Counterfactuals in Thought

There's an important connection between preferences, i.e. facts of the form:

> *making A true is better for me than making B true*

and counterfactual facts. Facts about:

> *how good things would be for me if A had been true* and *how good things would have been had B been true.*

However, it's commonly assumed that the direction of analysis and explanation goes from the latter sorts of facts to the former — the value facts are defined and explained in terms of the counterfactual facts. In this paper I will prove that counterfactuals are implicitly defined by their role in thought by way of their connection with rational preference and value (section 1-6). I will consider a number of different applications of this fact. In section 7 I'll consider the idea that counterfactual facts can be reduced to value facts. While I am of the view that analyses are always circular to some degree, and and that counts as a reduction is highly guise sensitive, I'll argue that the circle of analysis between counterfactuals and value is tighter, and has more wide ranging implications, than circles of analysis that involve similarity, or probability. In section 8 I argue that the fact that value pins down counterfactuals allows us to single out, from a plethora of contextually salient conditionals, a special conditional for doing deliberative reasoning by its role in thought. And in section 9-11 I will attempt to reinstate the logic of conditionals on a firmer basis than the usual body of linguistic intuitions about sentential truth and validity by describing a general correspondence between principles of conditional logic and principles of decision theory, which can be adjudicated by appeal to considerations about rational preference.

1 Objective Value

A standard formalism for modeling value and preference, having its origin in utilitarian ethics[1], theorizes in terms of one or both of two real valued functions: the objective

[1] Although not necessary wedded to it, see e.g. Wedgwood (2017).

value and/or the subjective value of a given prospect.[2] I will therefore begin by introducing these notions. I will introduce them as I think of them, namely as self-standing notions that can be given a direct and intrinsic axiomatization without reference to counterfactuals, or rational degrees of belief.

By a prospect I simply mean any consistent proposition that can be true or false, but typically a proposition that is relevant to whatever the agent cares about. Prospects will be identified with non-empty sets of worlds. The objective value of a prospect, A, for a given agent, is a measure of how good As being true is for the agent. This measure of the goodness of a prospect must make sense even for prospects that fail to happen because the agent ultimately rejects them, for otherwise it could play no role in the deliberation that happens *before* she decides to reject them, before she knows that they are false. By objective value I really mean objective *agential* value: it is objective because it is insensitive to your subjective degrees of belief, *not* because it is insensitive to individual preference. What is objectively valuable may still differ between people—whereas subjective value is a quantity that takes your subjective uncertainty into account.

To illustrate, suppose I suggest we both bet a dollar on whether yesterdays lottery number, which at present is unknown to us, is even or odd. I win and you lose if it is even, and I lose and you win if it is odd. As it happens it *is* even, so the objective value for me of accepting the bet is \$1. In worlds where it is odd the objective value for me is \$-1. Each person has their own objective value function: for you the objective value would be by described reversing these numbers. Objective value can in general be represented by a binary function $v.(\cdot)$ that takes a prospect (i.e. a non-empty set of worlds), A, and a world w and returns a real number $v_A(w)$ telling us how good it is for A to be true, for the agent, in world w. v_A is sometimes called a random variable: it represents an unknown number. In our example it represents 1 in some worlds, and in others -1.

Let's introduce some notions. A *total outcome*, for an agent, is a complete specification of everything the agent could possibly care about. For present purposes we can, without loss of generality, identify total outcomes with possible worlds—complete specifications of everything, and *ipso facto* everything the agent could care about. Different outcomes can be more or less valuable to an agent. The value of an outcome, w, is simply a measure of how good things are generally for the agent in that world; it is the value of things as they are and will be without intervention, *not* the value of any action we might perform to change the world in this way or the other. Above we have introduced the value of a prospect A at a world w, $v_A(w)$. *Outcome value* by contrast, is a function v_* that assigns each outcome (world) a real number representing how good the outcome is to the agent. Thus officially, an objective value will be identified with an ordered pair $(v.(\cdot), v_*(\cdot))$ of prospect value and outcome value. However, if we make

[2]Terminological warning. Elsewhere I have used "actual value" instead of "objective value". Note that objective value is still agent relative. It is objective in the sense that it is insensitive to your subjective degrees of belief, but not objective in the sense that it doesn't vary from person to person.

a very modest assumption, we can identify outcome value with the prospect value of the tautologous prospect—the action of doing nothing—which we call the *status quo* and write as $\top$, at that world. The value of an outcome w for an agent is thus defined as $v_\top(w)$. When we are talking about outcome value, as opposed to prospect value, we will sometimes drop the subscript and just write $v(w)$.

Outcome Value is the Status Quo For any rational objective values $(v_\cdot(\cdot), v_*(\cdot))$, $v_* = v_\top$.

In the presence of this assumption an agents objective value can be identified with a single function $v_\cdot(\cdot)$ and treat outcome value as defined by $v_\top$.

Although it is officially a primitive of the theory, the notion of objective (or subjective) value is not free floating and is tightly constrained by its relationship to other notions. If one is in a position to make one of several propositions true, the one with the highest value (objective or subjective, depending on the theory) is the one you *should to make true*. So there is a connection with prudential obligations. There are also intimate connections between value, credence and preference which will be drawn out shortly. Perhaps the notion of value itself is in turn reducible to one of these other notions (preference being a common choice).[3] I will not take a stand on that issue here, but will just remark that the notion of propositional value seems to be in at least as good a standing as these others notions in virtue of their relationship.

We will use the symbol $\mathcal{O}$ for the set of rationally permissible objective value functions. The space of permissible objective value functions, $\mathcal{O}$, is subject to a formal constraint. The objective value of a proposition, A at a world x, for an agent has to be the value of some possible outcome. This just means that the objective value of a prospect cannot be impossible: it cannot have a certain value if there is no possible outcome that has that value. Moreover, the outcome of a prospect is an objective agent independent thing. It cannot be that the objective value of a bet for me is given by the outcome of me winning \$1 off you, and that the objective value of the very same bet for you is given by the different outcome me losing \$1 to you. If, in every outcome I benefit iff you lose out somehow, then the objective value of a prospect cannot be good for both me and you — the outcome of a prospect is the same for me as for you.

Objective Outcomes For any prospect A and world x, there exists an outcome y "the actual outcome" of the prospect, with the property that: $v_A(x) = v_*(y)$ for every $v_\cdot(\cdot) \in \mathcal{O}$.

Objective Outcomes has some desirable consequences. While two people may value things differently, they cannot disagree in their objective values about how good a

[3]See, e.g., Jeffrey (1990), chapter 5. Such reductions are often motivated by concerns about psychological reality of real valued functions like V. Typically when an agent psychology can be modeled by one real valued function there will be a range of transformations of that function that will represent their psychology equally well. A qualitative ranking of preference between propositions, it is claimed, does not have this proliferation of equally good representations.

prospect is for them, without disagreeing about the value of some outcome. Or, putting it another way, if every possible way things could turn out looks equally good to two agents, then a given prospect cannot be better for one agent than the other. Thus, for any two objective value functions, $v_\cdot(\cdot)$ and $u_\cdot(\cdot)$:[4]

If $v_\top = u_\top$, then $v_A = u_A$ for any proposition A.

This says that objective values supervene on the values of the possible outcomes. This consequence is worth stating because it articulates an important way in which objective values will be different from subjective values: subjective values do not supervene on how the agent values the outcomes alone, they depend also on the agents credences in those outcomes.

There are a couple of further constraints that are, I think, undeniably true, but I will not build into the notion of an objective value function for reasons that will become clear later. The first is that any two truths should have the same objective value. The objective value of a prospect that actually obtains is just going to be a measure of how things will actually go for the agent, and there is only one way things will actually go for the agent. For instance, since the lottery number is in fact even, and I have in fact betted on it being even (the prospect has been accepted and it is in fact true), the prospects value is \$1, the value of the actual world. Thus, for any permissible objective value function $v_\cdot(\cdot)$, and propositions A and B that are true at x, the objective value of A at x is the same as the objective value of B at x:

Truth Indifference $v_A(x) = v_B(x)$ whenever $x \in A \cap B$.

Another, equivalent, way of putting this is that the objective value of a truth at a world is just the value of the actual outcome at that world: $v_A(w) = v_*(w)$ whenever $w \in A$.

The second is that the outcome of a prospect should compatible with that prospect. For any prospect A, the objective value of A is the value of a particular outcome compatible with A.

Prospect-Outcome Compatibility The outcome of a prospect is compatible with that prospect: for any prospect A and world x, there exists an outcome $y \in A$ such that $v_A(x) = v_*(y)$ for every $v_\cdot(\cdot) \in \mathcal{O}$.

A final principle about objective value is the principle that the objective value of a prospect that is a choice between A or B, $A \vee B$, is either the same as the objective value of A or the objective value of B

Prospect Choice $v_{A\vee B}(x) = v_A(x)$ or $v_{A\vee B}(x) = v_B(x)$, for any prospect A, world x and $v \in \mathcal{O}$.

I do not take this principle to have the same self evidence as the previous two principles, but it certainly possesses some degree of attractiveness.

[4]Suppose that $v_\top = u_\top$, and we want to show $v_A(x) = u_A(x)$. By Outcomes there exists a world y such that $v_A(x) = v_*(y)$ and $u_A(x) = u_*(y)$. And since $v_*(y) = u_*(y)$ the result follows.

2 Subjective Value

Let us now turn to subjective value. The subjective value of a proposition for an agent will also be represented by a real number measuring how (instrumentally) good the agent regards the truth of that proposition to be for them. We will suppose, naturally, that this number is sensitive to a body of evidence that the agent possesses. For instance, perhaps the prospect of buying a certain painting has high value until you learn it is a fake, upon which the value of the prospect decreases. So we shall represent an agents values across time with a real valued function taking two propositions as arguments: a proposition A that is being evaluated, and a body of evidence E:[5]

> $V_E(A)$ is a real number representing the value of A being true to a certain agent with respect to the body of evidence E.

Unlike objective value, the subjective value of a prospect does not just depend on what you care about, but also depends on your degrees of belief. For instance, while the objective value of the bet we described in the previous section is different for me and for you (\$1 for me and \$-1 for you), we may assign it the same subjective value, because we may both think it is as likely that we win the bet as lose.

The standard account of subjective value is based on expected utility theory (although the standard account is not universally accepted, see, for instance, Buchak (2013)). According to the standard theory, subjective value is a weighted average of the objective values at different possibilities, weighted by how likely those possibilities are. So if my credences assign 0.5 to the number being even, and to it being odd, then my subjective value is $0.5 \times 1 + 0.5 \times -1 = 0$. However, I will not begin with the assumption that subjective value is an expectation of another quantity, objective value. I will instead take subjective value as primitive, represented by a binary function V that takes a body of E, and a prospect A, and returns a real number $V_E(A)$ representing the subjective value of A being true to a certain agent with respect to the body of evidence E. We will use capital letters, like V and U, for subjective values. We will write $\mathcal{S}$ for the set of rationally permissible subjective values. As before, we will identify outcomes with worlds. Officially outcome value is an independent function u mapping outcomes to numbers, but we shall adopt an assumption that lets us define it as the value of doing nothing (the *status quo* prospect) conditional on that outcome obtaining: $u(w) := V_w(\top)$. $\mathcal{S}$ is intuitively subject to some formal coherence constraints. Write $P_V(E)$ for $\frac{V_\top(B)-V_{\neg E}(B)}{V_E(B)-V_{\neg E}(B)}$ for some arbitrary proposition B.

Priors and Evidence If $V_\top = U_\top$ then for any E, $V_E = U_E$

[5] In assuming that an agents evidence can be encoded by a single proposition we are assuming (i) that evidence is *propositional* and (ii) that it is closed under conjunction introduction. Those who deny (ii) but not (i) will typically identify a body of evidence with a set of propositions, in which case our formalism should be intelligible through the interpretation of E not as a piece of evidence the agent possesses but rather as the conjunction of their evidence (which may not itself be evidence).

