This is a post-peer-review, pre-copyedit version of an article published in Erkenntniss. The final authenticated version is available online at: http://dx.doi.org/10.1007/s10670-020-00357-7"

The material conditional is sufficient to model deliberation

Giacomo Bonanno*
University of California, Davis, USA
gfbonanno@ucdavis.edu

December 30, 2020

Abstract

There is an ongoing debate in the philosophical literature whether the conditionals that are central to deliberation are subjunctive or indicative conditionals and, if the latter, what semantics of the indicative conditional is compatible with the role that conditionals play in deliberation. We propose a possible-world semantics where conditionals of the form "if I take action a the outcome will be x" are interpreted as material conditionals. The proposed framework is illustrated with familiar examples and both qualitative and probabilistic beliefs are considered. Issues such as common-cause cases and 'Egan-style' cases are discussed.

1 Introduction

It is a trivial observation that making choices requires comparing alternatives. Making decisions – whether in a one-person situation or in an interactive one (that is, a game) – involves reasoning along the following lines: 'if I take action a, then the outcome will be x and if I take action b, then the outcome will be y'. Having reached an assessment of this form for every available action, the optimal or rational choice will then be the

*I am grateful to three anonymous reviewers for helpful and constructive comments.

one that yields the most desirable outcome. Despite the simplicity of this observation, it turns out to be not at all obvious how to obtain a satisfactory formalization of the conditionals involved in deliberation. It is a widely held opinion that the relevant conditionals are subjunctive conditionals or counterfactuals. For example, Gibbard and Harper write:

> "[R]ational decision-making involves conditional propositions: when a person weighs a major decision, it is rational for him to ask, for each act he considers, what would happen if he performed that act. It is rational, then, for him to consider propositions of the form 'If I were to do *a* , then *c* would happen'. Such a proposition we shall call a *counterfactual*." ([10, p. 153])

Along the same lines, Aumann writes:

> "[O]ne really cannot discuss rationality, or indeed decision making, without substantive conditionals and counterfactuals. Making a decision means choosing among alternatives. Thus one must consider hypothetical situations – what would happen if one did something different from what one actually does. [I]n interactive decision making – games – you must consider what other people would do if you did something different from what you actually do." ([2, p. 15])

In natural language conditionals can be expressed in different ways, conveying different meaning. For example, if I say 'if I had left the office at 4 pm I would not have been stuck in traffic', I convey information that – as a matter of fact – I did not leave the office at 4 pm and thus I am uttering a *counterfactual conditional*, that is, a conditional with a false antecedent (such a statement would not make sense if uttered *before* 4 pm). On the other hand, if I say 'if I leave the office at 4 pm I will not be stuck in traffic' I am uttering an *indicative conditional* and am conveying the information that I am evaluating the consequences of a possible future action (such a statement would not make sense if uttered *after* 4 pm). Concerning the latter conditional, is there a difference between the indicative mood and the subjunctive mood? If I said 'if I *were* to leave the office at 4 pm I *would* not be stuck in traffic', would I be conveying the same information as with the previous, indicative, conditional? On this point there does not seem to be a consensus in the literature. We agree with DeRose's claim that the

subjunctive mood conveys different information relative to the indicative mood: its role is to

> "call attention to the possibility that the antecedent is (or will be) false, where one reason one might have for calling attention to the possibility that the antecedent is (or will be) false is that it is quite likely that it is (or will be) false." ([4, p. 10])

Thus the indicative conditional signals that the decision whether to leave at 4 pm is still "open", while the subjunctive conditional seems to convey that the speaker is somehow ruling out that option: for example, he has made a tentative or firm decision not to leave at 4 pm (see Section 3.4 for further discussion of this point).

The focus of this paper is on the *conditionals of deliberation*, which, borrowing from Krzyżanowska, can be defined as follows:

> "The term conditional of deliberation is meant to apply to only those conditional sentences that concern actions that a deliberating agent considers undertaking, on the one hand, and events or states of affairs that depend on those actions, on the other hand." ([15, p. 3])

We agree with DeRose that the conditionals of deliberation are best understood as indicative conditionals. However, as Douven ([5]) points out, there is no agreed-upon understanding among philosophers concerning the nature of indicative conditionals. Some maintain that indicative conditionals ought to be understood as material conditionals, others claim that they ought to be understood in terms of Stalnaker-Lewis ([27, 21]) counterfactuals, while others go as far as claiming that indicative conditionals are non-propositional, that is, they lack truth conditions.[1] The objective of this paper is not to contribute to the debate concerning indicative conditionals in general, but rather to put forward the thesis that, **within the context of deliberation, material conditionals are sufficient**. In other words, sentences of the form "if I take action a then the outcome will be x" can be interpreted as material conditionals, that is, as equivalent to "either I don't take action a or the outcome will be x". We propose a possible-world semantic analysis of the conditionals of deliberation, where a possible world is described in terms of the external facts (or environment or context or

[1] For an account of the different views see [5] and [6].

background), the action taken and the corresponding outcome. Of course, we also need to model the beliefs and preferences of the Decision Maker (henceforth DM).

Before we proceed, we need to address a natural question, namely "why is it important to establish whether material conditionals are sufficient to model deliberation or whether counterfactuals are inherently needed?" From a pragmatic point of view, the type of reasoning involved in deliberation is clear; what is to be gained by analysing how one should interpret the conditionals that are entailed?[2] Of course, the same objection could be raised with respect to the entire literature on the topic of deliberation: why worry ([4, 7, 15]) about whether the conditionals of deliberation are indicative conditionals or counterfactuals? or whether ([9]) conditionals of deliberation can be assigned truth values? These are philosophical issues that, perhaps, do not have pragmatic value. It is clear, however, that the conceptual apparatus of Stalnaker-Lewis counterfactuals is more complex than propositional logic: at the semantic level one needs to postulate a family of similarity relations on the set of possible worlds (one for each possible world)[3] and at the syntactic level one needs to rely on *modal* logic; on the other hand, the semantics for material conditionals is merely the Boolean algebra of sets and, at the syntactic level, propositional logic is all that is needed. We believe that *it is important to clarify what level of logical complexity is needed to analyse an issue*: in this case the notion of deliberation.[4]

[2]As a reviewer pointed out, a decision/game theorist might argue as follows: "We have the decision matrix. It is understood that we can select an act but not a state, and that we have beliefs over states but not over acts. What more is needed?" The reviewer points out that one could read the decision matrix as a list of counterfactuals and take expected utility maximization to be a description of the process of choice. Why worry about more than that? The reviewer goes on to state "For many of us the answer is clear. We have inherent intellectual interest in modeling the reasoning process; we believe that it will be very useful to understand the psychology of decision making, and probably indispensable if we want to do Artificial Intelligence (AI)." However, the focus of this paper is not on the psychology of decision making, nor on its AI implementation, but rather, as explained below, on the complexity of the conceptual apparatus that one needs in order to model deliberation.

[3]One then declares the sentence "if ϕ were the case then ψ would be the case" to be true at a possible world ω if ψ is true at the most similar world(s) to ω where ϕ is true.

[4]An analogy might be useful: suppose that a mathematical theorem has been proved using the "heavy" apparatus of algebraic topology and somebody then offers a much simpler proof that uses only elementary tools. Rather than asking "why should we care

A separate, but related, issue is whether the material conditional can meaningfully capture the natural language indicative conditionals that are central to deliberation.[5] The significance of conditionals in the context of decision making stems from the fact that they capture causal dependencies between actions and their outcomes. Can the material conditional capture these dependencies? We hope to convince the reader that the answer is 'Yes'. As explained in Section 2, the causal dependencies between environment and action, on one side, and outcome, on the other, are captured by a function that assigns a unique outcome to every pair (e, a), where e is an environment and a an action. Since every possible world encodes these three items (environment, action and outcome) the causal dependencies are encoded in the set of possible worlds. Furthermore, since the crucial assumption in the proposed framework is that the DM considers every action possible (that is, for every action there is an accessible world where he takes that action), material conditionals are indeed sufficient to capture such dependencies: the material conditional "if I take action a the outcome is x" (or, in the probabilistic case, "the probability of outcome x is p") zooms in – through the lens of the DM's beliefs – on those worlds where action a is indeed taken and verifies that the outcome is indeed x (while the worlds where action a is not taken are an innocuous appendage). Whether the DM himself thinks of these conditionals as material conditionals is not the issue: the point we are trying to make is that, *from a modeling point of view*, the material conditional is sufficient to capture the essence of deliberation.

