foregone conclusion, the mechanical choice process itself continues to operate, and crucially so: you compare what would be the case if you took one action or another, and act according to which situation you prefer. The box content may inform you in advance of your choice (as does the oncoming traffic in the street-crossing scenario), but the box content does not somehow override your choice process to impose the announced choice upon you; you still just choose whichever action you decide is best. So you must still figure out which choice *is* the best (with respect to the presumed goal of maximizing the expected payoff).

The recommendation that it is best here to take just the large box, even though that box is transparent, is so violently counterintuitive as to make the original formulation of the problem seem almost banal by comparison.[27] Evidentialists and causalists alike will of course take both boxes here. Let us examine what makes this recommendation seem so bizarre, and why (I claim) the intuitions it counters are mistaken. The same problem, I believe, arises whenever there is foreknowledge of a choice and its outcome, even in mundane situations. Accordingly, I digress back to the street-crossing scenario before returning to the present problem.

Foreknowledge of the already-determined choice, and of its outcome, presents a difficulty beyond the non-means-end correlation problem addressed in section 6 (which was addressed by appeal to the explanatory-deferral principle). This additional difficulty is especially conspicuous in the transparent-boxes variation of Newcomb's Problem, but it appears as well in mundane situations like street-crossing, presuming determinism. The extra difficulty, I believe, goes to the heart of the fatalist intuition that choice and determinism are incompatible. The further difficulty is as follows.

Returning to the street-crossing scenario, suppose the choice machine also has the schemas

$S{\sim}T_1 : C \_ O{\sim}H$ (successful crossing)

$S{\sim}T_1 : {\sim}C \_ {\sim}O{\sim}H$ (non-crossing)

$*S{\sim}H_1 : C \_ O{\sim}H$ (fatalist)

where we define

> H_1 := The state of the world now (including me) is such that according to physics, I am not hit by traffic in a moment.

and say the choice machine also has

$: H_1 \_ H$, $: H \_ H_1$, $: {\sim}H_1 \_ {\sim}H$, $: {\sim}H \_ {\sim}H_1$.

Each of the above schemas expresses an exceptionless relation between its action and result, given its context. The fatalist schema, however, turns out to be misleading. That schema asserts that if I cross the street in a situation where I am not in fact about get hit by traffic, then I reach the other side—and am not hit by traffic. I thus achieve both my primary goal (not being hit) and my secondary goal (getting to the other side). The combined utility of the two result conditions of this currently applicable schema exceeds the utility of the action ~*C* (which achieves just the primary goal), so *C* is seemingly the preferable action, given $S{\sim}H_1$. And in fact, we *are* given $S{\sim}H_1$. That is, under the operative idealizations, I already know—even with dangerous traffic oncoming now—that the event of my getting hit by traffic in the next moment is not in fact present in the space-time of the actual universe (because I already know that I will not cross now in the presence of such traffic). And the actual absence of that event—regardless of the *reason* for its absence—is all that ~*H* or $\sim H_1$ asserts, so the context $S{\sim}H_1$ is indeed satis-

[27] Gibbard and Harper (1977) mention a transparent-boxes variation (though with no discussion of a simulation-based prediction, or the infinite-regress problem and its solution), but advocate taking both boxes.
Blackburn (1998) discusses a problem by Kavka (1983), the *toxin puzzle*—which, as Blackburn notes, is structurally close to the transparent-boxes problem, with a large reward bestowed iff you are predicted to forego a small reward—and advocates the equivalent of the one-box choice. But in Kavka's variant, the predictor tells you what the game is prior to making the prediction, and then reads your mind accurately enough to predict what your eventual choice will be, offering or withholding the large reward accordingly. Blackburn argues that you should "cultivate the disposition" to forego the small reward (for the benefit of that disposition's conventionally causal influence on the still-future mind-reading), and to not change your mind even after the large reward is obtained (or else, by assumption, your resolve would not have been firm enough to convince the predictor). Still, if the choice process operates as always when it comes time to take or forfeit the small reward—that is, by comparing what would be the case if one action or another were taken, and effecting the action whose consequence is preferred—Blackburn does not, I think, explain why you should then choose to abide by your earlier resolution. Moreover, in Newcomb's Problem, unlike in Kavka's variant, the entire problem is presented to you not in advance of the "mind-reading", but only afterward when your large reward has already been secured. (The Newcomb's-Problem snapshot and simulation might even have been conducted before you were born.) By the time you first confront the problem, your then-cultivated disposition can exert no causal influence over your large reward; the large reward has already been established before you make any resolution.

fied, and I know it. Thus, if I were to (mis)construe this schema's relation between action and result as a means-end connection now, a gravely incorrect action would follow.