Evidential Linearity P_V is a probability function and independent of B, and $V_E(A) = \sum_i P_V(E_i \mid E) \times V_{E_i}(A)$ for any partition E_i

Subjective Outcomes For any proposition A, there is an "outcome" partition $(E_w)_{w \in W}$ indexed by W, such that for any value $V \in \mathcal{S}$, world w, and $E \subseteq E_w$, $V_E(A) = V_w(\top)$.

Each of these coherence constraints can be motivated on intuitive grounds, much as we motivated the constraints on objective value. Priors and Evidence simply says that if two people have the same prior subjective values (i.e. the same values on the tautologous proposition), then they have the same values on any piece evidence: your values are determined by your priors and your evidence. Evidential Linearity is a consequence of the idea that the value of a prospect is an expectation. It can be derived from two more basic principles (due to Bolker (1966)), sometimes called Averaging and Impartiality, however, since the justification of Evidential Linearity is not my primary concern here I will not take this route.[6]And Subjective Outcomes is motivated in a way similar to Objective Outcomes. The subjective value of a prospect should be the expected value of the values of the possible outcomes. So for each prospect A, there is a partition of propositions, E_w, expressing the claim "the outcome of A is w", and conditional on this proposition, the value of the prospect should be the same as the value of the outcome w. An important consequence of these principles is:[7]

If $P_V = P_U$ and $V_w(\top) = U_w(\top)$ for every outcome w, then $V = U$.

this says that subjective value supervenes on the values the agent assigns to outcomes and their degrees of belief. (Contrast this with the result that the objective values of prospects depends only on the values of outcomes discussed in section 1.)

I have presented theories of both objective and subjective values by some internal coherence constraints, but I have not said how they relate. According to the standard account, expected utility theory, the subjective value of a prospect, with respect to some body of evidence, is simply the expectation of its objective value. Let us suppose that an agents rational degrees of belief, relative to a body of evidence E, can be represented by a probability function P_E. The connection is spelled out as follows:

Expected Utility Theory $V_E(A) = \sum_w P_E(w) \times v_A(w)$

[6]These principles are:

Averaging If E_1 and E_2 partition E into two, then the value of E lies between the value of E_1 and E_2: $V_{E_1}(A) \leq V_E(A) \leq V_{E_2}(A)$ or $V_{E_2}(A) \leq V_E(A) \leq V_{E_1}(A)$

Impartiality If E, E' and D are pairwise disjoint, and $V_E(A) = V_{E'}(A)$, and $V_D(A) \neq V_E(A)$, then if $V_{E \cup D}(A) \neq V_{E' \cup D}(A)$, then for any D' disjoint from both E and E' and such that $V_D(A) \neq V_E(A)$, $V_{E \cup D'}(A) \neq V_{E' \cup D'}(A)$

[7]One simply applies Linearity to V and U with respect to the partition of outcomes (i.e. worlds).

If we have a space of objective values $\mathcal{O}$ satisfying Objective Outcomes, and we define $\mathcal{S}$ as the set of all possible expectations of objective values using the above formula, $\mathcal{S}$ will satisfy our coherence constraints for subjective values. Conversely, if we have a space of subjective values $\mathcal{S}$ satisfying our constraints, we can for each $V \in \mathcal{S}$ define an objective value as $v_A(w) := V_w(A)$, and the set of such v will satisfy the constraint for objective value, Objective Outcomes.

Alternative accounts of the relation between objective and subjective value are possible; notably Buchak (2013). These theories agree that subjective value is determined, at least in part, by the agents credences and their objective values.

3 Counterfactuals

Let us now briefly recap some standard, and hopefully familiar, technology for modeling conditionals. The standard framework—developed, among other places, in Stalnaker (1968) and Chellas (1975)—takes a conditional to be modeled by a *selection function*. A selection function is a function, f, that takes a non-empty set of worlds, A, representing a possible antecedent proposition, a world w representing the world of evaluation, and returns a world, $f(A, w)$, representing the way things *would* have gone (according w), if A *had* been true.[8] While I won't build in any further constraints, there are further constraints that the selection function might satisfy. I will discuss three important candidates. The first is that the world that would have obtained if A had been true, should be a world where A is true:

ID $f(A, w) \in A$.

The second is that if A is true at w, (i.e. $w \in A$), then w is the world that would have been true if A had been true:

MP $f(A, w) = w$ if $w \in A$.

A final condition that is sometimes imposed on the selection function is that the world that would have obtained if either A or B had been true is either the world that would have obtained if A had been true, or the world that would have obtained if B had been true:

[8]Our selection functions are slightly less general than the treatments in Stalnaker and Chellas. Both these authors allow for conditionals that are vacuously true when the antecedents fall short of a contradiction, i.e. $A \boxrightarrow \bot$ might hold when A expresses a non-empty set of worlds. Stalnaker achieves this, for instance, by introducing an impossible world. By contrast, we dispense with the impossible world by assuming that $f(A, x)$ is a regular world for any non-empty A (and leaving f undefined on empty A). The generality lost by making this simplification seems to me to be a worthy cost for avoiding the complications that arise if we do not make it, but I believe nothing of substance would change if we generalized the framework along this dimension. Like Stalnaker, but not Chellas, we also assume that $f(A, x)$ is a unique world rather than a set of worlds, but this loss of generality is a bit more intentional.

DIS Either $f(A \cup B, x) = f(A, x)$ or $f(A \cup B, x) = f(B, x)$

Readers familiar with Stalnaker (1968) may be more familiar with the condition that if $f(A, x) \in B$ and $f(B, x) \in A$ then $f(A, x) = f(B, x)$. But these conditions are equivalent to DIS in the presence of ID.[9]

A selection function determines a conditional connective. A conditional proposition, which will write $A \boxrightarrow B$, is true (at a world w), if and only if either A is empty, or else B is true at the world that would have obtained if A had obtained (at w). Or more succinctly, the conditional is true at w when $f(A, w) \in B$ or $A = \emptyset$. This means that we can identify the counterfactual proposition expressed by $P \boxrightarrow Q$ with the set $\{w \mid f(A, w) \in B\}$ when $A \neq \emptyset$ and P and Q express A and B respectively. It is easily seen that certain inferences of counterfactual logic are guaranteed to preserve truth at all worlds.

Finite Conjunction $A \boxrightarrow B, A \boxrightarrow C \dashv\vdash A \boxrightarrow (B \wedge C)$

Logical Consequents $\vdash A \rightarrow B$ whenever B is a logical truth.

Logical Equivalence $A \boxrightarrow B \vdash A' \boxrightarrow B$ when A and A' are logically equivalent.

Conditional Excluded Middle $\vdash (A \boxrightarrow B) \vee (A \boxrightarrow \neg B)$

The most contentious of these principles is the principle of conditional excluded middle. It is guaranteed by the assumption, automatically built into the formalism, that there is a maximally specific way things would have been if A. (This assumption is relaxed in some formalisms by letting $f(A, w)$ denote a *set* of worlds — the worlds that *might* have obtained in A had been true, see Chellas (1975)). While the principle has been contested on metaphysical grounds—it seems to commit us to free-floating facts—there are some powerful arguments in its favor, both from natural language data and judgments about the probabilities of conditionals, and by far the dominant view from the current literature on conditionals is that it is valid.[10] Conditional Excluded Middle also plays a crucial role, we will see, in connecting counterfactuals to to value, and natural theses about value. We will assume conditional excluded middle going forward.[11]

[9] Assume $f(A, x) \in B$ and $f(B, x) \in A$. Then $f(A, x) = f((A \cap B) \cup (A \cap \overline{B}), x) = f(A \cap B, x)$ or $f(A \cap \overline{B}, x)$ by DIS. The latter disjunct cannot hold, because $f(A \cap \overline{B}, x) \notin B$, by ID, contradicting our assumptino that $f(A, x) \in B$, so $f(A, x) = f(A \cap B, x)$. By parallel reasoning $f(B, x) = f(A \cap B, x)$, so $f(A, x) = f(B, x)$. Conversely suppose the condition stated above holds of f, and that $f(A \cup B, x) \notin A$. By ID, it follows that $f(A \cup B, x) \in B$. Also by ID $f(B, x) \in B \subseteq A \cup B$. So by the condition $f(A \cup B, x) = f(B, x)$ as required.

[10] Stalnaker (1980), Williams (2010), Mandelkern (2018), Goodman (manuscript), Klinedinst (2011), Kratzer (2020), Cariani and Goldstein (2018), Schultheis (2024), Baron-Schmitt and Schultheis (manuscript), Santorio (2022), Khoo (2022), Bacon (Forthcomingb), Dorr and Hawthorne.

[11] Those who reject conditional excluded middle but wish to vindicate the general intuition linking counterfactuals and decision must seek more complicated formulations of causal decision theory. See Lewis (1981) for instance.

It is also straightforward to see that the optional conditions ID, MP and DIS respectively guarantee the truth at all worlds of the following

Identity $A \boxrightarrow A$

Modus Ponens $A, A \boxrightarrow B \vdash B$

Disjunction $A \boxrightarrow C, B \boxrightarrow C \vdash A \vee B \boxrightarrow C$

Lastly, there is a weakening of Modus Ponens, which I will call Modest Modus Ponens, which I believe is accepted even by those who reject Modus Ponens.

Modest Modus Ponens $\top \boxrightarrow A \vdash A$

it corresponds to the condition:

MMP $f(W, x) = x$

Before we move on, let me mention two further roles that selection functions are sometimes taken to satisfy. We will not by any means treat these as given, but we will have reason to refer to them later and their interaction with other roles that we will take the counterfactual to satisfy. The first connects the conditional selection function with a notion of "similarity" between worlds:

The Similarity Role $f(A, x)$ is the most similar A-world to x

Proponents of The Similarity Role must specify which notion of similarity must be plugged into this principle in which it is remotely plausible. I will not take up this task on behalf of proponents of the similarity role, but suffice it to say it cannot be a pretheoretic notion of similarity. For instance, if the role is satisfied there must always be a unique most similar A-world to any give world, so there cannot be similarity ties or infinite descending chains of ever similar worlds. And, as Fine (1975) points out, the most "similar" world where Nixon pressed the nuclear button in the 70s is radically different from ours, given plausible judgments about counterfactuals—it is similar only in a very circumscribed and theoretical sense of "similar". Nonetheless, even without a fully fleshed out theory of similarity, this constraint has substantive implications for the logic of conditionals. For if it is satisfied by any formally well-behaved ordering relation, whether intuitively tracking a notion of similarity or not, then f must satisfy the conditions ID, MP and DIS. For the most similar A world to x is trivially always an A world, whatever we mean by "similar". The most similar world to any world x is x itself, on any formally well-behaved interpretation of "similar", so that when $x \in A$ the most similar A world to x is x itself. And the most similar $A \cup B$ world to x is either an A world or a B world, and in either case, there cannot be a more similar A (B) world to x, because otherwise we would have a more similar $A \cup B$ world too.

The other role that the conditional selection function is sometimes thought to satisfy connects it with probability. For every selection function f and a proposition A,

there is a random variable f_A standing for a randomly selected A-world, selected by a (irreducibly counterfactual) random process.[12] The result of the selection at a world x, $f_A(x)$, is just defined as $f(A, x)$, and the proposition that the world w was randomly selected, written $f_A = w$, is just the set of worlds $\{x \mid f_A(x) = w\}$. The probability role says that, with respect to some special class of probability functions, the probability that w is randomly selected is proportional to the probability of w itself. That is, for a fixed proposition A:

The Probability Role $Pr(f_A = w) \propto Pr(w)$.