We will begin our analysis with two examples and then, in the next section, develop the general framework.

1.1 Example 1: Bob in the shower

Bob is visiting a foreign country and is now in the bathroom of his hotel, ready to take a shower; the room is very cold and he wants to enjoy a hot shower. There are two faucets: the one on the left is not labeled, while

how we prove a theorem, if we already know that it is true?", mathematicians would probably value the conceptual clarification provided by the elementary proof.

[5] A Reviewer observed that "Even though there seems to be no agreement whatsoever concerning the meaning of indicative conditionals, the idea that a natural language conditional is a material conditional is by far the hardest to maintain" and added "What would it mean for the proposal put forward in the paper? That a decision maker can make decisions by interpreting conditionals in a completely artificial way?"

the one on the right is labeled with a 'c'. He believes – as it happens, erroneously – that 'c' stands for 'cold' and infers that the unlabeled faucet on the left is the one that will deliver hot water. There are two possible actions: turn on the left faucet (L) and turn on the right faucet (R), and two outcomes: he gets hot water (H) or he gets cold water (C). Which outcome occurs depends on the action taken and on the environment, by which we mean the actual plumbing connections: either $(cold, hot)$ (cold water on the left and hot water on the right) or $(hot, cold)$. Figure 1 shows this in the familiar matrix form (inside each cell is recorded the outcome: C or H).

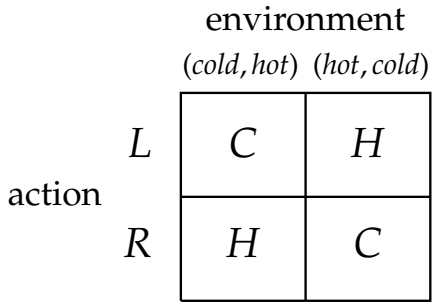


Figure 1: The relationship between action, environment and outcome

However, we will use a different representation, in terms of possible worlds. A possible world is meant to be a complete description of the situation. In this case: (1) what the environment is, (2) what action is taken, (3) what outcome occurs and (4) the utility of the outcome (Bob prefers outcome H to outcome C and thus we can assign utility 1 to the former and utility 0 to the latter). Note that, as is standard in game theory (see, for example, [1, 3]), in the description of the world we include the action taken. The possible-world representation is shown in Figure 2. Thus, for example, α is the possible world where – as a matter of fact – the left faucet delivers cold water and the right faucet delivers hot water, Bob ends up turning on the left faucet and the outcome is that he gets cold water, so that his utility is 0.

With slight abuse of notation, we will use the letter L to denote both the action of turning on the left faucet and the sentence "the left faucet is turned on"; similarly for the letter R, the letter C (denoting both the outcome that cold water flows and the sentence "cold water flows") and

possible world:	α	β	γ	δ
environment:	*(cold,hot)*	*(hot,cold)*	*(hot,cold)*	*(cold,hot)*
action:	L	L	R	R
outcome:	C	H	C	H
utility:	0	1	0	1

Figure 2: The possible-world representation

the letter H. Thus in Figure 2 the truth set of the sentence L, denoted by $\|L\|$, is the the set of worlds $\|L\| = \{\alpha, \beta\}$, and the material conditional "if the left faucet is turned on then hot water will flow", denoted by $L \rightarrow H$, is true not only at state β (where both L and H are true) but also at states γ and δ, where the antecedent is false: $\|L \rightarrow H\| = \{\beta, \gamma, \delta\}$.

To complete the picture, we need to add Bob's beliefs. Bob mistakenly believes "if I turn on the left faucet I will get hot water and if I turn on the right faucet I will get cold water"; that is, he believes that both $L \rightarrow H$ and $R \rightarrow C$ are true. According to our assumptions, Bob is wrong in thinking that the label 'c' means 'cold': he is in Italy where 'c' means 'caldo', which translates into 'hot'. Thus the true or actual world will be either α (if he ends up turning on the left faucet) or δ (if he ends up turning on the right faucet). However, Bob believes that the true or actual world is (depending on what action he takes) either β or γ.

We represent beliefs by means of a binary relation $\mathcal{B} \subseteq \Omega \times \Omega$ on the set of possible worlds $\Omega = \{\alpha, \beta, \gamma, \delta\}$. For every $\omega \in \Omega$ we denote by $\mathcal{B}(\omega)$ the set of worlds that are reachable from ω:

$$\mathcal{B}(\omega) = \{\omega' \in \Omega : (\omega, \omega') \in \mathcal{B}\}. \tag{1}$$

We shall throughout take beliefs to be consistent and to satisfy positive and negative introspection, that is, the relation $\mathcal{B}$ will be assumed to be serial ($\mathcal{B}(\omega) \neq \emptyset$, for every $\omega \in \Omega$), transitive (if $\omega' \in \mathcal{B}(\omega)$ then $\mathcal{B}(\omega') \subseteq \mathcal{B}(\omega)$) and euclidean (if $\omega' \in \mathcal{B}(\omega)$ then $\mathcal{B}(\omega) \subseteq \mathcal{B}(\omega')$). Thus the logic of belief will be taken to be KD45.[6]

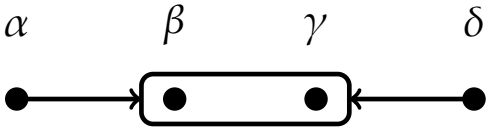
The crucial step in constructing the relation $\mathcal{B}$ is the following assump-

[6]Nothing depends on this assumption, but it makes for a more convenient graphical representation of the relation $\mathcal{B}$.

tion:

> For every possible world ω and every available action a, there is a world $\omega' \in \mathcal{B}(\omega)$ such that the action taken at ω' is a. (A1)

Assumption (A1) says that, *for every action that is available to him, the DM considers it possible that he takes that action.* Before we comment on this crucial assumption (see Section 1.2), let us complete the example. To represent Bob's beliefs we take $\mathcal{B}(\omega) = \{\beta, \gamma\}$, for every $\omega \in \{\alpha, \beta, \gamma, \delta\}$. Graphically we can represent $\mathcal{B}$ as shown in Figure 3, where we adopt the following convention: for any two possible worlds ω and ω', $\omega' \in \mathcal{B}(\omega)$ if and only if either ω and ω' are enclosed in the same rounded rectangle or there is an arrow from ω to the rounded rectangle containing ω'.[7]



environment	(*cold*, *hot*)	(*hot*, *cold*)	(*hot*, *cold*)	(*cold*, *hot*)
action	L	L	R	R
outcome	C	H	C	H
utility	0	1	0	1

Figure 3: Bob's belief relation: $\mathcal{B}(\alpha) = \mathcal{B}(\beta) = \mathcal{B}(\gamma) = \mathcal{B}(\delta) = \{\beta, \gamma\}$.

If ϕ is a formula, we denote by $B\phi$ the formula "the DM believes ϕ" and the validation rule is the standard one: $B\phi$ is true at possible world ω, denoted by $\omega \models B\phi$, if and only if ϕ is true at every possible world that the DM considers possible at ω; that is, letting $\|\phi\|$ denote the truth set of ϕ, $\omega \models B\phi$ if and only if $\mathcal{B}(\omega) \subseteq \|\phi\|$. Hence, in the example illustrated in Figure 3, where the truth set of the material conditional $L \rightarrow H$ is $\|L \rightarrow H\| = \{\beta, \gamma, \delta\}$, $B(L \rightarrow H)$ is true at every possible world, that is, at every world Bob believes "if I turn on the left faucet I will get hot water";

[7] In other words, for any two possible worlds ω and ω' that are enclosed in a rounded rectangle, $\{(\omega, \omega), (\omega, \omega'), (\omega', \omega), (\omega', \omega')\} \subseteq \mathcal{B}$ (hence the relation is total on the set of possible worlds contained in the rectangle) and if there is an arrow from a possible world ω to a rounded rectangle then, for every ω' in that rectangle, $(\omega, \omega') \in \mathcal{B}$.

if the true or actual world is α then Bob's beliefs are erroneous: as a matter of fact, if he turns on the left faucet, he will get cold water. Similarly, since the truth set of the material conditional $R \rightarrow C$ is $\|R \rightarrow C\| = \{\alpha, \beta, \gamma\}$, the formula $B(R \rightarrow C)$ is true at every possible world. Thus both $B(L \rightarrow H)$ and $B(R \rightarrow C)$ are true at every world.