Moreover, let us assume that there is a miniscule physical possibility of crossing successfully and without collision even in the presence of dangerous traffic. If I cross the street now, in front of the onrushing traffic, the result condition of the (actually-exceptionless) fatalist schema $*S{\sim}H_1{:}C\_O{\sim}H$ could then follow without contradiction—it is (barely) possible that I'd indeed escape collision. But it would still be a grave error to count on that result and to act accordingly, since the result would be extremely unlikely.

Unlike the misleading schema $*SD{:}C\_{\sim}T_2$ in section 5, the fatalist schema $*S{\sim}H_1{:}C\_O{\sim}H$ is not explained by a more-general schema that is currently inapplicable; hence, the fatalist schema is not subject to explanatory deferral. In particular, the fatalist schema does not defer to an explanation by the successful-crossing schema $S{\sim}T_1{:}C\_O{\sim}H$, because $S{\sim}T_1C$ is not more general than $S{\sim}H_1C$ (on the contrary, it is less so). Indeed, $*S{\sim}H_1{:}C\_O{\sim}H$ corresponds to a causal link—the action of crossing, when it does occur under the specified circumstances, actually causes reaching the other side without collision. As discussed in section 6, causal links tend to be maximally general and explanatory, so we would not expect that schema to defer to any others (except to those that describe the same causal process at a lower level, in terms of constituent parts; but such a description should be consistent with the higher-level description, and so does not alter the prediction as to what would occur).

Even if we suppose that the choice machine has the schema

$ST_1 : C\_H{\sim}O$ (0.99) (dangerous crossing)

which correctly predicts the outcome of crossing in front of dangerous traffic—namely, a strong likelihood of being hit and not reaching the other side—this dangerous-crossing schema neither overrides nor explains the fatalist schema, so the machinery (as specified so far) does not know which of those conflicting schemas to trust. The dangerous-crossing schema does not empirically override, because its activation does not constitute a special case of the fatalist schema's circumstances of activation (on the contrary, the two schemas' activations are almost mutually exclusive). It does not constitute a more-general explanatory schema, because it asserts a contrary result to that of the fatalist schema, and thus certainly does not explain any occurrence of the fatalist schema's result.

The schema $*S{\sim}H_1{:}C\_O{\sim}H$, if misconstrued as a means-end link when T_1, crystallizes the oft-perceived incompatibility between choice and determinism. The schema captures the fatalist intuition that asks rhetorically: since the past already determines that I'm not actually about to get hit by traffic now, why bother doing anything (such as not-crossing) for the sake of that already-guaranteed goal? Since (it is already determined that) $\sim H$ is true—and since I already know it—I can (seemingly, according to the misleading schema) cross now safely and successfully, despite the dangerous oncoming traffic. The foreseeable predetermination of the outcome $\sim H$ appears to lead to the (obviously absurd) conclusion that crossing now would be safe—i.e., that it would be futile to act (by not-crossing) for the sake of the already inalterably achieved goal of not being hit by traffic.[28] That apparent conclusion from the foreseeable predetermination is seemingly a *reductio ad absurdum* of the foreseeable predetermination.

There is an obvious intuitive reply to this seeming incompatibility between choice and determinism. Yes, I am not in fact about to be hit by traffic; yes, the present state of the universe already inalterably assures that; yes, I already know that the present state already assures that. And yes, anytime I actually cross when I am not in fact about to be hit by traffic, I reach the other side—and (tautologically) am not hit. Nonetheless, I *would* (contrary to actual, already-known fact) very likely be hit (and the present state *would*, contrary to already-known fact, be such that I will very likely be hit) *if*, also contrary to fact, I were to cross now (in front of the oncoming dangerous traffic).

This intuitive reply translates into a solution for a choice machine. An additional principle must be built into the machinery. To reiterate the problem as the proposed machinery stands so far: the fatalist schema $*S{\sim}H_1{:}C\_O{\sim}H$ is applicable when $S{\sim}H_1$ is true, and that schema (since it is not overridden or subject to explanatory deferral, even when T_1 is true) motivates taking action C for the sake of $O{\sim}H$ (since, as the above non-crossing schema correctly asserts, the alternative, $\sim C$, leads to $\sim O{\sim}H$, a result of lesser utility). To block the inappropriate influence of the fatalist schema, the mechanism needs a built-in principle saying that a given applicable, non-overridden schema's prediction is currently *self-revoked* if (according to other applicable, non-overridden schemas) the schema *wouldn't* be applicable now in the first place, nor would its result obtain now, if the schema's action were taken now. More specifically, this self-revocation occurs iff some other applicable, non-overridden schema, or composition of sche-