The probability role is equivalent to Stalnaker's equation, that $Pr(A \boxdot\!\!\rightarrow B) = Pr(B \mid A)$, with respect to probability functions Pr in the special class. Defenders of this role must say which class of probability functions must be plugged into this role to make it plausibly true. When f is the selection function expressed by an indicative conditional relative to some evidence E, some accept the thesis with respect to the class of probability functions that are rational given E as a body of evidence.[13] If f is a counterfactual selection function expressed relative to a contextually salient time, t, the equation may be satisfied by the chances function for the time t. Stalnaker's equation also imposes substantive constraints on the logic of conditionals—Finite Conjunction, Logical Consequents, Logical Equivalence, Identity and Conditional Excluded Middle must be validated (see, for instance, van Fraassen (1973)). Moreover, it also has negative consequences: a key one is that f can*not* satisfy DIS[14] meaning that the Probability Role and the Similarity Role cannot be both satisfied at once; one or the other must at least be restricted somehow.

4 Counterfactuals and Values

It is widely thought that the two value quantities are tightly constrained by their relation to counterfactuals: the better things *would* have been if a prospect A had been true, the higher the value of that prospect. This connection can be drawn between objective value, and between subjective value independently, although if the standard theory is correct, both connections can ultimately be reconciled with the help of Expected Utility Theory.

The connection with subjective value is most familiar, and was first proposed by Stalnaker, and defended in Gibbard and Harper (1978), and has sometimes been called "causal decision" theory.[15] Informally, it says that the value of a prospect A is obtained by summing the values of each possible *outcome* — a proposition settling any matter

[12] We are assuming an abundant conception of random selection processes in which any non-constant function from worlds into a set A counts as a different process of randomly selecting something from A. The constant functions are deterministic selection processes.

[13] [ANONYMIZED]

[14] This a consequence of the result of Stalnaker (1976).

[15] The original formulation is found in a letter from Stalnaker to Lewis, printed in Harper et al. (1981). The title "causal decision theory" is contentious unless there is a connection between counter-

you might care about — weighted by how likely you think that outcome is to occur if A were to be true. Recall that identified outcome propositions with world propositions.[16] And we also identified the value of an outcome with the value of the *status quo* prospect, $\top$, conditional on that outcome $V_w(\top)$.[17] For a fixed body of evidence, E, this means:

The Subjective Value Role $V_E(A) = \sum_w P_E(A \boxright w)V_w(\top)$

We will say that a connective satisfies the *subjective value role* if the equation above holds for every subjective value function $V \in \mathcal{S}$ and propositions A, E, with the connective in question replacing $\boxright$.

Notice the equation takes for granted our assumption of conditional excluded middle: that there is a way the world would have been, w, had A been true, but that we are (typically) uncertain about which way that is. The terms $P(A \boxright w)$ thus weight the goodness $V(w)$ of each world proposition, w, by how likely you take it to be that it is the way things would have been if A had been made true.

The second connection relates counterfactuals to objective value, and has been defended recently in Bacon (2022), Williams (2023).[18] The objective value of a prospect, A, is simply the value of the outcome that *would* have obtained if A had been true. That is to say, we should have that $v_A(w) = v_*(x)$ where x is such that $A \boxright x$ is true at w. Using earlier conventions this amounts to saying that for every permissible objective value $v.(\cdot) \in \mathcal{O}$:

The Objective Value Role $v_A(w) = v_*(f(A, w))$

We will say that a selection function satisfies the *objective value role* if the equation above holds for every objective value function $v \in \mathcal{O}$ and proposition A, with the selection function in question replacing f. We will frequently take the liberty of extending this property to connectives defined by selection functions that have the property.[19]

Observe that The Objective Value Role ensures our basic constraint on objective values, Objective Outcomes. For any A and x, there exists a y, namely $f(A, x)$, such that $v_A(x) = v_*(y)$ for all $v.(\cdot) \in \mathcal{O}$. Moreover, the optional conditions MP, ID and DIS

factuals and causation, but I will not take issue with the name here. See alse Hedden (2023), Gallow (2024).

[16] A space of outcomes for an agent, represented by a value function V, may be defined as a partition of logical space $(E_i)_{i \in I}$ such that the agent is indifferent between the different ways E_i could be true: for any $i \in I$ and pair of worlds $w, w' \in E_i$, $V(w) = V(w')$. It is a consequence of The Subjective Value Role and some minimal conditions on $\boxright$ that if $(E_i)_{i \in I}$ is a space of outcomes for the agent, then $V(A) = \sum_i P(A \to E_i)V(E_i)$. The condition on $\boxright$ is simply that it be representable by a selection function; more details given in section 3.

[17] Given a very plausible decision theoretic principle (Prefer Better Outcomes from section 11), or the Identity axiom for the counterfactual, $A \boxright A$, there is an alternative definition of the value of the world w: the value of making that world obtain, $V_\top(w)$. However, without these assumptions they are inequivalent.

[18] See also Hedden (2023) §6, Ahmed and Spencer (2020).

[19] More explicitly, a counterfactual connective $\boxright$ satisfies this condition when: $v_A(x) = \alpha$ iff, for some world y, $A \boxright y$ is true at x and $v_*(y) = \alpha$.

we imposed on selection functions ensure, respectively, that v_A, so defined, satisfies the two optional conditions we imposed on objective value functions, Truth Indifference, Outcome-Prospect Compatibility and Prospect Choice.[20]

Notice that Expected Value Theory relates objective value to subjective value, The Subjective Value Role relates subjective value to counterfactuals, and the The Objective Value Role relations objective value to counterfactuals. So we have two equations identifying V_E with two potentially different quantities – it should be checked that they are all consistent. But it turns out they are. Fix a selection function f satisfying ID and MP. For every "utility function", $u : W \to \mathbb{R}$, mapping outcomes to real number values, let us define the corresponding objective value function by setting $v_A(x) := u(f(A, x))$ for every A and x, and define $\mathcal{O}$ to be the set of all such functions obtained from a utility function in this way. Moreover, given a regular probability function P and utility function $u : W \to \mathbb{R}$, we may define a subjective value function by the expectation $V_E(A) = \sum_w P_E(w) \times u(f(A, w))$ for every A and E. Define $\mathcal{S}$ to be the set of all functions V defined this way. One can show from these definitions that $\mathcal{O}, \mathcal{S}$ and f satisfy Expected Utility Theory, the Subjective Value Role, the Objective Value Role and the coherence constraints we imposed on objective and subjective values with respect to each other.[21]

5 Defining Counterfactuals by the Objective Value Role

The orthodox approach to causal decision theory is to take as primitive the counterfactual in question, as well as the agents degrees of belief, and take The Subjective Value Role to be a *definition* of subjective value in terms of counterfactuals and degrees of belief. Similarly, in Bacon (2022) the The Objective Value Role is taken as an analysis of objective value. However, it is possible to reverse the order of analysis.

This is possible because it turns out that, with a modest assumption of permissivism about what one can rationally care about, there is a *unique* conditional that satisfies The Objective Value Role. Similarly, if the space of subjective values is sufficiently rich, there is a *unique* conditional that satisfies The Subjective Value Role. In this section we treat the former.

[20]For Truth Indifference suppose that $x \in A \cap B$. Then by the Modus Ponens condition $f(A, x) = x = f(B, x)$, so $v_A(x) = v_*(f(A, x)) = v_*(f(B, x)) = v_B(x)$. For Outcome-Prospect Compatibility, note that the witness in our argument for Objective Outcomes, $f(A, x)$, belongs to A by ID. For Prospect Choice, note that $v_{A\vee B}(x) = v_*(f(A \cup B, x)) \in \{v_*(f(A, x)), v_*(f(B, x))\}$ because $f(A \cup B, x) \in \{f(A, x), f(B, x)\}$ by DIS.

[21]For instance, The Subjective Value Role holds. Suppose V is determined by P and u in the way described. Then $V_E(A) = \sum_w P_E(w) \times u(f(A, w)) = \sum_w P_E(A \boxdot\!\!\rightarrow w) \times u(w) = \sum_w P_E(A \boxdot\!\!\rightarrow w) \times V_E(w)$. The last identity holds because ID ensures that $f(\{w\}, x) = w$, and so $V_E(w) = \sum_x P_E(x) \times u(f(\{w\}, x)) = u(w)$. The Objective Value Role also holds because MP guarantees that $f(A, w) = f(\top, f(A, w))$, and so $v_A(w) = u(f(A, w)) = u(f(\top, f(A, w)) = v_\top(f(A, w))$. Expected Utility Theory is immediate from the definitions of V and v in terms of u.

Let $\mathcal{O}$ denote the set of permissible objective values. Every $v.(\cdot) \in \mathcal{O}$ should therefore satisfy Truth Indifference and Outcomes. We will assume a form of *permissivism* about what one can rationally care about. I will state both a strong and weak form, although we will only need the weak form:

Strong Permissivism For any $u : W \to \mathbb{R}$, there is a $v.(\cdot) \in \mathcal{O}$ such that $v_\top = u$ (i.e. $v_\top(w) = u(w)$ for all $w \in W$).

Weak Permissivism If $v_\top(x) = v_\top(y)$ for every $v.(\cdot) \in \mathcal{O}$, then $x = y$.

Strong permissivism tells us that we can care as much or as little about anything we like: *any* function from worlds to values is the utility of some possible rational agent. Weak permissivism merely tells us that any two outcomes can be valued differently by some possible rational agent. One might think that it is irrational to both assign a very high value to being 6ft exactly, and to highly disvalue being 6.00001ft, other things being equal, but not irrational to care about the former a tiny bit more than the latter, thus refuting strong but not weak permissivism. While strong permissivism is overkill for what we are trying to show, I suspect at least some of the sorts of reasons one might have to reject strong permissivism extend to weak permissivism so it is worth having both in mind.[22]

Theorem 5.1. *Suppose that the set of objective value functions $\mathcal{O}$ satisfies Weak Permissivism (on top of Objective Outcomes). Then there is a unique selection function f that satisfies the The Objective Value Role:*

$$v_A(x) = v_*(f(A, x))$$

for all $v \in \mathcal{O}$. (Recalling that we write $v_*(x)$ for $v_\top(x)$.)

Or in other words, there is a unique counterfactual that satisfies the objective value role. A proof of this can be found in appendix A.

6 Defining Counterfactuals by the Subjective Value Role

Some philosophers are skeptical of objective value, at least as it is understood here. Broome, for instance, maintains that all value is subjective value (Broome (2004) chapter 6). Expected utility theory can be maintained at least verbally, because subjective value can be thought of as the expected value of itself. But recall that our principle Objective Outcomes implies that whatever objective value is, i is a quantity that is independent of our subjective degrees of belief. The skeptic of objective value might take the view that there is no quantity that is *independent* of our subjective degrees of

[22] [ANON]Anonymous footnote.

belief of which subjective utility is an expectation of.[23] Or one could take the position that such quantities could be cooked up, but are not theoretically important.

I therefore offer another reduction of counterfactuals to value, this time, to subjective value. This second reduction is of independent interest, for it gives us an alternative perspective on the first. But this reduction will also allow us to bring on board skeptics of objective value. The objective value cynic may take subjective value as the basic moral quantity, and still be able pin down the counterfactual by its interaction with subjective value.

Recall that $V_E(A)$ is a measure of the subjective value of a prospect A, given a body of evidence E, and $\mathcal{S}$ denotes the set of all rational subjective value functions. Recall also that we formulated some basic properties that the space of rational subjective values $\mathcal{S}$ might satisfy. Crucially these properties can be stated without reference to objective values, and without reference to counterfactuals: they are simply internal conditions that a set of subjective value functions can satisfy, which are internally motivated by intuitions directly about subjective value. Permissivism about how you value outcomes in the Subjective Value setting is formulated as follows:

Strong Permissivism For any $u : W \to \mathbb{R}$, there is a $V_{\cdot}(\cdot) \in \mathcal{S}$ such that $V_{\cdot}(\top) = u$ (i.e. $V_w(\top) = u(w)$ for all $w \in W$).