The DM's rationality at a possible world can be assessed by relating the action that he takes at that world to his beliefs. In the simple case where there are only two actions and only two outcomes we can define rationality as follows:[8]

> If at possible world ω the DM
>
> 1. believes that
> - if he takes action a the outcome will be z, and
> - if he takes action a' the outcome will be z',
> 2. weakly prefers z to z' (equivalently, the utility of z is greater than or equal to the utility of z'),
> 3. takes action a,
>
> then he is rational at ω.

Using this definition we conclude that, in the example illustrated in Figure 3, Bob is rational at worlds α and β, where he turns on the left faucet, and irrational at the remaining two worlds.

1.2 Discussion of Assumption (A1)

We now turn to a discussion of Assumption (A1), which says that, at every possible world, the DM considers it possible that he takes any of the available actions. In the proposed framework the DM's beliefs are modeled as "pre-choice" or "deliberation-stage" beliefs: when deciding what to do, the DM considers the consequences of *all* his actions, *without pre-judging his subsequent decision*; in other words, *the DM's beliefs are truly open to the possibility of taking any of the available actions*. Indeed, as pointed

[8] This basic – and very weak – definition of rationality seems to be uncontroversial. For example, it seems to be in accordance with the following definition proposed by Krzyżanowska ([15, p. 4]): "If A, in a context C, accepts 'If ϕ, ψ', and desires ψ to be the case, it would be rational for A in C to (attempt to) make ψ true."

out by several authors, it is the essence of deliberation that one cannot reason towards a choice if one already knows what that choice will be. For instance, Ginet [12, p. 50] claims that "it is conceptually impossible for a person to know what a decision of his is going to be before he makes it"; Goldman [13, p. 194] writes that "deliberation implies some doubt as to whether the act will be done" and Levi sates that "the deliberating agent cannot, before choice, predict how he will choose" [19, p. 65] and coins the phrase "deliberation crowds out prediction" [20, p. 81].[9]

A separate issue is whether it makes sense to deny the DM knowledge of his current choice, while at the same time allowing him to have beliefs about (or be certain of) what choice he will make *at a later time* (in the context of sequential decisions). This is an issue that has been addressed in the literature and several authors have maintained that there is no inconsistency between the principle that one should not attribute to the DM beliefs about his current choice and the claim that, on the other hand, one *can* attribute to the DM beliefs about later choices. For example, Gilboa writes:

> "[. . .] we are generally happier with a model in which one cannot be said to have beliefs about (let alone knowledge of) one's own choice *while making this choice*. [. . .] One may legitimately ask: Can you truly claim you have no beliefs about your own future choices? Can you honestly contend you do not believe – or even know – that you will not choose to jump out of the window? [. . .] The answer to these questions is probably a resounding 'No'. But the emphasis should be on timing: when one considers one's choice tomorrow, one may indeed be quite sure that one will not decide to jump out of the window. However, a future decision should actually be viewed as a decision by a different "agent" of the same decision maker. [. . .] It is only at the time of choice, within an "atom of decision", that we wish to preclude beliefs about it." [11, pp. 171-172]

In a similar vein, Levi writes

> "Agent X may coherently assign unconditional credal probabilities to hypotheses as to what he will do when some future

[9] Similar observations were made by other authors (e.g. Kyburg, [17, p. 80]). For a list of relevant references see [18].

> opportunity for choice arises. Such probability judgments can have no meaningful role, however, when the opportunity of choice becomes the current one." [20, p. 81]

Similarly, Spohn [25, p. 114] states the principle that "any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts" but also maintains [26, pp. 44-45] that in the case of sequential decision making, the DM *can* ascribe subjective probabilities to his future (but not to his present) actions.

An implication of the proposed approach is that, since – at the time of deliberation – the DM does not know what choice he is going to make, he cannot know that his forthcoming choice is rational. Thus at a possible world the DM may be rational and not know that he is rational. This is unavoidable if one wants to model pre-choice or deliberation-stage beliefs.[10]

1.3 Example 2: Pete and the card game

The second example is DeRose's ([4]) version of Gibbard's ([9]) riverboat example, which we further simplify by considering a deck of only 3 cards.[11]

> Pete and Gus are in the final round of a card game. They each draw one card from a deck of thee cards, numbered 1, 2 and 3. Each of the players looks at his own card, but not at the opponent's card. In this final round, it is up to Pete to decide whether to quit or play. If he decides to quit, he can keep the \$1,000 he won before this round (call this outcome a *Draw* and denote it by D). If Pete decides to play, both players have to show their cards. If Pete's card is higher than Gus's, Pete's winnings will be doubled (call this outcome a *Win* and denote it by W). If Pete decides to play, and his card is the lower one, he will lose everything (call this outcome a *Loss* and denote it by L). Pete has not decided yet whether to play or quit.
>
> An observer, Zack, is certain that Pete is not stupid enough to play if he knows that his own card is the lower one. Zack sees

[10] For further discussions on this issue see [19, 20].

[11] Unlike Gibbard's, DeRose's version has the advantage of not requiring familiarity with the game of poker. DeRose's example involves a deck with 100 cards.

that Gus (Pete's opponent) holds the card numbered 2.[12] He writes a note saying

(*) "If Pete plays he will win".

A second observer, Jack, sees not only Gus's card but also Pete's card, which is numbered 1. He writes a note that says

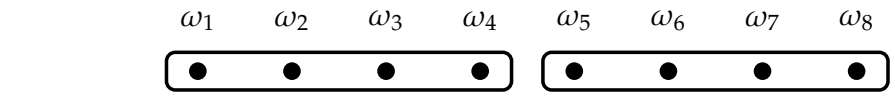
(**) "if Pete plays he will lose".

Gibbard ([9]) argues that if (*) and (**) have a truth value at all, they must both be true. This is because both Zack and Jack are warranted in their assertions and, furthermore, their assertions do not rest on any false beliefs about relevant facts. However, (*) and (**) cannot both be true, if one accepts the principle of Conditional Non-Contradition (CNC), according to which the same antecedent and contradictory consequents cannot both be true, unless the antecedent is inconsistent. Note that CNC rules out the material conditional account. Gibbard draws from this example the conclusion that indicative conditionals do not have truth conditions. Krzyżanowska et al ([16]) – while rejecting the material conditional account – propose a truth-conditional semantics for indicative conditionals that renders both (*) and (**) true by relativizing the truth of a conditional to the speaker's background knowledge.

Let us cast this example within our framework. First of all, in the description of a possible world we need to include the environment in the form of a pair (p, g) where p is the card in Pete's hand and g the card in Gus's hand. From the story we know that, as a matter of fact, $p = 1$ and $g = 2$; however, this fact is only known to Jack, while, for example, Pete considers both (1,2) and (1,3) possible and Zack (who only knows that $g = 2$) considers both (1,2) and (3,2) possible. In order to obtain a full representation of the situation described, we need eight possible worlds. We describe each possible world by specifying (1) the pair of cards (p, g), (2) the action taken by Pete (P for 'play' or Q for 'quit'), (3) the outcome (W

[12] In the original story, Zack is an accomplice of Pete's and secretly communicates to Pete the value of Gus's card. In our simplified version of the story, this is not necessary: Zack knows that Pete either has a 1 – in which case Pete knows that if he plays he loses – or Pete has a 3 – in which case Pete knows that if he plays he wins.

for 'win', D for 'draw', L for 'loss') and (4) the utility of the outcome for Pete: 10 for W, 5 for D and 0 for L.[13] We begin with Pete's belief relation $\mathcal{B}_P$ on $\Omega = \{\omega_1, \omega_2, \ldots, \omega_8\}$, which is shown in Figure 4.



	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
cards (p, g)	(1,3)	(1,3)	(1,2)	(1,2)	(3,2)	(3,2)	(3,1)	(3,1)
action	P	Q	P	Q	P	Q	P	Q
outcome	L	D	L	D	W	D	W	D
Pete's utility	0	5	0	5	10	5	10	5
Pete rational?	$\neg R$	R	$\neg R$	R	R	$\neg R$	R	$\neg R$

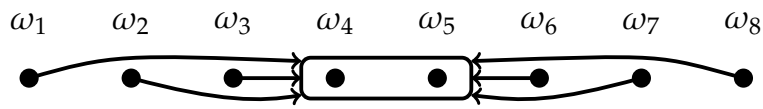
Figure 4: Pete's belief relation $\mathcal{B}_P$.