[28] Of course, a further application of the same reasoning would also argue for the futility of crossing for the sake of the (also already determined, one way or the other) outcome of getting to the other side. But fighting fatalism with fatalism will not produce a reasonable decision strategy.

mas—here, ST_1:*C_H~O*(0.99) composes with :*H_*H_1—would be activated if the given schema were activated (here, *C* would obtain, leading to *H* and in turn H_1), culminating in probable result conditions (here, H_1 and *H*) that contradict the given schema's context and also its result. I propose that a currently self-revoked schema be treated as currently inapplicable; it thus does not contribute to its designated action's current utility.[29]

The principles of self-revocation and explanatory deferral are complementary; neither suffices in the absence of the other. As just noted, **S~*H_1:*C_O~H* is not less general than *S~*T_1:*C_O~H*, thwarting a deferral via the latter's explanation of the former. And conversely, if not for explanatory deferral, self-revocation would be vacuously ambiguous. For instance, as discussed in section 6, explanatory deferral rescues the choice machine from the safely-disposed-crossing schema **SD*:*C_~*T_2. Without the explanatory deferral, **SD*:*C_~*T_2, composed with :~T_2_~T_1 (assuming the existence of that exceptionless schema too), asserts that T_1 wouldn't hold if *C* were to occur now; and **S~*H_1:*C_O~H*, which is also applicable, asserts *O~H* if *C*. So according to those schemas, the context and result of ST_1:*C_H~O* are both contradicted if that schema's action occurs now. Thus, without explanatory deferral, **S~*H_1:*C_O~H* could be construed to help make ST_1:*C_H~O* self-revoking, rather than vice versa.

Explanatory deferral deals with a schema that has not been (or cannot be) tested with respect to a given overriding condition—e.g., **SD*:*C_~*T_2 cannot be tested (by activating it) when T_1 obtains (because *D* prevents *C* when T_1). Self-revocation deals with a schema whose context incorporates the negation of a condition that is actually an outcome—in the subjunctive sense, not necessarily causal—of the schema's activation, even if there is no possible overriding condition under which the schema's result fails to obtain (e.g., with the fatalist schema **S~*H_1:*C_O~H*, there can be no additional condition under which crossing when not actually about to be hit actually results in being hit). Explanatory deferral keeps the means-end-recognizing criterion from being too lax, and thus protects the choice machine from a wildly exaggerated sense of the efficacy of its actions. Self-revocation keeps the criterion from being too strict (in that e.g. the fatalist schema's bogus means-end link from *C* to *O~H*, holding *~H* constant, is effectively a denial of the actual means-end link from *C* to *H*); self-revocation thus protects the machine from fatalist complacency about already-known outcome that the machine still in fact has a choice about.

The analysis of Newcomb's Problem with transparent boxes is exactly parallel to the foregoing. We confront the fatalist intuition that the visible $1M (like the world-state that guarantees I am not about to be hit by traffic) is already known to be inalterably there; and given its presence, taking both boxes box seemingly results in both the primary (already inalterably secured) goal of getting the $1M in the large box, and the secondary goal of getting the $1,000 in the small box (just as crossing the street—given that in fact I am not about to be hit—seemingly results in the primary, already assured goal of not getting hit, plus the secondary goal of getting to the other side); whereas taking just the large box results in only the primary goal of getting $1M, forfeiting the $1,000 (just as not-crossing still meets the primary goal of safety, but forfeits the secondary goal of reaching the other side). So it is apparently futile to act by refraining from taking the small box (or by not crossing) for the sake of the already-inalterable, already-known-to-be-secured primary goal, at the cost of forfeiting the secondary goal.

But that appearance is false; the action is not futile. If (contrary to fact) the alternative action were taken, then (also contrary to fact) the (albeit already known to be inalterably secured) primary goal would *not* be achieved. (Of course, this subjunctive link does not mean that the already secured goal—in Newcomb's Problem or the street-crossing problem—ever *changes* its state, in any actual situation, from one moment to the next, when the action is taken. It does not.)