Weak Permissivism If $V_x(\top) = V_y(\top)$ for all $V \in \mathcal{S}$, then $x = y$.

Given any space of values satisfying the coherence constraints and Weak Permissivism, there will be a unique selection function that satisfies The Subjective Value Role,

Theorem 6.1. *Let $\mathcal{S}$ be a space of subjective values satisfying the coherence conditions Evidence and Priors, Evidence Linearity, and Subjective Outcomes, and the assumption Weak Permissivism. Then there is a unique selection function, $f_{\mathcal{S}}$, satisfying The Subjective Value Role. That is, for any value $V \in \mathcal{S}$*

$$V_E(A) = \textstyle\sum_w P_{V_E}(A \boxdot\!\!\to w \mid E) V_w(\top)$$

where $\boxdot\!\!\to$ is the connective determined by $f_{\mathcal{S}}$, and P_{V_E} is the probability function determined by V_E, $P_{V_E}(A) = \frac{V_\top(\top) - V_{\neg(E \wedge A)}(\top)}{V_{E \wedge A}(\top) - V_{\neg(E \wedge A)}(\top)}$

7 A Reductive Account of Counterfactuals?

While the Objective and Subjective Value Roles are usually read as furnishing us with a definition of objective or subjective value in terms of counterfactual implication, the theorems we have just stated raise the possibility of reversing the order of analysis,

[23] Bacon (2022) shows that Jeffrey's decision theory has this feature: the news value of a prospect is not the expectation of any crecence independent quantity. Although news value can be expressed as the expectation of credence dependent quantities — indeed there are many such quantities, including itself, or the quantity of "praxic good" discussed in Konek and Levinstein (2019).

extracting from our results a definition of counterfactuals in terms of value. Perfectly reductive analyses are rare, however, and there is a spectrum of positions you could take about the web of analysis and the positions of value and counterfactuality in that web. In this section I will examine the significance of our result for the circle of analysis. I will begin by addressing some objections to the idea that value could be *prior* to counterfactuality, and then I will argue, more modestly, that value is more central and more tightly connected within the circle of analysis to counterfactuality than other notions that have sometimes been appealed to in this endeavor, such as similarity and probability.

Objection 1. The point of a decision theory, the objection goes, is to deliver us with verdicts concerning what we are supposed to do. The two value roles achieve this by assuming that we already know, or know how likely it is, that various things would happen if A were true, and on this basis delivers verdicts about the values of propositions and recommendations concerning what to make true. One might object that if we reverse the order of analysis — take values as our starting point and use them to determine the facts about what would happen if A — it looks as though a decision theory, given by the The Subjective Value Role or by the Objective Value Role, cannot serve the advisory role just outlined, which requires one to have views about the counterfactual facts prior to the value facts.

This line of reasoning appears to be premised on the idea that the two value roles should do more than simply describe a relation between a rational agents values, utilities and credences in counterfactuals, but that it should rather provide instructions that an agent can follow to *get* the value of an action proposition *from* their utilities and credences in counterfactuals.[24] But even granting that the equation is not merely descriptive and is more like a recipe for determining value that describes rational psychological processes, I think the objection is easily met by the simple observation that the order of metaphysical analysis need not coincide with the order in which we employ the concepts in our practical reasoning. The fact that we often first figure out which good or bad things would happen if A were true in order to figure out the value of A has no straightforward bearing on the direction of analysis. To illustrate, one could easily imagine someone who has learned to identify peacocks and then independently to identify peahens — two kinds of birds that look quite dissimilar. If they are then told that a peafowl is a peacock or a peahen they can use their prior ability to identify peacocks and peahens to figure out which birds are peafowl: first figure out whether the bird is a peacock or a peahen, and use that to decide whether it is a peafowl. The order in which we came to grasp the concepts, and the order in which they appear in our reasoning here has no bearing on the order of metaphysical analysis. Presumably a peacock is metaphysically defined as a male peafowl, and a peahen as a female peafowl (cf. the standard example of an analysis of a vixen as a female fox). Indeed, it is hard to see how any claim about the order of analysis could possibly prevent one from

[24] [There are some quite general reasons to be skeptical of this way of reading the equation.] note on Williamson and operationalizing epistemology.

moving back and forth between value and counterfactuals in any order in our practical reasoning.

Objection 2. The value-first approach to counterfactuals rests on the notion of objective value (and perhaps by extension subjective value). But this notion is not in good standing because it often involves assigning objective values to false prospects in a way that seems to float free from the physical facts, or indeed, respectable facts of any sort.[25] Suppose that we have made a bet of a dollar on the flip of a particular coin: I will lose a dollar if it lands heads, and gain one if it lands tails. Accordingly, the objective value of this bet is either \$1 or \$-1. But now suppose that I do not accept the bet, and consequently the coin is not flipped—indeed, it is never flipped again. Then what determines which of the two possible objective values the bet has? If we are living in an indeterministic world it does not seem to be something that can be grounded in the past, present or future physical aspects of the coin or the environment.[26] Yet it presumably has to be one or the other.

The objection here is, I think, also far from decisive. It has been established that the objective value of a prospect is sometimes a vague or indeterminate matter. But I think this is not grounds for either rejecting the notion, or dismissing it as a basis for understanding counterfactual facts. After all, most concepts are vague, and consequently most concepts that figure in analysis have cases where it is indeterminate whether they apply. And in the case of value statements, in particular, there are many strategies for making sense of their non-factuality, such as expressivism or relativism. According to the former, for instance, value statements are not to be identified with expression of beliefs, or to be evaluated for correctness in the same way as beliefs, but rather with the expression of desires or some other non-cognitive attitude.

Indeed, if the objection were good it would apply equally to the reverse analysis, that takes counterfactuals first and attempts to analyse objective and subjective value in terms of them, along with degrees of belief and desire over outcomes. For given conditional excluded middle counterfactuals are just as mysterious and indeterminate as objective values in these cases. For consider the coin of the previous example that's never been flipped, and will never be flipped either. I may ask, nonetheless, whether it *would* have landed heads or tails if it had been flipped. The law of Conditional Excluded Middle states that either the coin would have landed heads, or that it would have landed tails, if it had been flipped. But inspection of the coin and its environment don't seem to offer any clue as to which of these two counterfactuals is true. Objective value facts are in no worse standing than the counterfactual facts given the law of Condition Excluded Middle.

Objection 3. One cannot analyse counterfactuals in terms of value on the grounds that value facts are "thin" facts, and that counterfactual facts are "thick" and somehow more objective ("worldly", "metaphysically substantive", etc.), and that reductions

[25] The following example is discussed in Bacon (2022).

[26] And even if we are living in a deterministic world, where a very exact specification for flipping a coin might have its result determined by the momenta and positions of the air particles, we have not specified how the coin is to be flipped in sufficient exactness.

cannot start with the thin and end up with the thick.

There is, however, a venerable tradition, following Adams (1965), of taking conditional facts to be thin in something like the way that value and other moral facts are supposed to be. Expressivists about some subject matter who are happy to make sense of the propositions of that subject matter will often explain the metaphysical "thinness" of those propositions by their having an epiphenomenal role in thought. For indicative conditionals, they say that having a degree of belief in a conditional proposition $A \to B$, where A and B are ordinary propositions, amounts to nothing more than having a certain pattern of degrees of belief towards "thick" (i.e. categorical) propositions appearing in consequent and antecedent position, A and B.[27] It is completely determined by the ratio of of the degrees of belief in two thick propositions, $A \wedge B$ and A. This sort of strategy is of a piece with other approaches to thin propositions. Gibbard has suggested that "thin" moral propositions can be similarly elucidated by their doxastic role; a belief in a moral proposition is a deliberative state with respect to a "thick" non-moral proposition.[28] Some have thought that in order to employ the notion of objective chance requires one to make a substantive metaphysical posit. But Lewis (1980) says that we can make sense of the notion of objective chance by way of its role in thought, by saying what it is to have a credence in a chance proposition. Others have made similar claims with respect to epistemic modals, and vagueness.[29]

Expressivists about conditionals are usually concerned with indicative conditionals. However, we saw above that, like indicatives, there is often no fact of the matter about counterfactuals with false antecedents. It would, at any rate, be *prima facie* very puzzling if there was both a distinction between thick and thin propositions and that indicative and subjective conditionals fell on opposite sides of it.[30] For both sorts of conditional share a syntax, and logic, and across most languages are expressed using the very same form of words.

Objection 4. There is not enough distance between counterfactuality and the notion of objective value (and thus, perhaps, subjective value) for the latter to provide an informative analysis of the former. For instance, there is very little difference between figuring out what the objective value of a bet is, and figuring out how good things *would be* if the bet is accepted.

Here, of course, we teeter at the brink of the paradox of analysis. Perhaps the lesson of that paradox is just that some circles of analysis are more informative, central, or theoretically fruitful than others. In this respect, we can at least make some comparisons with other theories of conditionals.

Consider, first, theorists who make extensive use of a notion of similarity in their theorizing about counterfactuals. These theorists typically accept the principle we earlier called the The Similarity Role, or something like it, and maintain that a counterfactual $A \mathrel{\square\!\!\to} B$ is true when the *most similar* A-world to actuality is a B-world

[27]This tradition is includes, for instance, Edgington (1986), Appiah (1984), Bennett (2003).

[28]Gibbard (2003) , Schroeder (2011).

[29]See, e.g., Schulz (2010) , and Bacon (Forthcominga) respectively.

[30]Notwithstanding the positions of Edgington (1995), Lewis (1976), and others to the contrary.

(Stalnaker (1968)).[31] These principles merely state a relationship between counterfactuality and similarity: one might take the stronger position that counterfactuality can be *reductively analyzed* in terms of similarity. But this stronger position about the direction of analysis has received much resistance, and is typically disavowed by the similarity theorists.[32] As noted in section 3, the notion of similarity at issue is not a pretheoretic one, but rather one that is somewhat doctored to fit the analysis. Perhaps the easiest way to arrive at it is by reverse engineering our judgments about counterfactuals (Lewis (1973) pp54,63, Stalnaker (1984) p126-7). Indeed, the circle of analysis can be closed in this manner, for the operative notion of similarity admits a definition in terms of counterfactuals: a world x is more similar (to actuality) than y, in the operative sense, when it is x, not y, that would have obtained if one of them had obtained. Proponents of similarity based theories of counterfactuals typically do not take themselves to have given a reductive account of counterfactuality in terms of similarity, but they do accept that counterfactuals satisfy The Similarity Role, and this gives important information about the selection function.

A more recent theory of conditionals understands conditionals in terms of "random selection": $A \boxdot\!\!\rightarrow B$ is true when a certain A-world that was randomly selected is a B-world (Schulz (2017), Bacon (Forthcomingb)). There are, however, many different ways to randomly select something — consider different mechanisms for picking a number at random (dice, tickets in a hat, etc). The Probability Role constrains us to selection processes such that: $Pr(f_A = w) \propto Pr(w)$ for every probability function Pr in a suitable class, which, recall, is equivalent to "Stalnaker's thesis" for probability functions in the class. But even that doesn't fully pin down the selection mechanism: if there is at least one process then there are typically many selection processes satisfying this constraint. Bacon (Forthcomingb) takes the mechanism to be irreducibly conditional—not something that we had independently of our notion of conditionals. So The Probability Role cannot provide us with a reductive analysis, whatever that means, but it does widen the circle of analysis: it is a principle connecting conditionals to other concepts like rational degrees of belief and chance.[33]

The connection between counterfactuality and value does better than either of the constraints discussed above. First, the notion of objective value and subjective value that we are using here is the pretheoretic one. If we grant, as we have been, that causal decision theory in the style of Gibbard and Harper (1978) is, at minimum, a materially correct account of the pretheoretic notion of value, we obtain, also at minimum, a materially correct definition of counterfactual implication in terms of a pretheoretic notion. In the web of analysis, value is connected with all sorts of key ethical concepts. While the *pretheoretic* notion of similarity may have similarly wide ranging connections,

[31] Alternative similarity analyses include Lewis (1973) and Pollock (1976); not very much rests on which we use for this discussion.