Using the definition of rationality given above, we can determine that Pete is rational at worlds ω_2, ω_4, ω_5 and ω_7 and irrational at the remaining worlds: we have marked this in Figure 4 by associating with every possible world either an R for 'rational' or a $\neg R$ for 'not rational'. Thus the truth set of the sentence 'Pete is rational' is $\|R\| = \{\omega_2, \omega_4, \omega_5, \omega_7\}$. Denote by P the sentence 'Pete plays', by L the sentence 'Pete loses' and by W the sentence 'Pete wins'; the truth sets of these sentences are $\|P\| = \{\omega_1, \omega_3, \omega_5, \omega_7\}$, $\|L\| = \{\omega_1, \omega_3\}$ and $\|W\| = \{\omega_5, \omega_7\}$, respectively. Thus the truth set of the material conditional $P \rightarrow W$ ('if Pete plays he will win') is $\|P \rightarrow W\| = (\Omega \setminus \|P\|) \cup \|W\| = \{\omega_2, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8\}$ and the truth set of the material conditional $P \rightarrow L$ ('if Pete plays he will lose') is $\|P \rightarrow L\| = (\Omega \setminus \|P\|) \cup \|L\| = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_6, \omega_8\}$. Denoting by $B_P\phi$ the formula 'Pete believes ϕ', we have that the truth set of $B_P(P \rightarrow L)$ is $\{\omega_1, \omega_2, \omega_3, \omega_4\}$ and the truth set of $B_P(P \rightarrow W)$ is $\{\omega_5, \omega_6, \omega_7, \omega_8\}$. It follows that, as noted above, the possible worlds at which Pete is rational are $\omega_2, \omega_4, \omega_5$ and ω_7.

Let us now represent the beliefs of Zack who (i) knows that Gus's card is 2 and (ii) "is certain that Pete is not stupid enough to play if he knows that his own card is the lower one"; we take this last sentence to mean that *Zack believes that Pete is rational*. Then Zack's belief relation $\mathcal{B}_Z$ is shown in

[13] We have chosen utility numbers that would not create confusion with the numbers on the cards. Since the preferences of Pete are taken to be merely ordinal preferences – in the sense that all that is expressed by utilities is that Pete prefers W to D and D to L – any three numbers would do.

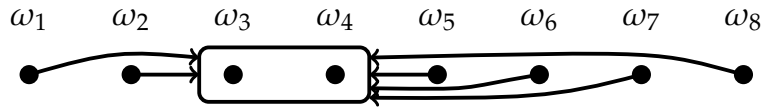
Figure 5. Note that it is true at every possible world that (i) Zack believes that Pete is rational: $\mathcal{B}_Z(\omega_i) \subseteq ||R||$, for every $i = 1, 2, \ldots, 8$ and (ii) Zack believes that 'if Pete plays he will win': $\mathcal{B}_Z(\omega_i) \subseteq ||P \rightarrow W||$, for every $i = 1, 2, \ldots, 8$.



	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
cards (p, g)	(1,3)	(1,3)	(1,2)	(1,2)	(3,2)	(3,2)	(3,1)	(3,1)
action	P	Q	P	Q	P	Q	P	Q
outcome	L	D	L	D	W	D	W	D
Pete's utility	0	5	0	5	10	5	10	5
Pete rational?	$\neg R$	R	$\neg R$	R	R	$\neg R$	R	$\neg R$

Figure 5: Zack's belief relation $\mathcal{B}_Z$.

Finally, let us represent the beliefs of Jack who knows that Gus's card is 2 and Pete's card is 1. Jack's belief relation $\mathcal{B}_J$ is shown in Figure 6. Note that it is true at every possible world that Jack believes that 'if Pete plays he will lose': $\mathcal{B}_J(\omega_i) \subseteq ||P \rightarrow L||$, for every $i = 1, 2, \ldots, 8$.



	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
cards (p, g)	(1,3)	(1,3)	(1,2)	(1,2)	(3,2)	(3,2)	(3,1)	(3,1)
action	P	Q	P	Q	P	Q	P	Q
outcome	L	D	L	D	W	D	W	D
Pete's utility	0	5	0	5	10	5	10	5
Pete rational?	$\neg R$	R	$\neg R$	R	R	$\neg R$	R	$\neg R$

Figure 6: Jack's belief relation $\mathcal{B}_J$.

Since – as a matter of fact – Pete holds the card numbered 1 and Gus the card numbered 2, after Pete makes his decision, the true or actual world will be either ω_3 or ω_4 and at *either* of these two possible worlds all of the following formulas are true:

1. $P \rightarrow L$,

2. $B_P(P \rightarrow L)$,

3. $B_J(P \rightarrow L)$,

4. $B_Z(P \rightarrow W)$,

5. $B_Z R$.

Given that, as a matter of fact, Pete has the card numbered 1, if he is rational he will quit and thus the true or actual world will be ω_4, in which case to the above list we can also add, trivially,

6. $(P \rightarrow W)$.

The material conditionals $P \rightarrow W$ (which corresponds to sentence (*) above) and $P \rightarrow L$ (which corresponds to sentence (**) above) will be "objectively" true at world ω_4 as well as "subjectively true", that is, believed to be true – the former by Zack and the latter by Jack.[14] Like Krzyżanowska et al ([16]) we take the view that the interpretation of the conditionals (*) and (**) is relative to the speaker's background knowledge/beliefs, but unlike these authors we interpret the conditionals as material conditionals.

2 A semantics for deliberation

Examples 1 and 2 above were particularly simple, in that the DM's beliefs were such that he associated a unique outcome to each of the available actions. In general, the DM might be uncertain about the environment within which he is operating. For example, suppose that I have been invited to a party but am not feeling well and I wonder if it is because I caught a contagious disease or merely because I was not able to sleep last night. If I do have a contagious disease, then my attending the party will result in my infecting other people, while if I don't have a disease then my presence at the party will not harm other people. My uncertainty about

[14]Suppose that, contrary to what Zack believes, Pete is *not* rational and chooses to play, so that he loses: the true or actual world turns out to be ω_3. Can we still claim that Zack is in a position to assert 'if Pete plays, he will win'? Those who require knowledge in order to validate an assertion as permissible, would answer 'No'; however, we side with those (e.g. [24]) who take the view that the norm of assertion is justified belief: truth or knowledge are not reqired.

the cause of my feeling unwell makes it unwarranted for me to assert 'if I go to the party I will infect other people' as well as to assert 'if I go to the party I will *not* infect other people'. When contemplating whether or not to go to the party I would have to consider both the possibility that I will infect other people and the possibility that I will not. I might be able to attach probabilities to the two outcomes. For example, since I just returned from a trip abroad, I might be inclined to think that it is very likely that I do have a contagious disease and that, therefore, I would infect other people.

In this section we provide a formulation of the framework illustrated in the previous section that is general enough to accommodate uncertainty as well as probabilistic beliefs, that is, conditionals of the form "if I take action a then the outcome will be z_1 with probability p_1, z_2 with probability p_2, $\ldots$, z_m with probability p_m". We begin with the case of qualitative beliefs and ordinal utility and consider probabilistic beliefs and expected utility in Section 2.2.

2.1 Qualitative beliefs and ordinal utility

Let

1. Ω be a set of *possible worlds*,
2. A a finite set of *actions* that the DM believes to be available to him,[15]
3. E a finite set of *external facts* or *environments*,
4. $Z = \{z_1, \ldots, z_m\}$ a finite set of *outcomes*,
5. $f : \Omega \to E$ a function that assigns a unique environment to every possible world ('f' stands for 'facts'),
6. $c : \Omega \to A$ a function that assigns a unique action to every possible world ('c' stands for 'choice'),

[15] If an action is available to the DM, but he is unaware of it, then it cannot enter into his deliberation. On the other hand, the DM might mistakenly believe that he can perform an action which – as a matter of fact – is *not* available to him (e.g. shooting a gun that he believes to be loaded, whereas in fact it has no bullets), in which case he *will* consider the consequences of taking that action (and, if he attempts to take it, he will be surprised by the outcome).