Just as in the opaque-box case, the means-end link in the transparent-boxes problem comes from the composition of two other such links:

- If (contrary to fact) you were to take both boxes, then (also contrary to fact) the snapshot-time past state of the universe would be such that (according to physics) you will choose both boxes if presented with $1M in the large box. Just as in the hand-raising past-predicate example, you do not cause the past to be what it was; yet you do have a choice about this particular aspect of the past, exactly as you have a choice of which action to take now.
- If (contrary to fact) that past state were such that you will choose both boxes if presented with $1M in the large box, then (also contrary to fact) the benefactor would put no money in the large box. Achieving that aspect of

[29] By the present proposal, a schema is self-revoking only if its result as well as its context is contradicted by the schema's action (according to other applicable schemas). To see why, consider a *chaperoned* street-crossing scenario in which I would certainly have been kept away from the street unless I had the competence and desire to cross safely. The schema ST_1:*C_H~O*(0.99) is not self-revoking even though, if *C* could occur given ST_1, circumstances ST_1 would not have obtained in the first place.

the past state is a means to the goal of there being \$1M in the large box; this means-end link is conventionally causal, mediated by the simulation.

Schemas expressing those two composed-together means-end links (similar to the schemas in the previous section) describe the structure here just as in the opaque-box version of the problem:

$NM_1 : B \_ \sim P, \quad NM_1 : \sim B \_ P,$

$NN_{100} : P \_ M_1, \quad NN_{100} : \sim P \_ \sim M_1, \qquad NN_{99} : P \_ M_1\,(.99), \quad NN_{99} : \sim P \_ \sim M_1\,(.99)\ ,$

$: M_1 \_ M\ , \quad : M \_ M_1\ , \quad : \sim M_1 \_ \sim M\ , \quad : \sim M \_ \sim M_1\ ,$

where we define some of the predicates slightly differently than in section 7:

N := I am now presented with a transparent-boxes Newcomb's Problem choice.

P := The past state of the universe at the time of the snapshot is such that (according to physics) if I see \$1M in the large box, I will take only the large box, forfeiting the \$1,000 box.

M_1 := The large transparent box in front of me contains \$1M.

M := I obtain \$1M in the large box in front of me.

K := I obtain the \$1,000 in the small box in front of me.

And, just as in the street-crossing example, these schemas render self-revoking the schema

$*NM_1 : B \_ MK\ ,$

because a result (in the subjunctive, not necessarily causal, sense) of that schema's action B, according to the schemas above, would be $\sim P$ and in turn $\sim M_1$ and $\sim M$, which respectively contradict the context and result of $*NM_1 : B\_MK$.

The foregoing analysis applies even if we consider a fallible simulator, as in the N_{99} case. It is then possible (though unlikely) for the benefactor to (mistakenly) place \$1M in the large box even if you will take both boxes—just as, in the street-crossing problem, it is possible (though unlikely) to cross successfully and without collision even as dangerous traffic approaches. Even though the \$1M thus does not imply the impossibility now of taking both boxes (similarly, even though not being about to be hit does not imply the impossibility of crossing now), it is very likely that if you were to take both boxes, the simulation would have so predicted, and the large box would have been left empty, even though in reality it wasn't (it is very likely that if I were to cross now, the traffic would hit me, even though in reality it does not), and so you should not do so (ditto). As with the opaque-box version of the problem, the expected utility of what *would* be the case if you were to take both boxes, or just the large one, is computed with respect to the subjunctive probability of a correct simulation in each case (because of the partly acausal means-end link from the action to the simulation result); even a modest probability of correctness (say, 0.9, or even 0.6 or less) suffices to justify the one-box choice.[30]

One might try to distinguish the street-crossing example from the transparent-boxes problem by appeal to *how you know* that the goal, or a guarantee thereof, already obtains. In the street-crossing problem, you know $\sim H_1$ in advance because you know your choice $\sim C$ in advance, whereas in Newcomb's Problem, you know M_1 because you *see* the \$1M. But you might have an additional basis for knowing M_1: as Gibbard and Harper note (regarding the opaque box, but it applies to the transparent box as well) you might figure out which choice you will make and what, accordingly, has been put in the box. And conversely, in the street-crossing problem, you might have an additional way to know $\sim H_1$: suppose a rightly trusted third party assures you that you are not in fact about to be hit by traffic (perhaps the third party is just extrapolating from trillions of instances in which you, or others of safe disposition, stand waiting to cross carefully with a clear view and are never then struck by traffic, even if dangerous traffic approaches; the third party need not even know whether the action of crossing occurs in those instances). In both problems, then, you could have two simultaneous, independent bases for your foreknowledge of the actual outcome—one basis that derives from knowing what your actual choice will be, and one that is otherwise derived.