[32] Stalnaker (1984) chapter 7, Lewis (1973) §4.3.

[33] Schulz (2017) chapter 6 suggests reducing selection to the notion of arbitrary reference in Breckenridge and Magidor (2012). I predict, however, that many will find this no less obscure than a primitively counterfactual notion of random selection.

the highly theoretical notion of counterfactual similarity, by contrast, does not seem to, and so The Similarity Role has fewer wider implications than the The Objective or Subjective Value Roles. Second, both of the two value roles completely pin down the conditional by its connection to value. While The Probability Role does better by partially pinning down conditionals to other theoretically central pretheoretic concepts, such as rational degrees of belief or chance, it is only a partial characterization.

An advantage that has often been attributed to the similarity theory of conditionals is that, even if it does not provide us with a reductive analysis, the existence of a necessary and sufficient conditions for a counterfactual in terms of *any* well-behaved notion of similarity of the sort described above, formally constrains the logic of the counterfactuals as discussed in section 3. But such implications are not unique to the Similarity Role: both the value roles, and the Probability Role discussed above also tightly constrain the logic of conditionals, and may do so in a way that is incompatible with the constraints imposed by the similarity analysis. Recall that The Similarity Role implies the condition DIS, and the validity of Disjunction, a principle whose validity is incompatible with The Probability Role. Moreover, in section 11 below, we will see that the validity of Disjunction corresponds to a principle of decision theory: a principle which, if rejected, would require this inference to be invalid if we are to maintain the link between objective/subjective value and counterfactuals.

8 Contextualism and the Conditional of Deliberation

On what I am calling the 'standard' way of reading the two value roles, the relative priority of counterfactuals and value both in analysis and in practical reasoning begins with counterfactuals and ends with value. In the previous section I resisted the first claim about the order of analysis. However the second idea, that judgments of counterfactuals enter into our practical reasoning before judgments about objective value, also requires scrutiny.

It is a widely acknowledged fact that counterfactual sentences are context sensitive. The sentence 'if A were the case then B would be the case' can express one proposition when uttered in one context, and a completely different proposition in another context, even when A and B are not themselves context sensitive. It follows from a natural compositionality assumption that there are lots of propositional connectives that can be expressed by the English counterfactual conditional in different contexts. What happens if plugging one of these into The Subjective (or Objective) Value Role yields the recommendation to take one course of action, and plugging in another yields the recommendation to take a different course of action? Presumably there is at most one set of permissible actions, and so at most one of the many propositional connectives expressed by the English counterfactual yields the right values when plugged into the

relevant role.[34] Let us call this connective *the conditional of deliberation*. If there is a unique such conditional, however, how do we get a fix on it?[35]

The possibility of inconsistent verdicts across contexts is not merely theoretical. Consider the following scenario:

> An evil scientist has strapped me to an electric chair. My wife must press one of two buttons: button A or button B. One of these buttons will kill me, the other will do nothing. I do not know which button does what, but I know that my wife does.

Since my wife and I are on good terms—she doesn't want me dead—she is not going to press a button that she knows will kill me. I think I can be confident in the following counterfactual:

> If my wife were to press button A, then it would not kill me.

The context of the above assertion is one in which I am paying special attention to what my wife would do given various facts about her knowledge and our relationship. (Those sympathetic to similarity analyses of the counterfactual might describe this as a context where the relevant similarity relation is based on what is normal behaviour for my wife — the worlds where she is out to kill me are very distant.)

On the other hand, suppose we prime ourselves with a different version of the same story before making our judgment:

> An evil scientist has strapped me to an electric chair. My wife is presented with two buttons, A and B, and told what they will do. There is a disconnected circuit connecting the chair and a 10000 volt power source, and one, but not the other, of these buttons is attached to the circuit and will close it if pressed. The mechanism is completely foolproof: whichever of the two buttons is attached to the circuit will kill me if it is pressed!

[34] This follows from the the uniqueness portion of theorems 5.1 and 6.1, and the possible worlds framework. By relaxing the latter assumption, we could end up with multiple necessarily equivalent connectives satisfying the objective and subjective value roles. But the context sensitivity phenomena implies the existence of many connectives that can be expressed by the English conditional with different extensions.

[35] There is one style of response to this sort of argument that is worth mentioning briefly. It relies on the observation that the word 'ought' is also context sensitive, and so one one might postulate that for each resolution of the context sensitivity of 'ought' there is a corresponding notion of value in which the proposition you ought to make true, and the action proposition with maximal value line up. One might even go as far as to maintain that the context sensitivity in deontic modals and value march in lockstep with the context sensitivity in counterfactuals, and do so in a way that makes The Subjective Value Role comes out true in every context. However this idea does not appear to be born out by actual data — below we'll consider a contextual resolution of the counterfactual which, when plugged into The Subjective Value Role, would yield recommendations that you clearly shouldn't follow, on any reasonable resolution of 'shouldn't'.

I think if we now ask ourselves what would happen if my wife were to press button A, a different judgment arises, in which we are uncertain whether it would kill me or not. After all, we don't know whether button A or B is connected to the circuit. The second description gives us more detail about the mechanism, but this cannot on its own explain why we find the counterfactual more likely in the first case—we didn't need that detail to know that it would kill me if pressed, we took it on trust. When the mechanism is emphasized like this, then I think the counterfactual 'If my wife were to press button A, then it would not kill me' expresses a proposition that I am 50-50 on because I'm simply unsure whether button A is wired.[36]

What value does the proposition that my wife presses button A get according to The Subjective Value Role when we use these two different interpretations?[37] Well, on the interpretation where I'm certain that if she were to press button A, it would not kill me, then the value of this proposition according to The Subjective Value Role will not be especially worse than the status quo. Whereas if I'm 50-50 in this counterfactual, the value of this proposition is very low — it is equivalent to taking a gamble that has a 50-50 chance of death, and no positive upshot. Now it can't be that both of these are the value of this proposition — only one of the two counterfactual connectives expressed in these two contexts is the right one to plug into The Subjective Value Role. And I think it is obvious that the value predicted by the first interpretation is absurd: the right counterfactual to use is the one involved in the second context.

Considerations such as these indicate that we cannot figure out which connectives may be plugged into The Subjective Value Role in complete isolation of considerations of value. We arrived at our verdict about which connective was the right one to use in our practical reasoning by looking at which delivered the right sort of value.

So, once we have recognized that context sensitivity gives us a proliferation of counterfactual connectives, we have another reason to be interested in the results of section 5 and section 6, even if we ultimately reject the application of these results for reductive purposes. We have argued that there exists a special counterfactual connective that satisfies The Subjective Value Role, *the counterfactual of deliberation*. But we have also shown that this connective is not uniquely determined by the way in which English counterfactual sentences are used. How, if not through natural language, do we latch on this counterfactual? We have shown that the counterfactual of deliberation has a rich role in thought that is independent of the English counterfactual words.

[36]Contrast with the case described in Jackson (1977). See also Lewis (1979), Downing (1959).

[37]Note that the proposition that my wife presses the button A is presumably not a proposition I am in a position to make true or false — it is not an 'action proposition'. But it still receives a *value* according to causal decision theory which is related to concepts such as preference, or the degree to which you'd like a proposition true, even if less directly related to concrete action in this case.

9 The Methodology of Conditional Logic

The conventional methodology for figuring out if an inference involving conditionals is valid is to find instances of the inference involving natural language conditional sentences, and evaluating the premises and conclusion for truth or falsity. However, applying the standard methodology can be quite a delicate matter. By examining the role of counterfactuals in thought we can also shed some light on these tricky issues.

Let me illustrate. There has been a long-standing debate in the literature on counterfactuals about the validity of the inference Antecedent Strengthening: the inference $A \boxrightarrow C \vdash A \wedge B \boxrightarrow C$. It is a central principle because it is accepted by material and strict accounts of the conditional, but rejected by the probabilistic and similarity accounts.[38] The *prima facie* case against it is based on the judgments that 1 and 2 are true.

1. If the match had been struck, it would have lit.

2. If the match had been soaked in water and struck, it would not have lit.

However, if Antecedent Strengthening were valid, 1 would license the inference to

3. If the match had been soaked in water and struck, it would have lit.

which seems to be incompatible with 2 in this context. However, many people have noted, in defence of antecedent strengthening, that these judgments are subject to order effects. If we evaluate this pair of conditionals in a different order—that is we initially assert 2, and then consider 1—some people are less inclined to outright assert that 1 is true.

2. If the match had been soaked in water and struck, it would not have lit.

1. If the match had been struck, it would have lit.

There is a large literature trying to figure which of these judgments is the correct one — with some theories (e.g. Von Fintel (2001), Gillies (2007a)) arguing that Antecedent Strengthening is valid, and others (e.g. Moss (2012), Boylan and Schultheis (2017)) arguing it is not. Some will attempt to explain the differing judgments by appealing to a contextual shift in one of the two pairs. Others working in the dynamic tradition have sometimes gone as far as to reject the notion of a valid argument that is independent of the order of the premises altogether. But in my view the debate seems to be at an impasse insofar as the linguistic data is concerned.

Order effects are just the tip of the iceberg. Similar debates exist about most of the central principles of conditional logic: simplification of disjunctive antecedents (the inference $(A \vee B) \rightarrow C \vdash A \rightarrow B$), or-to-if ($A \vee B \vdash \neg A \rightarrow B$), Import-Export

[38] Proponents of antecedent strengthening then include: Jackson (1987), Williamson (2020), Von Fintel (2001), Gillies (2007b). Opponents include: Lewis (1973), Stalnaker (1968), Adams (1965).

$(A \to B \to C \dashv\vdash A \wedge B \to C)$, CSO $(A \leftrightarrow B \vdash (A \to C) \equiv (B \to C))$, and even modus ponens.[39] A standard move in these debates appeal to the context sensitivity of conditionals to explain away the apparent validities and the apparent invalidities that conflict with the preferred semantic theory. This has lead some philosophers and linguists to reject the discipline of conditional logic altogether as resting on a mistake, to be replaced instead by a quasi-pragmatic relation that holds when asserting some premise A in a context c, puts one in a context c' in which it is OK to assert the conclusion C (see, e.g., Gillies (2011)).This relation is a far cry from the inferential relations usually studied in logic, which satisfy basic properties like reflexivity ($A \vdash A$), and transitivity (if $A \vdash B$ and $B \vdash C$ then $A \vdash C$).[40]

Against these pronouncements, I would like to defend what I think of as the orthodox view of conditional logic. Logic is concerned with the inferential relationships between language independent propositions. The inferential relationships between *sentences* is complicated by the presence of context sensitivity and dynamic phenomena, and thus so is the relationship between sentential validity and our judgments of truth and falsity. By contrast, the notion of propositional entailment is uncomplicated. One proposition entails another when, necessarily, if the one is true, so is the other (in the broadest sense of 'necessarily', whatever that might be). Once we have a definite propositional connective, C, in mind we can ask what the logic of that connective is — for instance, it satisfies Antecedent Strengthening if the proposition Cpq entails $C(p \wedge r)q$. Propositions are language independent entities: they are not the sorts of thing that can be context sensitive, and a single proposition can be thought by a wide range of different beings in different languages, and with different psychological makeup. Non-human animals can believe conditional propositions without having any linguistic competence at all. The fox might believe that the river rat would come out of hole in the river bank, if the water were to get to a certain height. The fox can evidently make this judgment about the truth or falsity of a conditional proposition without first making judgments about the truth or falsity of any English, or other natural language, sentences. Consequently questions of propositional entailment may in principal be settled without first figuring out how exactly the propositions expressed by conditional sentences depend on context.