7. $R : A \times E \to Z$ a function that expresses the causal link from environment and action to outcome ('R' stands for 'result').[16] Let $r : \Omega \to Z$ be a function that assigns an outcome to each possible world and define it as follows:

$$\forall \omega \in \Omega, \quad r(\omega) = R\left(c(\omega), f(\omega)\right).$$

8. $\succsim$ a binary relation on Z representing the DM's *preferences* over outcomes. The interpretation of $z \succsim z'$ is that the DM considers z to be at least as good as z'. Define (1) $z \succ z'$ as $z \succsim z'$ and not $z' \succsim z$ and interpret it as "the DM prefers z to z' " and (2) $z \sim z'$ as $z \succsim z'$ and $z' \succsim z$ and interpret it as "the DM is indifferent between z and z' ". Let $U : Z \to \mathbb{R}$ (where $\mathbb{R}$ denotes the set of real numbers) be a *utility* function that represents $\succsim$, in the sense that, for all $z, z' \in Z$, $U(z) \geq U(z')$ if and only if $z \succsim z'$. Let $u : \Omega \to \mathbb{R}$ be a function that assigns a utility to each possible world and define it as follows:

$$\forall \omega \in \Omega, \quad u(\omega) = U\left(r(\omega)\right).$$

9. $\mathcal{B} \subseteq \Omega \times \Omega$ a (serial, transitive and euclidean) binary relation on Ω representing the qualitative beliefs of the DM. As before, we denote by $\mathcal{B}(\omega)$ the set of possible worlds that are reachable from ω: $\mathcal{B}(\omega) = \{\omega' \in \Omega : (\omega, \omega') \in \mathcal{B}\}$.

As illustrated in the examples of the previous section, each possible world is described in terms of the external facts or environment, the action taken, the corresponding outcome and the utility of that outcome: this is the role of items 5-8 above.

Concerning the DM's beliefs (item 9) we continue to assume that, for every action that is available to him, the DM considers it possible that he takes that action; this is Assumption (A1), which was discussed in Section 1.2 and which is reproduced below:

$$\forall \omega \in \Omega, \forall a \in A, \quad \exists \omega' \in \mathcal{B}(\omega) \ : \ c(\omega') = a. \tag{A1}$$

[16] Edgington ([7, p. 84] writes that "A properly causal decision theory should be up-front about causation". The function R captures the objective causal link from environment and action to outcome. Kyburg ([17]) might say that the function R represents the DM's *power* to bring about outcome z by taking action a in environment e. The constraints on how the DM should perceive this causal link are discussed below.

For all $\omega \in \Omega$ and $a \in A$, let $\mathcal{B}(\omega, a)$ be the set of possible worlds reachable from ω at which the action taken is a:

$$\mathcal{B}(\omega, a) = \{\omega' \in \mathcal{B}(\omega) : c(\omega') = a\} \tag{2}$$

(thus, by (A1), $\mathcal{B}(\omega, a) \neq \varnothing$).

As stated in item 7 above, we take the function R to be an expression of the objective causal relationship between action/environment and outcome. Any misconception on the part of the DM about the result of taking a given action ought to be a reflection of his incorrect beliefs about the environment, as in the case of Bob in Example 1 who mistakenly believes that the left faucet is connected to hot water.

The DM might be uncertain about what outcome will obtain if he takes a certain action, because he might be uncertain about the environment he is facing. Are there any "rationality" constraints that we should impose on the DM's beliefs in this regard? We impose the following constraint ('I' stands for 'independence'): $\forall a \in A, \forall e \in E, \forall \omega, \omega' \in \Omega$,

$$\begin{array}{l} \text{if } \omega' \in \mathcal{B}(\omega, a) \text{ and } f(\omega') = e, \text{ then,} \\ \forall b \in A, \exists \omega'' \in \mathcal{B}(\omega, b) \text{ such that } f(\omega'') = e. \end{array} \tag{I}$$

According to (I), if the DM considers it possible that he takes action a in environment e then he must also consider it possible that he takes any other available action in that same environment. An equivalent way of stating (I) is as follows. Given a possible world ω and an action a, let $E(\omega, a)$ be the set of environments that the DM considers possible at ω conditional on taking action a:

$$E(\omega, a) = \{e \in E : e = f(\omega') \text{ for some } \omega' \in \mathcal{B}(\omega, a)\}. \tag{3}$$

Then (I) is equivalent to

$$\forall \omega \in \Omega, \forall a, b \in A, \quad E(\omega, a) = E(\omega, b). \tag{I'}$$

Assumption (I) (equivalently, (*I'*)) is a requirement of back and forth independence between action and environment. In one direction, (I) rules out a belief that the action causally determines the environment; for example, in Newcomb's problem ([22]) it rules out the belief "if I take both boxes then the opaque box will be empty and if I take only the opaque box then it will

contain $1,000,000", and in the Prisoner's Dilemma (see, for example, [23]) it rules out the belief "if I cooperate then my opponent will cooperate and if I defect then my opponent will defect".[17] In the other direction, (I) rules out a belief that the environment pre-determines what action will be taken; for example, in Egan's ([8]) psychopath example (see Section 3 for details) it rules out the belief "if I am a psychopath then I am bound/prone to press the button". Even though the DM might be delusional about this, we take it that a *fundamental belief* in a deliberation context is that one is *free to choose*, that is, that one's choice is not pre-determined by the environment. As Kyburg ([17, p. 80]) puts it, "to the extent that I am actually making a choice, I must regard that choice as free". We will further discuss this issue in Section 3.

One could object that, often, actions actually do change the environment; for example if the environment is that the window is closed then my action of opening the window changes the environment, so that if I now take a further action, I will do so in the environment of an open window. This is true, but such a change should be recorded in the *outcome* and, in situations of sequential decisions, one would then specify the new environment as a function of the old environment and the outcome of the previous action.

A theory of rationality when the DM is uncertain about the environment can be developed without postulating that the DM has probabilistic beliefs. There are several notions of rationality that can be used when beliefs are not probabilistic, but merely qualitative, as assumed so far.

One possibility is to define the DM to be rational at possible world ω, where he takes action a, if it is not the case that he believes that another available action b *guarantees* a better outcome in every environment, that is, if – according to his beliefs – it is not the case that an alternative action b strictly dominates action a. Formally, we say that – according to the DM's beliefs at possible world ω – action b *strictly dominates* action a if:

$$\begin{array}{c} \forall \omega_1, \omega_2 \in \mathcal{B}(\omega), \text{ if } c(\omega_1) = a,\ c(\omega_2) = b, \\ \text{and } f(\omega_1) = f(\omega_2) \text{ then } r(\omega_2) \succ r(\omega_1). \end{array} \tag{4}$$

That is, the DM believes that, in every environment that he considers pos-

[17] In this vein, Ahmed rewords Egan's psychopath example as follows (emphasis added): "[...] If you do push button A then it is 99 to 1 that you are a psycho. *This is not because pushing the button makes you a psychopath (it does not)* [...]"

sible, action b yields an outcome that he prefers to the outcome associated with action a.

A stronger (that is, more demanding) definition of rationality is in terms of *weak* dominance: the DM is rational at possible world ω, where he takes action a, if it is not the case that he believes that another available action b weakly dominates action a. Formally, we say that – according to the DM's beliefs at possible world ω – action b *weakly dominates* action a if the following is the case:

$$\begin{gathered} \forall \omega_1, \omega_2 \in \mathcal{B}(\omega), \text{ if } c(\omega_1) = a,\ c(\omega_2) = b \\ \text{and } f(\omega_1) = f(\omega_2) \text{ then } r(\omega_2) \succsim r(\omega_1); \\ \text{furthermore, there exist } \omega', \omega'' \in \mathcal{B}(\omega) \text{ such that} \\ c(\omega') = a,\ c(\omega'') = b,\ f(\omega') = f(\omega'') \text{ and } r(\omega'') \succ r(\omega). \end{gathered} \tag{5}$$

That is, the DM believes that, in every environment that he considers possible, action b yields an outcome which is at least as good as the outcome yielded by action a and, furthermore, there is an environment that he considers possible where b yields a better outcome than a does.

2.2 Probabilistic beliefs and cardinal utility

As noted above, it is common in the literature to define rationality in terms of expected utility maximization, thereby postulating probabilistic beliefs and assuming that the DM has preferences over the set of lotteries on (that is, probability distributions over) the set of outcomes Z that satisfy the axioms of Expected Utility Theory. In this section we will assume this.