[30] Another way to justify the one-box choice is to note that for all you know, you might be the simulated you; hence you should act in part for your (causal) influence on the simulation outcome. This view is consistent with but does not obviate the present subjunctive approach. Say the real you assumes that it is the real you. Nothing false can logically follow from that true (even if unjustified) assumption; in particular, nothing false follows as to which choice is in fact more lucrative for you to make. (The simulated you might, of course, infer false conclusions from its false assumption that it is the real you.) So the one-box choice is more lucrative for you only if you cannot infer otherwise by assuming that you are indeed the real you.

Ultimately, then, the dominance argument in Newcomb's Problem—the argument that the large box already contains either $1M or is empty, and either way, you get $1,000 more if you take both boxes than if not—fails because on the contrary, if there is $1M in the large box, you *would* (probably) get more if you were to take just the large box than if you were to take both—because in the latter case the simulation would (probably) have so predicted and thus the large box would (probably) be empty (even if in fact it is not), whether or not the large box is opaque.

As an (albeit inconclusive) point of confirmation that the one-box choice is correct even in the transparent-boxes version, note that in that version, as in the opaque-box formulation of the problem, an individual who would choose just the large box (provided—in the transparent case—that it contains $1M) has a higher expected gain from a Newcomb's Problem encounter than someone who would take both—the former will indeed be offered $1M in the large box (or probably so, given a fallible simulation), the latter (probably) an empty large box. Put another way, if you could contemplate the situation in advance—before the benefactor even takes the snapshot, or runs the simulation—you would hope that when the time comes, you will take only the large box, even though you will already know what it contains, because otherwise the world at the time of the snapshot will be such that the simulation will probably predict your taking both boxes. If there were no rational justification for taking just the large box when the time comes, then rationality would paradoxically compel you to act in a manner contrary to how you correctly wished (in advance) you would act under the very circumstances that you know have now arisen.[31, 32]

The fatalist intuition that it is futile to act for the sake of an already-known, already-inalterable goal can be induced by focusing on the known prior guarantee in the deterministic street-crossing scenario; but the intuition asserts itself far more insistently in Newcomb's Problem, especially with transparent boxes. I suspect the disparity comes from the differing obviousness of the present-time, already fixed, already known guaranteeing conditions in the various problems. With the transparent large box, the guaranteeing condition is simple, concrete, and plainly visible; with the street-crossing example, the condition is complex and abstract, and we are easily oblivious to it. The original version of Newcomb's Problem is intermediate: there, the guaranteeing condition is simple and concrete, but not visible.

With the transparent large box, *seeing* a virtually certain prediction of your choice—and seeing the goal state itself (or a guarantee thereof)—makes it impossible to ignore or de-emphasize that the choice and its outcome are already established. The seeming conflict between choice and determinism intrudes before our very eyes, and any lingering intuitive allegiance to fatalism makes its final stand. But the goal state in the transparent-boxes problem is no more pre-established than goals always are, given determinism. It is just more *blatantly* pre-established.

9. The Prisoner's Dilemma

As Lewis (1979), Horgan (1981), Leslie (1991)[33] and others have pointed out, Newcomb's Problem is structurally similar to the Prisoner's Dilemma (though Lewis advocates taking both boxes in Newcomb's Problem, and correspondingly recommends the uncooperative choice in the Prisoner's Dilemma). The details spill well beyond the bounds of this paper, but the gist is easily stated and important enough to warrant mention. In one version of the (non-iterated) Prisoner's Dilemma, two mutually-incommunicado players face symmetric, simultaneous choices, and are assumed to be able to figure out the correct choice (and to have mutual knowledge of their ability and knowledge; Hofstadter 1985 terms such agents *superrational*). Each faces a binary choice between cooperating or defecting. Given that the other cooperates, each does better to defect; given that the other defects, each also does better to defect. But both do better if both cooperate than if both defect. By assumption, each of the two acts only for the

[31] Much the same dissonance arises in a suggestion by Rodrigo Vanegas (personal communication) in response to the transparent-boxes variation. Proposing the converse of Nozick's peeking-friend ploy, Vanegas recommends that you *keep your eyes closed*, transforming the problem back to an opaque-box problem, allowing you to take just the large box and reap $1M. Otherwise, the evidentialist argument goes, you would be rationally compelled to take both boxes, and would thus see an empty large box.

[32] Consider another transparent-boxes variation of Newcomb's Problem. The benefactor conducts *two* simulations, one showing you presented with $1M in the large box, the other simulation showing you presented with an empty large box. Both simulations are highly (but not perfectly) reliable as to what you would do if you in fact encountered the specified large-box content. The benefactor places $1M in the large box iff *both* simulations show you taking just the large box. The subjunctive reasoning above then implies (correctly, I claim) that you should take just the large box even if it is empty. And once again, taking just the large box—even if it is empty—is what you would correctly wish in advance that you would do. Once again, someone who would do that reaps a larger payoff, on average, from Newcomb's Problem encounters than someone who would not.