I believe the picture I have just articulated reflects a common conception, going back to C.I. Lewis, of what the subject matter of conditional logic is, and that it is within this framework that Lewis (1973) chapter 6, Stalnaker (1968), Chellas (1975) and others have raised and settled questions about the soundness and completeness of various conditional logics with respect to semantics. These logics are not supposed to be sound with respect to the set of sentences accepted in any given conversation.

Skeptics of conditional logic will no doubt object that, even assuming that natural

[39]

[40]Kratzer famously says, in various places, that there is no such thing as a two-place conditional connective in natural language; see e.g. Kratzer (2012) chapter 4, which gives us a different way in which one might think that conditional logic rests on a mistake from broad considerations about conditional language.

language conditionals express binary conditional connectives in certain contexts, our access to the logical properties of these connectives must come by way of judgments about the felicity of natural language sentences that express them, if they are to come from anywhere. Indeed, it may be worse than a mere epistemological problem about how we can obtain access to facts about conditional connectives, but a problem of dilineating the scope of the discipline of conditional logic in the first place. The skeptic may concede that the logical properties of a certain connective C, is independent of the behaviour of the language that expresses that connective. However, there are many connectives, whereas we are only interested in the logic of conditionals. Surely, they will object, our way of distinguishing between conditional connectives and other connectives comes by way of their relationship to conditional language?

In my view, however, the wider role of conditionals vastly outstrips its relationship with natural language. They have a rich role in non-linguistic thought, through their relation to degrees of belief, chance, causation and deliberation, and these roles tightly constraint their logic. Perhaps this role could replace the standard appeal to natural language judgments?

To illustrate, consider again Stalnaker's thesis. We observed in section 3 that it has many implications for the logic of conditionals. It ensures the validity of Identity, Finite Conjunction, Logical Equivalence, and Conditional Excluded Middle. As well as ensuring certain inferences, this role also makes strong predictions about which inferences are *invalid*. Returning to our example of antecedent strengthening, it is clear that the conditional probability of C on A can be high while the conditional probability of C on $A \wedge B$ low (see figure).

For instance, the probability of a match lighting conditional on being struck can be high, while also low conditional on being both struck and wet. We thus come down in favour of the invalidity of antecedent strengthening without having to rely on the initial set of judgments regarding the felicity of sentences 1 and 3. One can evaluate these conditional probabilities without evaluating any natural language conditionals; one just needs to have a credence in the proposition that the match has been struck and lit, and a non-zero credence in the proposition that the match has been lit in order to have a conditional credence in the proposition that the match lit conditional on it being struck. The link to conditional probability also has some less obvious consequences. For instance, we mentioned previously that in its unrestricted form, it requires Disjunction to be probabilistically invalid in the sense that one can be more rationally confident in the conjunction of the premises than the conclusion (although it may have other good probabilistic statuses, such as preserving rational certainty.)

10 Counterfactual logic and value

Given the Subjective and Objective Value Roles, it can be shown that each principle of counterfactual logic corresponds to a principle of decision in the following sense: the principle A holds at every possible world if and only if the corresponding decision theoretic principle holds. This gives us a route to settling questions about the logic of conditionals that does not rely on linguistic judgments. In this section we'll describe the general recipe for getting from a conditional logic to a decision theory, and from a decision theory to a conditional logic. The general correspondence does not, however, deliver perspicuous descriptions of decision theories. In the next section, we'll look at some of the decision theoretical principles corresponding to different principles of conditional logic, and describe them in more intuitive preference theoretic terms.

We start by describing a correspondence between objective decision theories and conditional logic. A decision theory can equivalently be described by a set of coherent preferences, or by a set of real valued measures that respect the preferences; I will take the latter route for technical simplicity.[41] By an objective decision theory we will mean a space of permissible value measures, $\mathcal{O}$, over a set of world W, assigning each possible prospect (subset of W) a real numbered value at each world. A decision theory in this sense tells us, in quantititive terms, how objectively good various actions are in different circumstances.

Definition 10.0.1. *$(W, \mathcal{O})$ is decision theory when $\mathcal{O}$ is a set of "value" functions $v : P(W) \times W \to \mathbb{R}$, satisfying the following conditions.*

Objective Outcomes *For every $A \subseteq W$ and $x \in W$, there exists $y \in W$ such that $v_A(x) = v_\top(y)$ for every $v_\cdot(\cdot) \in \mathcal{V}$.*

Weak Permissivism *If $v_\top(x) = v_\top(y)$ for every $v \in \mathcal{O}$ then $x = y$.*

The language of propositional conditional logic is the smallest set containing an infinite stock of sentence letters, and containing the sentences $(A \to B), \neg A$ and $(A \wedge B)$ whenever it contains A and B. Given a decision theory, $(W, \mathcal{O})$, an interpretation function for that decision theory is a mapping from sentences to subsets of W satisfying the following constraints.

- $[\![A \wedge B]\!] = [\![A]\!] \cap [\![B]\!]$
- $[\![\neg A]\!] := W \setminus [\![A]\!]$
- $[\![(A \to B)]\!] := \{x \mid f_{\mathcal{O}}([\![A]\!], x) \in [\![B]\!]\}$.

where $f_{\mathcal{O}}$ is the unique function such that

[41] A coherent set of objective value preferences can be identified with the set of relations $A \preceq_w B$ between pairs of prospects at worlds, defined by the condition $v_A(w) \leq v_B(w)$, where v is in set $\mathcal{O}$ of real valued objective value measures of some numerically described decision theory.

$v_A(x) = v_\top(f_\mathcal{O}(A, x))$ for all $v \in \mathcal{O}$

i.e. $f_\mathcal{O}(A, x) = y$ iff $v_A(x) = v_\top(y)$ for all $v \in \mathcal{O}$.[42]

Definition 10.0.2. *Given a conditional logic* $\mathbf{L}$*, we say that an objective decision theory,* $(W, \mathcal{O})$*, is a* $\mathbf{L}$*-decision theory when* $[\![A]\!] = W$ *for every sentence* $A \in \mathbf{L}$ *and interpretation function* $[\![\cdot]\!]$ *satisfying the above constraints.*

Conversely, given an objective decision theory $(W, \mathcal{O})$*, the conditional logic of that decision theory is the set of sentences* A *such that* $[\![A]\!] = W$ *for all interpretation functions* $[\![\cdot]\!]$*.*

Pleasingly we obtain nice correspondences between our three optional principles of conditional logic our three optional principles about objective value:

Proposition 10.1.

- $(W, \mathcal{O})$ *is a Modus Ponens decision theory just in case it satisfies Truth Indifference.*
- $(W, \mathcal{O})$ *is an Identity decision theory just in case it satisfies Prospect-Outcome Compatibility.*
- $(W, \mathcal{O})$ *is a Disjunction decision theory just in case it satisfies Prospect Choice.*

Proofs may be found in appendix A. Having set this up, it is natural to ask what the logic of the class of all decision theories is. One might ask whether it is the same as the logic of arbitrary selection functions, which can be given a straightforward axiomatization.[43] The answer is not quite, at least given the simplifying assumption of identifying outcome value with the prospect value of the status quo (Outcome Value is the Status Quo). For then the condition MMP, $f_\mathcal{O}(W, x) = x$, holds trivially because $v_\top(w) = v_\top(w)$ for every $v \in \mathcal{O}$. Thus Modest Modus Ponens is valid for all decision theories considered, but not. However, if we adopt the official position of taking an objective value to be an ordered pair $(v.(\cdot), u(\cdot))$ of a prospect value and an outcome value, the logics of arbirary decision theories and selection functions coincide. Indeed, identifying $\mathcal{O}$ with such a set of ordered pairs we obtain the following

Proposition 10.2. $(W, \mathcal{O})$ *is a Modest Modus Ponens decision theory just in case it satisfies Outcome Value is the Status Quo.*

[42] Objective Outcomes ensures the existence of a y meeting the condition on the right-hand-side of this biconditional, and Weak Permissivism ensures its uniqueness.

[43] In this paper selection functions are somewhat more restricted than the general class found in, say, Chellas (1975), because they pick out a *unique* world, $f(A, x)$ for *every* non-empty proposition A. The uniqueness part ensures that CEM is valid, and being defined on every non-empty set A ensures that counterfactual necessity, defined as $\Box A := (\neg A \mathrel{\Box\!\!\rightarrow} \bot)$, has a logic of $\mathsf{S5}$. If we add both these things to the basic logic CK of Chellas (1975) we have an axiomatization of the class of selection functions described here.

Similar points apply to the completely parallel correspondence that holds between subjective decision theories and conditional logics. I describe that correspondence now without discussion.

Definition 10.2.1. *A decision theory $(W, \mathcal{S})$ is a space of subjective values $\mathcal{S}$ containing functions $V : P(W) \times P(W) \to \mathbb{R}$, satisfying the coherence conditions Priors and Evidence, Evidence Linearity, and Subjective Outcomes, and Weak Permissivism.*

We define a selection function $f_{\mathcal{S}}$ as the unique function such that

$$V_x(A) = V_{f_{\mathcal{S}}(A,x)}(\top) \text{ for all } V \in \mathcal{S}.$$

i.e. $f_{\mathcal{S}}(A, x) = y$ iff $V_x(A) = V_y(\top)$ for every $V_{\cdot}(\cdot) \in \mathcal{S}$, and we define the notion of an interpretation function as above, and introduce the notion of a **L**-decision theory and the conditional logic of a subjective decision theory as above.

11 Preference and Conditional Logic

The correspondence between conditional logics and decision theories described in the previous section is very general, but not very perspicuous. We wish to know, in terms of intuitive principles about preferences, what the decision theoretic upshots of specific principles of conditional logic are. I will discuss here the three principles we have singled out: Identity, Modus Ponens, and Disjunction.

Subjective decision theories are usually presented as theories of preferences (Jeffrey (1990), Bolker (1967), Joyce (1999), Savage (1954)). We will say that a prospect B is preferred to another A, on evidence E when $V_E(A) < V_E(B)$. We will write this $A \prec_E B$, $A \preceq_E B$ for the notion of non-strict preference, and $A \approx_E B$ for the associated notion of indifference. A proposition A is an *outcome* for an agent iff it settles everything the agent cares about: i.e. the value of the *status quo* on any piece of evidence stronger than A is the same. $V_{A'}(\top) = V_A(\top)$ for all $A' \subseteq A$. An outcome A is better than an outcome B iff $V_A(\top) > V_B(\top)$.

With these primitives we can formulate various principles about preference. For instance $\prec$ should be transitive, irreflexive, and should satisfy certain averaging principles.[44]

Let us consider some key principles about preference. The first tells us that if one of two outcomes is better for you, then bringing about the better outcome is to be preferred over bringing about the other.

Prefer Better Outcomes If A and B are outcomes and A is better than B, then you should prefer A to B conditional on any piece of evidence.

[44] Joyce (1999) chapter 7 contains axiomatizations of the preference relation of causal decision theory, and some representation theorems.

The principle is obvious. We may then justify the principle Identity for counterfactuals, $A \mathrel{\square\!\!\rightarrow} A$, without appeal to linguistic intuitions as follows. We appeal instead to the Subjective Value Role, and apply the following mathematical fact (for proofs, see appendix A):

Proposition 11.1. *A subjective decision theory satisfies Prefer Better Outcomes iff it validates Identity.*

Next consider the idea that you should be indifferent to actions you know have been taken.

Indifference to Known Truths Be indifferent between propositions that follow from your evidence: $E \vee B$ and $E \vee B'$ for any B and B'.

In our notation, this is written: $E \vee B \approx_E E \vee B'$. Intuitively, you shouldn't care about alternative actions that you know will not make a difference to the world because they are weaker than what you already know to be true. If I know that A is true, I know there's no difference between what is going to happen, what would have happened if I made A true, and what would have happened if I made any proposition weaker than A true. The principle may be equivalently stated as saying that we should be indifferent between any proposition that follows from you evidence and the *status quo*. We may use this principle to justify Modus Ponens.