We continue to assume that, for every possible world ω, the set of environments that the DM considers possible at ω, conditional on an action, is equal to the set of environments that he considers possible conditional on taking any other action, that is, we continue to assume (I'), which is reproduced below:

$$\forall \omega \in \Omega, \forall a, b \in A, \quad E(\omega, a) = E(\omega, b). \tag{I'}$$

Note that, in virtue of (I'), we can unambiguously define the set of environments that the DM considers possible at world ω, which we denote by $E(\omega)$, as follows:

$$E(\omega) = E(\omega, a) \quad \text{for some } a \in A. \tag{6}$$

For every possible world ω and action a, let $P_{\omega,a} : E(\omega) \to [0,1]$ be the DM's probabilistic beliefs *about the environment*, at ω and conditional on taking action a. Assumption (I') requires that, for any two actions a and b, the support of $P_{\omega,a}$ be the same as the support of $P_{\omega,b}$ (for every possible world ω). We strengthen (I') by requiring that the probabilities also be the same:

$$\forall \omega \in \Omega, \forall a, b \in A, \quad P_{\omega,a} = P_{\omega,b}\,. \tag{I^*}$$

That is, we rule out the possibility that, at any given possible world ω, the DM's probabilistic beliefs *about the environment*, conditional on taking action a, could be different from his beliefs conditional on taking a different action b. Note that, in virtue of (I^*), we can unambiguously define, for every possible world ω, the DM's probabilistic beliefs at ω about the environment as follows:

$$P_\omega : E(\omega) \to [0,1] \text{ is given by } P_\omega = P_{\omega,a} \text{ for some } a \in A. \tag{7}$$

It is worth stressing that *we have not postulated a probability distribution over the set of possible worlds*: doing so would undermine the point stressed in Section 1.2, namely that a theory of deliberation cannot attribute to the DM (probabilistic) beliefs about his own current choices. Since propositions are sets of possible worlds, it follows that we cannot assign probabilities to propositions, in particular, we cannot assign a probability to the conditional "if I take action a the outcome will be z". However, as shown below, it *is* meaningful to ask "if I take action a what is the probability that the outcome will be z?"

Recall from Section 2.1 that, given a possible world ω, $f(\omega) \in E$ is the environment associated with ω and $c(\omega) \in A$ is the action taken at ω; furthermore, $r(\omega) \in Z$ is the outcome associated with ω by means of the causal link from environment and action to outcome expressed by the function $R : A \times E \to Z$, that is, $r(\omega) = R\,(c(\omega), f(\omega))$. Thus, by (7), for every possible world ω and action a, we can derive a probability distribution $P'_{\omega,a} : Z \to [0,1]$ over the set of *outcomes* that the DM considers possible at ω, conditional on taking action a, as follows:

$$P'_{\omega,a}(z) = \begin{cases} 0 & \text{if there is no } e \in E(\omega) \text{ such that } z = R(a,e) \\ P_\omega(e) & \text{if } e \in E(\omega) \text{ and } z = R(a,e) \end{cases} \tag{8}$$

where P_ω is given by (7).

Note that, while – as noted above – (I^*) rules out the possibility that, at a given world, the DM's beliefs about the environment conditional on taking an action could be different from his beliefs conditional on taking a different action, it *is* possible that the DM's beliefs about the *outcomes*, conditional on taking action a, are different from his beliefs about the outcomes conditional on taking a different action b. For example, it is possible that the DM believes that if he smokes he has a higher probability of getting cancer than if he does not smoke. For a simple illustration of this possibility, suppose that there are two "environments": e, representing a genetic disposition by which smoking triggers cancer, and e' representing a genetic makeup where smoking does not lead to cancer. Suppose also that one develops cancer if and only if the environment is e and one smokes, as shown in Figure 7 (where S means 'smoking', $\neg S$ 'not smoking', C 'cancer' and $\neg C$ 'not cancer'). Finally, suppose that the actual world is ω_1 and that, at ω_1, the DM assigns probability 0.7 to e and 0.3 to e'. Then, at ω_1, the DM believes that if he smokes he has a 70% chance of developing cancer while if he does not smoke he is certain of not getting cancer.[18]

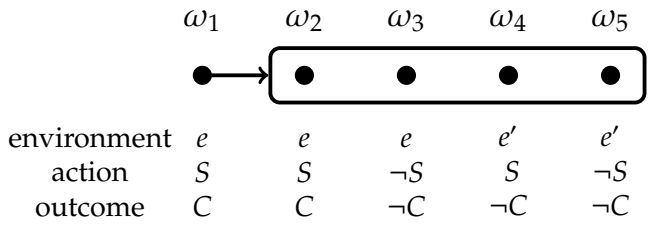


Figure 7: Smoking and cancer.

The utility function $U : Z \to \mathbb{R}$ is now taken to be a von Neumann-Morgenstern utility function that represents the DM's preferences, in the sense that, if $L = \begin{pmatrix} z_1 & \dots & z_m \\ p_1 & \dots & p_m \end{pmatrix}$ and $M = \begin{pmatrix} z_1 & \dots & z_m \\ q_1 & \dots & q_m \end{pmatrix}$ are two lotteries over the set of outcomes $Z = \{z_1, \dots, z_m\}$, then the DM considers L to be at least as good as M if and only if $\mathbb{E}[U(L)] \geq \mathbb{E}[U(M)]$ where $U(L) = \begin{pmatrix} U(z_1) & \dots & U(z_m) \\ p_1 & \dots & p_m \end{pmatrix}$ and $\mathbb{E}[U(L)]$ is the expected value of $U(L)$

[18] As a matter of fact, at ω_1 the DM decides to smoke and gets cancer.

(that is, $\mathbb{E}[U(L)] = p_1U(z_1) + \cdots + p_mU(z_m)$) and is called the *expected utility* of L (similarly for $U(M)$ and $\mathbb{E}[U(M)]$).

Fix a possible world $\omega \in \Omega$ and an action $a \in A$ and let $\mathcal{B}(\omega, a) = \{\omega_1, \ldots, \omega_n\}$. Then, at ω, the DM believes that if he takes action a he will face the lottery

$$L_{\omega,a} = \begin{pmatrix} r(\omega_1) & \ldots & r(\omega_n) \\ P'_{\omega,a}(r(\omega_1)) & \ldots & P'_{\omega,a}(r(\omega_n)) \end{pmatrix}$$

(where $P'_{\omega,a}$ is given by (8)), whose expected utility is (recall that $u(\omega)$ is defined as $U(r(\omega))$)

$$u(\omega_1)P'_{\omega,a}(r(\omega_1)) + \cdots + u(\omega_n)P'_{\omega,a}(r(\omega_n)).$$

In general, the expected utility of taking action a at possible world ω is

$$EU_\omega(a) = \sum_{\omega' \in \mathcal{B}(\omega,a)} u(\omega')P'_{\omega,a}(r(\omega')). \tag{9}$$

Rationality can then be defined as follows.

Definition. *Consider a possible world ω and let $a = c(\omega)$ be the action taken at ω. We say that the DM is* rational at *ω if and only if $EU_\omega(a) \geq EU_\omega(a')$ for every $a' \in A$.*

3 Discussion

In this section we discuss a number of issues that have attracted considerable attention in the literature on deliberation.

3.1 Common causes and Egan cases

Our independence assumption (I) (or (I^*) in the probabilistic case) rules out beliefs that allow the DM to extract evidential value from his own decisions. We now discuss this issue in more detail. The famous Newcomb problem ([22]) sparked an extensive debate in the literature giving rise to two competing decision theories: Evidential Decision Theory (EDT) and Causal Decision Theory (CDT). According to the former, the probabilities used in deliberation should be evidential probabilities, that is, they should

reflect the likelihood, given the agent's total available evidence, that outcomes will occur given acts. According to the latter, the probabilities used in deliberation should be causal probabilities, that is, they should reflect the propensity for acts to produce outcomes. The difference in recommendations between the two theories becomes clear when dealing with common-cause cases. For example, suppose that smoking is strongly correlated with lung cancer because of a common cause, namely a genetic defect that tends to cause both the desire to smoke and lung cancer. Cases like this, where there is correlation due to a common cause for both action and outcome (but no causal link between action and outcome) are taken to be counterexamples to EDT.[19] Recently, Egan ([8]) showed how to modify those counterexamples in such a way that they can be turned into what he takes to be counterexamples to causal decision theory. The "trick" is to make the common cause affect the action as well as an "enabling factor", which, in turn – together with the action – causes the outcome. For example, Egan's version of the smoking example ([8, p. 103]) is as follows: a desire to smoke is produced by a genetic condition which also causes one's lungs to be vulnerable to cigarette smoke, so that smoking causes cancer in those with the genetic condition, but not in those without. In this case, Egan argues, CDT gets it wrong: the DM should follow the recommendation of EDT and decide not to smoke, since the decision to smoke is evidence that the genetic condition is present and, therefore, smoking would cause cancer.[20]

[19] As Edgington explains [7, p.77], under the hypothesis that smoking and cancer are effects of a common cause, but one does not cause the other,

> The conditional probability of getting cancer, given that you smoke, may still be considerably higher than the conditional probability of getting cancer, given that you don't smoke (because smoking is a sign that you have the bad gene). But this is no longer a reason not to smoke. You either have the gene or you don't, and [if you do] refraining from smoking isn't going to reduce your chances of getting cancer.