[33] Leslie (1991) proposes that *quasi-causation* connects the behavior of two or more causally independent entities that operate according to "similar" causal factors, justifying the one-box choice in Newcomb's Problem, and cooperation in the Prisoner's Dilemma. But in order to do the work that Leslie requires of it, quasi-cause must invoke a very broad, abstract sense of similarity—the simulator may be implemented in a different technology (transistors vs. neurons) than its subject, for example, so that only at a high enough level of abstraction do the simulator and subject correspond sufficiently to construe one as a representation of the other. We must then explain why, for instance, if I will cross the street iff there is no dangerous traffic, crossing is not abstractly similar to—and thus quasi-causative of—the absence of dangerous traffic, given the correlation by which we could construe either as a representation of the other in the specified situation.

sake of doing better personally, with no inherent regard for the other's welfare; and unlike in the iterated Prisoner's Dilemma (addressed below), one's choice to cooperate or defect has no relevant effect other than the immediate pay-off to the two players.

From a given player's standpoint in the Prisoner's Dilemma, the other player's choice process, due to its postulated symmetry, is effectively a high-level simulation of the given player's choice—not a particle-by-particle simulation, but what Dennett (1987) has called a *folk-psychological* model (which, for all its pre-scientific informality, is often quite reliable; my effort to add a particular pair of three-digit numbers, for instance, effectively simulates your own effort to do so, and reliably predicts your answer, even if we use different—but correct—arithmetic algorithms). As in Newcomb's Problem, the chooser knows that the simulation's prediction (i.e., the other player's choice) would very probably correspond to whichever her own choice is, despite the absence of a causal link. Crucially, beyond the (evidential) prediction, there is a choice-supporting subjunctive link, a means-end link, to the other player's "simulation" of the given player's choice, as there is to the simulation outcome in Newcomb's Problem.[34, 35]

As is widely recognized, the non-iterated Prisoner's Dilemma encapsulates the fundamental question of whether it can be construed as rational, from the standpoint of self-interest, for you to cooperate with another individual when doing so *causes* net disadvantage to yourself. Analyzing the Prisoner's Dilemma as a variant of Newcomb's Problem argues that your cooperative action indeed stands in a means-end relation to the goal of the other's cooperation (and the benefits thereof to you), even in the absence of a causal link.[36] And of course, the Prisoner's Dilemma is easily generalized to situations where A cooperates with B because A desires C's cooperation, who symmetrically desires D's... which suggests a rationale for treating others well because you want (possibly different) others to treat you well (even in the absence of a causal link from your behavior to theirs).

But suppose you already know approximately how well, on average, others will treat you—as you arguably do know by projecting from past experience. The original version of Newcomb's Problem fails to provide a justification for your cooperation given that foreknowledge (in cases where your cooperation does not cause others' cooperation). The transparent-boxes version, however, does speak to this point, providing a more robust foundation for cooperation—a foundation that is not undermined by already knowing the other players' moves and thereby knowing the extent to which your goal has already been achieved.[37]

Of course, other factors promote cooperation too: punishments and rewards of various sorts, likely built-in inclinations to empathy and altruism at least under some circumstances, etc. Axelrod (1984) has shown that genetic algorithms in *iterated* Prisoner's Dilemma situations (in which your choice in one trial can cause others to punish or reward you in later trials) can evolve cooperative strategies, suggesting a possible origin of biological entities' inclination toward self-sacrificing cooperation in some circumstances. Frank (1998) has observed that human emotions, when they impel conduct seemingly contrary to rational self-interest, can often be seen as promoting rationally cooperative behavior; an innate tendency to such emotions might help implement an Axelrod-like evolved cooperative inclination. Pinker (2002) and Dennett (2003) have elaborated these themes with a wealth of evidence and argument. But the implications of an iterated Prisoner's Dilemma do not explain why we should not try to overcome whatever built-in (or socially inculcated) inclination toward cooperation we have (just as we may try, often successfully, to overcome our inclination to eat large quantities of fat and sugar) in the *non*-iterated Prisoner's Dilemmas that we face whenever we can profit from behaving badly *and not get caught*. Beyond any built-in (or externally imposed) inclination to cooperate, we still need a *reason* not to regard that inclination as pointlessly and undesirably self-

[34] Here, though, the subjunctive link is not via what the past universe-state would be (hence what input the simulator would receive), but rather via what choice a competent player would regard as correct (hence what the other player would so regard). One consequence of this difference is that the Prisoner's Dilemma "simulation" only works to the extent that both players have choice machinery competent to solve the problem correctly.