Proposition 11.2. *A decision theory satisfies Indifference to Known Truths iff it validates Modus Ponens.*

Our final principle articulates the following idea.

Irrelevant Outcomes Suppose A and B are outcomes, you prefer one over the other, and are indifferent between A and a gamble between A and B. Then you should be indifferent between any gamble A' with A as an outcome (i.e. with $A \subseteq A'$) and a gamble between A' and B.

In our notation, this means that if $A \prec_E B$, and $A \vee B \approx_E A$, then $A' \vee B \approx_E A'$ for any proposition A' weaker than A.

The intuition, and connection to Disjunction, is this. If you are indifferent between an outcome A and a gamble between A and another preferable (or less preferable) outcome B, then B is irrelevant to any other gamble involving the outcome A. This is because if you're indifferent between A and the gamble between A and B, it means you know that if one of A or B were to be the outcome, it would be A. But then by Disjunction, you know that if one of A, B and some other outcomes obtained, it would either be A or one of the other outcomes that obtained, it wouldn't be B. Moreover, Disjunction tells us that this outcome would have been the same if A, B and some other outcomes obtained, as if one of A and those other outcomes had obtained. Thus you should be indifferent between any gamble A' that has A as an outcome, and the same gamble except with B as an additional possible outcome, $A' \vee B$.

Proposition 11.3. *A decision theory satisfies Irrelevant Outcomes if it validates Disjunction, and validates Disjunction if it satisfies Irrelevant Outcomes and validates Identity.*

There are other ways of expressing Disjunction decision theoretically that can be helpful. Another would be that if A, B and C are outcomes with different value and you are indifferent between $A \vee B \vee C$ and A, you should be indifferent between $A \vee B$, $A \vee C$ and A. For the first indifference tells us that if $A \vee B \vee C$ obtained, it would be A that obtained, and Disjunction thus tells us that it is also A that would have obtained out of A and C, and out of A and B.

A decision theorist who accepts Irrelevant Outcomes can use it as a reason to accept Disjunction, and thus also refute those who like The Probability Role. However, I feel that the principle lacks the obviousness of the previous discussed principles. While the other two principles, Prefer Better Outcomes and Indifference to Known Truths are accepted by all decision theories I know of, including the primary alternative to causal decision theory, evidential decision theory (Jeffrey (1990)), the same cannot be said of the decision theoretic equivalents of Disjunction. The alternative principle mentioned above, for instance, is explicitly rejected by evidential decision theorists.[45] There are two sorts of causal decision theorists, on the other hand: those that agree with the evidential decision theorist that this principle of preference is not rationally required (and thus reject Disjunction), and those who accept it (and thus accept Disjunction). This is not the place to adjudicate between these two positions. It illustrates, however, that there is substance to the debate about the validity of Disjunction beyond linguistic intuitions that, due to involving disjunctive antecedents, are subject to a lot of noise.[46]

A Proofs of theorems

Theorem A.1. *Suppose that the set of objective value functions $\mathcal{O}$ satisfies Weak Permissivism (on top of Objective Outcomes). Then there is a unique selection function f that satisfies the The Objective Value Role:*

$$v_A(x) = v_*(f(A, x))$$

for all $v \in \mathcal{O}$. (Recalling that we write $v_*(x)$ for $v_\top(x)$.)

Proof. Given a space of objective value functions $\mathcal{O}$ satisfying weak permissiveness, we will define a function $f(A, x) = y$ iff $v_A(x) = v_*(y)$ for every $v_\cdot(\cdot) \in \mathcal{O}$. First we must show that this is indeed a function. Objective Outcomes ensures that for every A and x, there exists y such that $v_A(x) = v_*(y)$ for all $v_\cdot(\cdot) \in \mathcal{O}$. For the uniqueness of this y,

[45] Suppose A, B and C are three equally likely outcomes, such that the news value of A is 3, B is 6, C is 0. Then the news value of $A \vee B \vee C$ is 3, the the same A. But the news value of $A \vee B$ is 4.5, and of $A \vee C$ is 1.5.

[46] The literature on disjunctive antecedents is vast, but see, for instance, Cariani and Goldstein (2018), Khoo (2018) and the references therein.

we must show that if $v_A(x) = v_*(y) = v_*(y')$ for all $v.(\cdot)$ then $y = y'$. But this follows immediately from weak permissivism.

It is clear from the definition of f that for any objective value function, $v.(\cdot) \in \mathcal{O}$, $v_A(x) = v_*(f(A, x))$. Moreover, f is unique, for if f' also satisfied this role for every value function, then, for arbitrary A and x, $v_*(f(A, x)) = v_*(f'(A, x))$ for every value function, ensuring that $f(A, x) = f'(A, x)$ by weak permissivism. □

Theorem A.2. *Let $\mathcal{S}$ be a space of subjective values satisfying Evidence and Priors, Evidence Linearity, Subjective Outcomes, and Weak Permissivism. Then there is a unique selection function, $f_{\mathcal{S}}$, satisfying The Subjective Value Role. That is, for any value $V \in \mathcal{S}$*

$$V_E(A) = \sum_w P_{V_E}(A \boxdot\!\!\rightarrow w \mid E)V_w(\top)$$

where $\boxdot\!\!\rightarrow$ is the connective determined by $f_{\mathcal{S}}$, and P_{V_E} is the probability function determined by V_E, $P_{V_E}(A) = \frac{V_\top(\top) - V_{\neg(E \wedge A)}(\top)}{V_{E \wedge A}(\top) - V_{\neg(E \wedge A)}(\top)}$

Proof. A selection function $f_{\mathcal{S}}$ may be defined as follows:

$$f_{\mathcal{S}}(A, x) = y \text{ iff } V_x(A) = V_y(\top) \text{ for every } V \in \mathcal{S}$$

The right-hand-side specifies a well-defined function. Subjective Outcomes ensures that there exists a partition $(E_w)_{w \in W}$ such that $V_E(A) = V_w(\top)$ for any evidence E entailing E_w and any $V \in \mathcal{S}$. Since the propositions E_w are a partition, and x is a world proposition, there exists a unique world y such x entails E_y. So for every $V \in \mathcal{S}$, $V_x(A) = V_y(\top)$. Now suppose that $V_x(A) = V(y')$ for every value function V. Then $V_y(\top) = V_{y'}(\top)$ for every value, and thus $y = y'$ by Weak Permissivism.

By Evidential Linearity we know that $V_E(A) = \sum_w P_V(w \mid E)V_w(A)$. But by definition $V_w(A) = V_{f(A,w)}(\top)$, so we get that the previous sum is $\sum_w P_V(w \mid E)V_{f(A,w)}(\top) = \sum_x (\sum_{f(A,w)=x} P_V(w \mid E)V_x(\top)) = \sum_x P_V(\{w \mid f(A, w) = x\} \mid E)V_x(\top) = \sum_x P_V(A \boxdot\!\!\rightarrow x \mid E)V_x(\top)$.

Now suppose that f is another selection function satisfying The Subjective Value Role with respect to every value function, V:

$$V_E(A) = \sum_w P_V(A \boxdot\!\!\rightarrow_f w \mid E)V_w(\top)$$

Where $A \boxdot\!\!\rightarrow_f w = \{z \mid f(A, z) = w\}$. If $f(A, x) = y$ then $x \in A \boxdot\!\!\rightarrow_f y$, and $P_V(A \boxdot\!\!\rightarrow_f w \mid x) = 1$ if $w = y$ and $= 0$ otherwise. So for any value function V, $V_x(A) = \sum_w P_V(A \boxdot\!\!\rightarrow_f w \mid x)V_w(\top) = V_y(\top)$. So for every V, $V_x(A) = V_y(\top)$, which by the definition of $f_{\mathcal{S}}$, means that $f_{\mathcal{S}}(A, x) = y$. Since A, x and y were arbitrary, $f = f_{\mathcal{S}}$ as required.

□

Proposition A.3.

- *$(W, \mathcal{O})$ is a Modus Ponens decision theory just in case it satisfies Truth Indifference.*

- $(W, \mathcal{O})$ *is an Identity decision theory just in case it satisfies Prospect-Outcome Compatibility.*
- $(W, \mathcal{O})$ *is a Disjunction decision theory just in case it satisfies Prospect Choice.*
- $(W, \mathcal{O})$ *is a Modest Modus Ponens decision theory just in case it satisfies Outcome Value is the Status Quo.*

Proof. The proofs here are routine. I will establish the case of Modus Ponens and Disjunction.

Suppose that $(W, \mathcal{O})$ is a Modus Ponens decision theory. A standard argument shows that $f_{\mathcal{O}}$ satisfies MP (to show $f_{\mathcal{O}}(A, x) = x$ evaluate a case of Modus Ponens where $[\![P]\!] = A$ and $[\![Q]\!] = \{x\}$). In section 3 footnote 20 we showed that this implies that v satisfies Truth Indifference. If $x \in A$, $v_A(x) = v_\top(f_{\mathcal{O}}(A, x)) = v_\top(x)$ my MP, and similarly if $x \in B$, so $v_A(x) = v_B(x)$.

Conversely, suppose that $(W, \mathcal{O})$ satisfies Truth Indifference. We want to show that if $x \in A$, then $f_{\mathcal{O}}(A, x) = x$, i.e. that $v_A(x) = v_\top(x)$ for every $v \in \mathcal{O}$, but this follows by Truth Indifference since $x \in A \cap \top$.

Suppose that $(W, \mathcal{O})$ is a Disjunction decision theory. A standard argument shows that $f_{\mathcal{O}}$ satisfies DIS, so as in footnote 20 this implied that Prospect Choice holds.

Conversely, suppose that $(W, \mathcal{O})$ satisfies Prospect Choice. We want to show that $f_{\mathcal{O}}(A \cup B, x) = \{f(A, x), f(B, x)\}$. By Prospect choice, for any v $v_{A \cup B}(x) = v_A(x)$ or $v_B(x)$, and we know that if one of the disjuncts holds for some v it holds for all v, by the definition of $f_{\mathcal{O}}$. so $f_{\mathcal{O}}(A \cup B, x) \in \{f(A, x), f(B, x)\}$ as required. □

Proposition A.4. *A subjective decision theory satisfies Prefer Better Outcomes iff it validates Identity.*

Proof. Suppose Prefer Better Outcomes holds. A is a proposition, and x a world. Consider a value function that that assigns every world in A a value of 1 and every other world 0 (i.e. $V_x(\top) = 1$ when $x \in A$, and $V_x(\top) = 0$ otherwise). It follows from Averaging that $V_B(\top) = 1$ for any non-empty $B \subseteq A$, and $V_B(\top) = 0$ for any non-empty B disjoint from B [TO DO: need archemidean axiom?]. So A and $B = W \setminus A$ are outcomes, A is a better outcome than B. So by Prefer Better Outcomes $V_x(A) > V_x(B)$. By definition of $f_{\mathcal{S}}(A, x)$, $V_x(A) = V_{f_{\mathcal{S}}(A,x)}(\top)$, so that $V_x(A) = 1$ or $= 0$, and similarly for $V_x(B)$, thus $V_x(A) = V_{f_{\mathcal{S}}(A,x)}(\top) = 1$, whih means $f_{\mathcal{S}}(A, x) \in A$.