[20] In Egan's smoking scenario, assuming that the DM attaches sufficiently high probability to *not* having the genetic condition, smoking has a higher causal expected utility than not smoking; this is the reason why CDT would recommend smoking. However, Egan claims that the rational decision in this case is to refrain from smoking because, even if the DM thinks that he probably does not have the genetic condition (and thus that smoking would not cause cancer), matters are different if the DM supposes that he will smoke. For, *on the assumption that he will smoke*, he is likely to have the genetic condition,

The central question is thus: *should a decision theory "allow" the DM to extract evidential value from his own decisions*? We will address this issue in the context of Egan's famous psychopath example ([8, p. 97]) (the wording is taken from [7, p. 81]):

> Paul is debating whether to press the "kill all psychopaths" button. It would, he thinks, be much better to live in a world with no psychopaths. He is discussing this with a friend, who says "Only a psychopath would press such a button." Paul comes to think that might well be true. Paul strongly prefers living in a world with psychopaths to dying. Should Paul press the button?

Egan thinks that Paul should *not* press the button and so does Edgington:

> "[He should not press the button]. He thinks to himself: suppose I do press the button: then that is evidence that I am a psychopath, in which case I will thereby kill myself." ([7, p. 81])

I will argue against such conclusion. Let us remove the sentence

> «Paul is discussing this with a friend, who says "Only a psychopath would press such a button"»

from the above description and let the friend enter into the scene later, in two separate appearances. Paul is smart enough to realize that, if he is a psychopath, then pressing the button will cause also his own death. He thinks carefully about this, reviews his entire life and reaches the conclusion that he is *not* a psychopath. Thus he tentatively decides to press the button, but before doing so he tells his friend "I am inclined to press the button because *I believe that I am not a psychopath*". His friend tells him:

> "I am reading an article, published in a reputable academic journal, stating that there is a highly reliable psychological test to check if a person is a psychopath or not. A very large number of individuals were administered the test, but before undertaking the test they were asked to check one of the three boxes below:

and thus is likely to get cancer. It should be noted that, while some authors agree with Egan (e.g. [7]), others do not (e.g. [14, 28]).

- □ I am *not* a psychopath.
- □ I *am* a psychopath.
- □ I have some uncertainty as to whether or not I am a psychopath.

> Interestingly, **all** the subjects who – based on the result of the test – were later classified as psychopaths, checked the first box, that is, expressed the firm belief that they were not psychopaths. On the other hand, only the fraction q, with $0 \leq q < 1$, of those whom the test classified as non-psychopaths, checked the first box."

Armed with this piece of information, should Paul change his mind? Note that we are asking the question *whether Paul should attach evidential value to his initial beliefs*, not to his planned action. Our answer would be 'No', but we expect that some people will disagree. Let us see where an affirmative answer to this question takes us. Let p be the fraction of the tested population that turned out to be classified as psychopathic, that is, the *base rate* of being a psychopath. Then the empirical conditional probability of being a psychopath (denote this event by S and its complement by $\neg S$) given the belief of *not* being a psychopath (denote this event by B) is

$$\begin{aligned} P(S|B) &= \frac{P(B|S) \times P(S)}{P(B|S) \times P(S) + P(B|\neg S) \times P(\neg S)} \\ &= \frac{1 \times p}{1 \times p + q \times (1-p)} \\ &= \frac{p}{p + (1-p)q} \end{aligned} \tag{10}$$

which is a number between p (when $q = 1$) and 1 (when $q = 0$). Let us have Paul revise his beliefs by now assigning probability $\frac{p}{p+(1-p)q}$ to his being a psychopath and suppose that q is sufficiently small (or the utility of dying is sufficiently low) for the optimal decision now to be *not* to press the button. He then tells his friend: "thank you for the information; I have now revised my beliefs and am now inclined to *not* press the button". His friend replies:

> "I just finished reading this very interesting article. After seeing the result of each test, the experimenter put each subject in front

> of a (fake) button and – lying – told the subject that if he pressed the button all the psychopaths would be killed. Before asking the subject to make a decision, the experimenter asked him to state his preferences over the three possible outcomes and all the subjects had the same preferences as you: 'living without psychopaths' better than 'living with psychopaths' better than 'dying'. Interestingly, all the psychopaths were rational, given their beliefs: they decided to press the button. Of the remaining non-psychopathic subjects, some pressed the button and some did not."

Thus the empirical conditional probability of being a psychopath given the decision of *not* pressing the button is zero! Should Paul attach evidential value to his planned choice of not pressing the button? If we say 'Yes' – as we should, since we allowed him to attach evidential value to his initial beliefs – we should conclude that Paul should now revise his beliefs once more and return to his initial certainty that he is not a psychopath and thus should plan to press the button! We find this circular reasoning to be strongly suggestive that a theory of deliberation should prevent a DM from attaching evidential value to either his beliefs or his action. Thus we agree with Kyburg:

> "The idea is this: to the extent that I am actually making a choice, I must regard that choice as free. To regard it as free is exactly to regard it as *without* the evidential relevance it would have if it were regarded, not as an *act*, but as a bit of behavior. I cannot construe my act as evidence without depriving it of the character of an *act*. [...] But that is not to say that it does not have evidential value for us. *We* can see it as a mere piece of behavior, and as such it has evidential value [...]" ([17, pp. 80-81]).

Should we then conclude that it would be rational for Paul to ignore the information provided by his friend? Perhaps not, but the information to which Paul should pay attention ought to be the base rate p of being a psychopath and if p is large then he should consider switching to more cautious beliefs. However, the base rate is an "objective" piece of information and reacting to it has nothing to do with attaching evidential value to his own beliefs or action.

3.2 Causation, beliefs and propensity to act

It could be objected that the semantics proposed in Section 2 is too restrictive, in that – after all – it seems unable to accommodate "pure-common-cause" cases as well as "Egan" cases. For example, in the pure-common-cause smoking example, it is postulated that the genetic factor causes both lung cancer and *the desire to smoke*, while in the Egan version of the example the genetic factor causes the desire to smoke as well as vulnerability of the lungs to cigarette smoke (so that smoking then causes cancer in the presence of the genetic factor, but not without). Can we model the fact that the environment causes a *desire* or *propensity* to act in a certain way? The answer is affirmative: all we need to do is make a "small" change by allowing the utility function to be dependent on the actual environment, via the actual world: instead of postulating a utility function $U : Z \to \mathbb{R}$ we would postulate a utility function parameterized by the environment $e \in E$:

$$U_e : Z \to \mathbb{R},$$

so that, at world ω the utility function is $U_{f(\omega)} : Z \to \mathbb{R}$. For example, if ω is a possible world where the environment consists of the presence of the genetic factor, then the utility of the outcome of smoking would be larger than the utility of the outcome of smoking at a different possible world ω' where the environment is such that the genetic factor is absent. Thus the act of smoking would be more desirable for the DM at the former possible world than at the latter. This raises subtle issues concerning the DM's awareness of how the environment affects his preferences, which are left for future research.

3.3 Conditional probability versus probability of the conditional

The debate between EDT and CDT boils down to whether the beliefs used in deliberation should be conditional probabilities or probabilities of counterfactual conditionals. Such notions are not relevant within the framework proposed in this paper. First of all, in the probabilistic case *we do not postulate a probability distribution over the set of possible worlds*: doing so would undermine the point stressed in Section 1.2, namely that a theory of deliberation cannot attribute to the DM (probabilistic) beliefs about his

own current choices. Since propositions are sets of possible worlds, it follows that we cannot assign probabilities to propositions, in particular, we cannot assign a probability to the conditional "if I take action a the outcome will be z". However, as explained in Section 2.2, it *is* meaningful to ask "if I take action a what is the probability that the outcome will be z?"

The probabilities that we postulated are probabilities about the environment and can be thought of as conditional probabilities, that is, conditional on a given action; however, our independence assumption (I^*) requires these probabilities to be the same, conditional on any other action. As shown in Section 2.2, however, (I^*) is compatible with the conditional probabilities over *outcomes* to be action dependent.

3.4 'If I take ...' versus 'if I were to take ...'

We stated in Section 1 that we take the indicative mood "If I take action a the outcome will be x" and the subjunctive mood "If I were to take action a the outcome would be x" to express essentially the same conditional, with only a pragmatic difference: the former signals that the decision whether to take action a is still "open", while the latter seems to convey that the speaker is somehow ruling out taking action a. We can make this more precise, by identifying three stages in the deliberation process: (1) the pre-choice stage, (2) the after-choice, but pre-action, stage and the (3) after-action stage.