[35] Hofstadter (1985) defends cooperative action in the non-iterated Prisoner's Dilemma only by appeal to an explicitly evidentialist criterion: what you do informs you of what others like you will do in like circumstances, so you should do what you want them to do, he claims.

[36] Game-theoretic analyses of the non-iterated Prisoner's Dilemma are tangential to the fundamental issue. Before game theory is brought to bear, the problem must be formalized so as to designate the consequences of a player's potential moves. In the conventional formulation, your choice and your opponent's are independent, and game theory trivially endorses defection as the dominant strategy (see e.g. Binmore 1994). But if your choice were able to (probably) *cause* the other player to make the same choice, the game would be formalized such that game theory trivially endorses cooperation. The present claim is that the acausal subjunctive link to the other player's choice likewise makes the other's choice a (probabilistic) "consequence" in the sense relevant to means-end analysis, so the game formalization should be just as though the link were causal. Game theory per se does not address the means-end analysis; given that analysis, the game (so to speak) is over before game theory even makes its move.

[37] Further, the two-simulation version of the problem (note 32), in the case where the large box is empty, helps explain why one should treat others well even if one has been treated improbably badly by (different) others. Pursuing self-harmful retribution (a la Frank 1998) is also analogous to choosing just the empty large transparent box: the prospect of retribution failed to have its probable deterrent result (just as the actual one-box choice improbably failed to be predicted), but one may still reasonably act for the sake of that (retroactive, acausal, contrary-to-fact) outcome.

destructive, as a temptation to be resisted rather than cultivated. An acausal means-end link to others' reciprocity may provide that reason.

10. Summary

To understand how genuine choice could be mechanical—to reconcile choice with determinism, or even with approximate determinism—we must confront the compelling fatalist intuition that it is futile to act for the sake of that which our action cannot alter—the intuition that inalterability implies futility. Contemplating the operation of a simple choice machine demonstrates that choice is a particular mechanical process which, like any other such process, is no less real for having been predetermined. Contemplating the choice of a distant-past state (e.g. in the hand-raising example) shows that, once inalterability is no longer deemed a prohibitive obstacle, we can be seen to have a choice about some aspects of the world over which we lack any causal influence. Newcomb's Problem simply harnesses our ability to choose some such aspects of the world, using an acausally chosen state as a sub-goal to the goal of the eventual reward (the latter step via a merely causal path). Similarly with the non-iterated Prisoner's Dilemma.

Newcomb's Problem distills the challenge posed by deterministic choice. If inalterability does not imply futility, then being unable to alter the box's content does not necessarily undermine the desirability of acting for the sake of its content being one way or another. The transparent-boxes version goes one crucial step further. If an outcome is already determined anyway, then already knowing (or even literally *seeing*) what the outcome will be does not further undermine the efficacy of an action with respect to that outcome. If it can make sense to act for the sake of the unknown, inalterable box content, the same holds even if instead that inalterable content is already known. The transparent box makes the conflict between choice and determinism more blatant, without changing its essential character. If the conflict is not resolved in its most blatant form, then it is not resolved, but rather just partly concealed.

Whereas much of the literature construes arguments against evidential means-end relations as supporting exclusively causal means-end relations and vice versa, I propose here an intermediate approach. I construct a subjunctive sense of means-end relations—a choice-supporting sense of what *would* be the case if this or that action were taken—broad enough to include some acausal evidential relations, but narrow enough to exclude others. The machinery I sketch for recognizing means-end relations uses schemas that express actual correlations among specified conditions. A default presumption of conditional independence from other conditions can be (unremarkably) superseded by an empirical-override provision. An explanatory-deferral provision can defeat the default presumption that a schema's evidential relation is also a means-end relation, that its conditional probability is also a subjunctive probability. And finally, self-revocation can defeat the same presumption with regard to a schema whose context depends subjunctively on the action itself.

I attempt to justify the proposed means-end recognizing machinery by appeal to mundane situations (e.g. the street-crossing scenario) that help to isolate and examine the relevant principles—the sort of situations that our cognitive machinery must have evolved to deal with. Thus grounded, the means-end machinery can then be applied to esoteric or controversial scenarios, such as Newcomb's Problem or the Prisoner's Dilemma.