Now suppose $f_{\mathcal{S}}(A, x) \in A$ for any nonempty $A \subseteq W$ and $x \in W$, and that A and B are outcomes with A better than B: $V_A(\top) > V_B(\top)$. By Averaging an the Archemedian axiom, it suffices to show that we should prefer A o B conditional on any world proposition, x: $V_x(A) > V_x(B)$. $V_x(A) = V_{f_{\mathcal{S}}(A,x)}(\top)$ by the definition of $f_{\mathcal{S}}$, and the right-hand-side is identical to $V_A(\top)$ since $f(A, x) \in A$, and A is an outcome. By similar reasoning $V_x(B) = V_B(\top)$, so $V_x(A) > V_x(B)$ as required. □

Proposition A.5. *A decision theory satisfies Indifference to Known Truths iff it validates Modus Ponens.*

Proof. In the one direction, we have $V_A(A \vee B) = \sum_{w \in A} P_A(w) \times V_{f_{\mathcal{S}}(A \vee B, w)}(\top) = \sum_{w \in A} P_A(w) V_w(\top)$, since $f_{\mathcal{S}}(A \vee B, w) = w$ when $w \in A$. Calculating $V_A(A \vee B')$ in the same way we get the same result.

Conversely, suppose $w \in A$. To show that $f_(A, w) = w$ we must show that for arbitrary V, $V_w(A) = V_w(\top)$. Since $w \in A$, $A = A \vee w$, so $V_w(A) = V_w(w \vee A)$, moreover $V_w(w \vee A) = V_w(w \vee \top) = V_w(\top)$ by Indifference to Known truths, and this is the required result.

□

Proposition A.6. *A decision theory satisfies Irrelevant Outcomes if it validates Disjunction, and validates Disjunction if it satisfies Irrelevant Outcomes and validates Identity.*

Proof. Suppose that A and B are outcomes, $V_E(A) < V_E(B)$ and $V_E(A \vee B) = V_E(A)$. Note that the first two stipulations imply that A and B be disjoint.

The stipulation that $V_E(A \vee B) = V_E(A)$ means that for any $x \in E$ with positive probability, $f(A \cup B, x) = f(A, x)$. Write $A' = A \cup C$. Now $f(A \cup C \cup B, x) = f(A \cup C, x)$ for any $x \in E$. For by DIS, either $f(A \cup C \cup B, x) = f(A \cup B, x)$ or $f(A \cup C \cup B, x) = f(C, x)$. But in the former case we know that $f(A \cup B, x) = f(A, x)$ by assumption, and by DIS, $f(A, x) = f(A \cup C, x)$. And in the latter case $f(C, x) = f(A \cup C)$ also by DIS. Now $V_E(A \vee C \vee B) = \sum_w P(w) \times V_{f(A \cup C \cup B, w)}(\top) = \sum_w P(w) \times V_{f(A \cup C, w)}(\top) = V_E(A \vee C)$ as required.

Now suppose that DIS fails: for some X, Y, w, $f(X \cup Y, w) \notin \{f(X, w), f(Y, w)\}$. By ID $f(X \cup Y, w) \in X \cup Y$, so without loss of generality suppose $f(X \cup Y, x) \in X$. Divide X into two disjoint sets W and Z, such that $f(X, w) \in Z$, and $f(X \cup Y, w) \in W$ and $W \cap Y = \emptyset$ (e.g. $W = \{f(X \cup Y, w)\}$, $Z = X \setminus W$). By Permissivism, we can assign worlds in W value of 1, and worlds in $(Z \cup Y) \setminus W$ value 2. Because Y is disjoint from W, this makes W and Y outcomes. Now either $f(Y \cup W, w)$ belongs to Y or it belongs to W. In the former case set $A = Y, B = W$ and $A' = (Z \cup Y) \setminus W$. For then $V_w(A) = 2 = V(A \vee B)$, but $V_w(A') = 2$ and $V(A' \vee B) = 1$. In the latter case, set $A = Y, B = W$, and $A' = Y \cup Z$. Then $V_w(A) = 2 = V_w(A \vee B)$, but $V_w(A') = 2$ and $V_w(A' \vee B) = 1$. In either case we get a counterexample.

□

References

Ernest Adams. The logic of conditionals. *Inquiry: An Interdisciplinary Journal of Philosophy*, 8(1-4):166–197, 1965. doi: 10.1080/00201746508601430.

Arif Ahmed and Jack Spencer. Objective value is always newcombizable. *Mind*, 129 (516):1157–1192, 2020. doi: 10.1093/mind/fzz070.

Anthony Appiah. Generalising the probabilistic semantics of conditionals. *Journal of Philosophical Logic*, 13(4):351–372, 1984. doi: 10.1007/bf00247710.

Andrew Bacon. Actual value in decision theory. *Analysis*, 82(4):617–629, 2022. doi: 10.1093/analys/anac014.

Andrew Bacon. *Vagueness and Thought*. Oxford University Press, Forthcominga.

Andrew Bacon. Stalnaker's thesis in context. *Review of Symbolic Logic*, Forthcomingb.

Nathaniel Baron-Schmitt and Ginger Schultheis. Progressive specificity. manuscript.

J.F. Bennett. *A philosophical guide to conditionals*. Oxford University Press, 2003.

Ethan D Bolker. Functions resembling quotients of measures. *Transactions of the American Mathematical Society*, 124(2):292–312, 1966.

Ethan D. Bolker. A simultaneous axiomatization of utility and subjective probability. *Philosophy of Science*, 34(4):333–340, December 1967.

David Boylan and Ginger Schultheis. Strengthening principles and counterfactual semantics. In Thom van Gessel & Floris Roelofsen Alexandre Cremers, editor, *Proceedings of the 21st Amsterdam Colloquium*, pages 155–164. 2017.

Wylie Breckenridge and Ofra Magidor. Arbitrary reference. *Philosophical Studies*, 158 (3):377–400, 2012. doi: 10.1007/s11098-010-9676-z.

John Broome. *Weighing Lives*. Oxford University Press, New York, 2004.

Lara Buchak. *Risk and rationality*. OUP Oxford, 2013.

Fabrizio Cariani and Simon Goldstein. Conditional heresies. *Philosophy and Phenomenological Research*, 2(2):251–282, 2018. doi: 10.1111/phpr.12565.

B.F. Chellas. Basic conditional logic. *Journal of philosophical logic*, 4(2):133–153, 1975.

Cian Dorr and John Hawthorne. If...: A theory of conditionals.

P. B. Downing. Vii.–subjunctive conditionals, time order, and causation. *Proceedings of the Aristotelian Society*, 59(1):125–140, 1959. doi: 10.1093/aristotelian/59.1.125.

Dorothy Edgington. Do conditionals have truth conditions? *Crítica: Revista Hispanoamericana de Filosofía*, 18(52):3–39, 1986.

Dorothy Edgington. On conditionals. *Mind*, 104(414):235–329, 04 1995.

Kit Fine. Critical notice of lewis, counterfactuals. *Mind*, 84(335):451–458, 1975.

J. Dmitri Gallow. Counterfactual decision theory is causal decision theory. *Pacific Philosophical Quarterly*, 105(1):115–156, 2024. doi: 10.1111/papq.12451.

Allan Gibbard. *Thinking How to Live*. Harvard University Press, 2003.

Allan Gibbard and William L. Harper. Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach, and E. F. McClennen, editors, *Foundations and Applications of Decision Theory: Vol.II: Epistemic and Social Applications*, pages 125–162. D. Reidel, 1978.

Anthony Gillies. Indicative conditionals. In Gillian Russell and Delia Graff Fara, editors, *Routledge Companion to Philosophy of Language*. Routledge, 2011.

Anthony S Gillies. Counterfactual scorekeeping. *Linguistics and Philosophy*, 30(3): 329–360, 2007a.

Anthony S. Gillies. Counterfactual scorekeeping. *Linguistics and Philosophy*, 30(3): 329–360, 2007b. doi: 10.1007/s10988-007-9018-6.

Jeremy Goodman. Consequences of conditional excluded middle. manuscript.

William Leonard Harper, Robert Stalnaker, and Glenn Pearce, editors. *Ifs*. D. Reidel, Dordrecht, 1981.

Brian Hedden. Counterfactual decision theory. *Mind*, 132(527):730–761, 2023. doi: 10.1093/mind/fzac060.

Frank Jackson. A causal theory of counterfactuals. *Australasian Journal of Philosophy*, 55(1):3–21, 1977. doi: 10.1080/00048407712341001.

Frank Jackson. *Conditionals*. Blackwell, New York, 1987.

Richard C. Jeffrey. *The logic of decision*. University of Chicago Press, 1990.

J.M. Joyce. *The foundations of causal decision theory*. Cambridge Univ Pr, 1999.

Justin Khoo. Disjunctive antecedent conditionals. *Synthese*, 198(8):7401–7430, 2018. doi: 10.1007/s11229-018-1877-6.

Justin Khoo. *The Meaning of "If"*. Oxford University Press, New York, USA, 2022.

Nathan Klinedinst. Quantified conditionals and conditional excluded middle. *Journal of Semantics*, 28(1):149–170, 2011.

Jason Konek and Ben Levinstein. The foundations of epistemic decision theory. *Mind*, 128(509):69–107, 2019. doi: 10.1093/mind/fzw044.

Angelika Kratzer. *Modals and Conditionals. New and Revised Perspectives*. Oxford University Press, Oxfords, 2012.

Angelika Kratzer. Chasing hook. *Conditionals, Probability, and Paradox. Oxford University Press, forthcoming*, 2020.

David Lewis. Counterfactual dependence and time?s arrow. *Noûs*, 13(4):455–476, 1979. doi: 10.2307/2215339.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30, 1981. doi: 10.1080/00048408112340011.

David K. Lewis. *Counterfactuals*. Blackwell, 1973.

David K. Lewis. Probabilities of conditionals and conditional probabilities. *The philosophical review*, 85(3):297–315, 1976.

David K. Lewis. A subjectivist's guide to objective chance. *Studies in inductive logic and probability*, 2:263–293, 1980.

Matthew Mandelkern. Talking about worlds. *Philosophical Perspectives*, 32(1):298–325, 2018. doi: 10.1111/phpe.12112.

Sarah Moss. On the pragmatics of counterfactuals. *Noûs*, 46(3):561–586, 2012.

J.L. Pollock. *Subjunctive reasoning*. Reidel Dordrecht, 1976.

Paolo Santorio. Path semantics for indicative conditionals. *Mind*, 131(521):59–98, 2022. doi: 10.1093/mind/fzaa101.

Leonard Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

Mark Schroeder. Two roles for propositions: Cause for divorce? *Noûs*, 47(3):409–430, 2011. doi: 10.1111/j.1468-0068.2011.00833.x.

Ginger Schultheis. ?might? counterfactuals. *Linguistics and Philosophy*, 47(5):839–865, 2024. doi: 10.1007/s10988-024-09416-6.

Moritz Schulz. Wondering what might be. *Philosophical Studies*, 149(3):367–386, 07 2010.

Moritz Schulz. *Counterfactuals and Probability*. Oxford University Press, Oxford, United Kingdom, 2017.

R. Stalnaker. A theory of conditionals. *Studies in logical theory*, 2:98–112, 1968.

R. Stalnaker. Letter to van fraassen. *WL Harper and CA Hooker (1976)*, pages 302–306, 1976.

Robert Stalnaker. A defense of conditional excluded middle. *Ifs*, pages 87–104, 1980.

Robert Stalnaker. *Inquiry*. The MIT Press, 1984.

Bas C. van Fraassen. Probabilities of conditionals. In William L. Harper and Clifford Allan Hooker, editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, volume 1, pages 261–308. D. Reidel Publishing Co., 1973.

Kai Von Fintel. Counterfactuals in a dynamic context. *Current Studies in Linguistics Series*, 36:123–152, 2001.

Ralph Wedgwood. Must rational intentions maximize utility? *Philosophical Explorations*, 20(sup2):73–92, 2017. doi: 10.1080/13869795.2017.1356352.

J Robert G Williams. Defending conditional excluded middle. *Noûs*, 44(4):650–668, 2010.

J. Robert G. Williams. Aptness and means-end coherence: A dominance argument for causal decision theory. *Synthese*, 201(2):1–19, 2023. doi: 10.1007/s11229-022-04017-x.

Timothy Williamson. *Suppose and Tell: The Semantics and Heuristics of Conditionals*. Oxford University Press, Oxford, England, 2020.