The details of the means-end machinery here are proposed tentatively; I do not expect to find that they are complete and correct. But I am hopeful as to the merit of the general approach that those details illustrate—the methodology of using thought experiments that presume determinism and zero-probability idealizations, and the attempt to derive candidate means-end links from contrasts among actual situations, then winnow them in part by the deference of some candidate links to more-general explanatory links, and in part by deference of a link whose context depends (in the appropriate subjunctive sense) on the very action under consideration.

The nature of the means-end relation speaks to the perennial question of whether and how we can derive what *ought* to be from what *is*. Even with respect to pursuing purely self-centered goals, deciding what action one *ought* to take for those goals requires, as a starting point, some built-in kernel of means-end recognition—a way to derive what *would* be from what is. But the reduction of the Prisoner's Dilemma to Newcomb's Problem (especially with transparent boxes) argues that a deliberative choice machine—even with just self-centered goals and with just the built-in means-end-recognizing principles that the machine needs to pursue those goals in mundane situations—could in principle—without any *further*, specifically altruistic presupposition or inclination—derive *ought* from *is* in a way that prescribes cooperation with others, even when cooperation causes no personal benefit.

Acknowledgements

I pursued much of this work as a Visiting Fellow at the Center for Cognitive Studies at Tufts University. I am grateful for spirited discussions with Daniel Dennett and Uri Wilensky, and with Jim Davis, Gabriel Love, Will Lowe, Oliver Selfridge, Christopher Taylor, and Rodrigo Vanegas.

References

Allais, M. and Hagen, O. eds. (1979) Expected Utility Hypotheses and the Allais Paradox. Reidel.

Axelrod, R. (1984) *The Evolution of Cooperation*. Basic Books.

Binmore, K. (1994) Playing Fair: Game Theory and the Social Contract I. MIT Press.

Blackburn, S. (1998) *Ruling Passions: A Theory of Practical Reasoning.* Clarendon Press.

Dennett, D. (1980) The Milk of Human Intentionality, *Behavioral and Brain Sciences*, 3, pp. 428-430.

(1984) *Elbow Room: The Varieties of Fee Will Worth Wanting.* MIT Press.

(1987) *The Intentional Stance.* MIT Press.

(2003) *Freedom Evolves.* Viking.

Drescher, G. (1991) *Made-Up Minds: A Constructivist Approach to Artificial Intelligence*. MIT Press.

Eells, E. (1982) *Rational Decision and Causality.* Cambridge University Press.

Fine, K. (1975) review of Lewis, *Mind*, 84, 451-8.

Flavell, J. and Markman, E., eds (1983) *Child Psycholog: Volume III, Cognitive Development*. Wiley.

Frank, R. (1998) *Passions Within Reason: The Strategic Role of the Emotions*. Norton.

Gibbard, A. and Harper, W. (1977) Counterfactuals and Two Kinds of Expected Utility, *Foundations and Applications of Decision Theory,* ed. C.A.Hooker et al. Dordrecht:Reidel.

Goodman, N. (1983) *Fact, Fiction, and Forecast*, Fourth Edition. Harvard University Press.

Hofstadter, D. (1985) *Metamagical Themas: Questing for the Essence of Mind and Pattern.* Basic Books.

Horgan, T. (1981) Counterfactuals and Newcomb's Problem, *Journal of Philosophy*, 78, 331-356.

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

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

Kavka, G. (1983) The Toxin Puzzle, *Analysis*, 43, 33-6.

Leslie, J. (1991) Ensuring Two Bird Deaths With One Throw (Quasi-causation and Newcomb's Problem), *Mind*, 100, 73-86.

Lewis, D. (1973) *Counterfactuals.* Harvard University Press.

(1979) Prisoners' Dilemma Is a Newcomb Problem, *Philosophy and Public Affairs*, viii, 3, Spring 1979.

MacKay, D. (1960) On the Logical Indeterminacy of a Free Choice, *Mind*, 69, pp. 28-41.

Minsky, M. (1968) *Semantic Information Processing*. MIT Press.

(1986) *The Society of Mind.* Simon and Schuster.

Nozick, R. (1969) Newcomb's Problem and Two Principles of Choice, in *Essays in Honor of C.G. Hempel*, ed. N. Rescher et al. Dordrecht: Reidl.

(1993) *The Nature of Rationality.* Princeton University Press.

Pearl, J. (2000) *Causality*. Cambridge University Press.

Piaget, J. (1952) *The Origins of Intelligence in Children.* Translated by Margaret Cook. Norton.

Pinker, S. (2002) *The Blank Slate.* Viking.

Shubik, M. (1982) *Game Theory in the Social Sciences: Concepts and Solutions.* MIT Press.