The semantics proposed in Section 2 is intended to model the pre-choice (or deliberation) stage, during which the DM considers the consequences of all his actions, without pre-judging his subsequent decision. At this stage the DM has not made a decision yet and thus all his options are still "open". This is the stage at which Bob (of Section 1.1) is looking at the faucets and tells himself "if I open the right faucet I will get cold water".

In the post-choice, but pre-action, stage the DM has made up his mind, but not acted yet. At this stage Bob has reached the decision to turn on the left faucet, has extended his hand to do so and – if asked to explain the reason for his intended action – he would have to say "if I were to open the other faucet (the right faucet) I would get cold water". It would not make sense for him to utter a different sentence (e.g. that he would expect the right faucet to deliver hot water), since he has received no new relevant information. How should we model the DM's beliefs at the post-choice but

pre-action stage? Clearly they should be related to his pre-choice beliefs. Denote by $\mathcal{B}_0$ the belief relation at the pre-choice (or deliberation) stage (time 0) and by $\mathcal{B}_1$ the belief relation at the post-choice stage (time 1). Recall that, for every possible world $\omega \in \Omega$, $c(\omega) \in A$ is the action taken at ω. We suggest that the relation $\mathcal{B}_1$ should be defined as follows:

$$\forall \omega \in \Omega, \quad \mathcal{B}_1(\omega) = \mathcal{B}_0(\omega, c(\omega)) \tag{11}$$

where, as in (2), $\mathcal{B}_0(\omega, c(\omega)) = \{\omega' \in \mathcal{B}_0(\omega) : c(\omega') = c(\omega)\}$.

Thus the relation $\mathcal{B}_1$ captures the fact that, after deliberation, the DM *gets to know what action he will take*. For the case of Example 1 of Section 1.1 (Bob in the shower) the relations $\mathcal{B}_0$ and $\mathcal{B}_1$ are shown in Figure 8 (the top part, namely the relation $\mathcal{B}_0$, reproduces Figure 3).

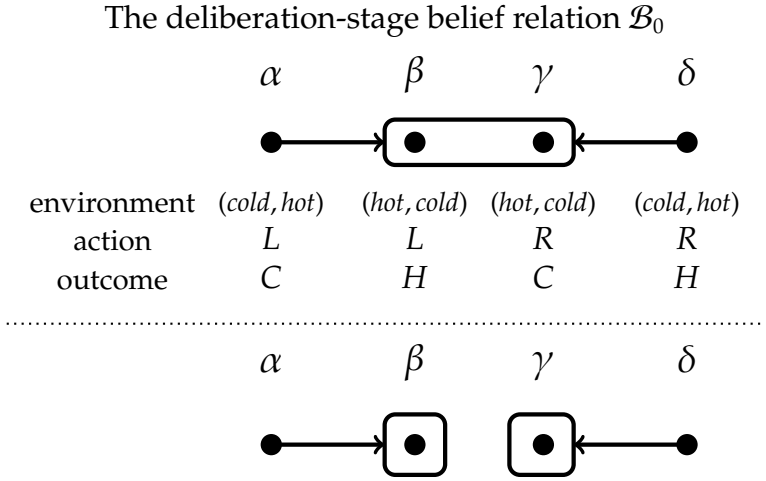


Figure 8: The deliberation-stage and post-choice-pre-action stage beliefs in Example 1 (Bob in the shower).

Let B_0 be the belief operator associated with the relation $\mathcal{B}_0$ and B_1 the belief operator associated with the relation $\mathcal{B}_1$. Thus, as before, at a world ω it is true that, at time $t \in \{0, 1\}$, the DM believes formula ϕ, denoted by $\omega \models B_t\phi$, if and only if $\mathcal{B}_t(\omega) \subseteq \|\phi\|$. Since, by (11), for every world ω, $\mathcal{B}_1(\omega) \subseteq \mathcal{B}_0(\omega)$ it follows that everything that the DM believed at state ω

in the pre-choice stage he still believes in the after-choice stage: for every formula ϕ, if $\omega \models B_0\phi$ then $\omega \models B_1\phi$. However, there are formulas that were not believed at date 0 but are believed at time 1. For instance, if a is an action *different* from the one taken at state ω, that is, if $a \neq c(\omega)$ (so that $\omega \models \neg a$) then, for every outcome x, the material conditional $a \rightarrow x$ is believed at time 1, because of the false antecedent. For example, in the case illustrated in Figure 8, we have that $\alpha \not\models B_0(R \rightarrow H)$ while, trivially, $\alpha \models B_1(R \rightarrow H)$.

Thus sentences of the form "if I were to take action a the outcome would be x" can no longer be treated as material conditionals in the post-choice stage, because of the possibility of a (known) false antecedent: they have to be construed as *subjective counterfactuals*. Denote the counterfactual "if I were to take action a, the outcome would be x" by $a \rightsquigarrow x$. The question is: what should the validation rule for such counterfactuals be? Should we make use of the Stalnaker-Lewis theory of counterfactuals and declare $a \rightsquigarrow x$ to be true at world ω if and only if the most similar world to ω where a is true is such that x is also true?[21] We claim that the answer is 'No': the conditional "if I were to take action a, the outcome would be x" ought to be interpreted as a *re-statement, in the subjunctive mood, of the pre-choice beliefs after the choice is made*; in other words, "if I were to take action a, the outcome would be x" is assertable at time 1 if and only if "I believe that if I take action a the outcome will be x" was true at time 0. Thus we propose that the validation rule for $a \rightsquigarrow x$ should be as follows:

$$\omega \models (a \rightsquigarrow x) \text{ if and only if } \omega \models B_0(a \rightarrow x). \tag{12}$$

That is, $a \rightsquigarrow x$ is true at world ω (at time 1) if and only if the material conditional $a \rightarrow x$ was believed to be true at ω in the pre-choice stage (at time 0). Note that it follows from transitivity of $\mathcal{B}_0$ that, for every possible world $\omega \in \Omega$, if $\omega \models a \rightsquigarrow x$ then $\omega \models B_1(a \rightsquigarrow x)$.[22]

[21] With the usual understanding that if $\omega \models a$ then ω itself is the unique closest world where a is true.

[22] By transitivity of $\mathcal{B}_0$ (positive introspection of belief), for every formula ϕ, the formula $B_0\phi \rightarrow B_0B_0\phi$ is valid, that is, true at every possible world; in particular, $B_0(a \rightarrow x) \rightarrow B_0B_0(a \rightarrow x)$ is valid. As noted above, for every formula ϕ, the formula $B_0\phi \rightarrow B_1\phi$ is valid; in particular, $B_0B_0(a \rightarrow x) \rightarrow B_1B_0(a \rightarrow x)$ is valid. Finally, since, by (12), $B_0(a \rightarrow x)$ is equivalent to $a \rightsquigarrow x$, it follows that $B_1B_0(a \rightarrow x)$ is equivalent to $B_1(a \rightsquigarrow x)$. Thus, for every $\omega \in \Omega$, $\omega \models a \rightsquigarrow x$ if and only if $\omega \models B_0(a \rightarrow x)$ only if $\omega \models B_0B_0(a \rightarrow x)$ only if $\omega \models B_1B_0(a \rightarrow x)$ if and only if $\omega \models B_1(a \rightsquigarrow x)$.

Our proposal is thus to model the after-choice-pre-action beliefs as a suitable adaptation of the pre-choice beliefs in such a way as to give the DM knowledge of the action that he will perform, while appropriately restating the believed links between actions and outcomes as "counterfactuals", without altering their substantive content.[23]

Finally, there is also a post-action stage. Here we need to distinguish between the case where the outcome is the one that the DM was expecting and the case where the observed outcome takes the DM by surprise. For example, modify the story of Bob by giving him correct beliefs, so that if he turns on the left faucet then, indeed, he gets hot water, as he thought he would. In this case – if asked to justify his action – Bob would have to say 'if I had turned on the right faucet I would gave gotten cold water', that is, he re-states his original belief but in the form of a counterfactual; after all, he has not received any information that contradicts those initial beliefs.[24] In the alternative case of surprise, where the outcome of the action performed is not as expected, then the DM would most likely not re-assert his pre-choice beliefs (although he might: Bob could infer, correctly or incorrectly, that all the hot water had been used and that it would still be true that turning on the right faucet would have yielded cold water).

Thus what we are suggesting is that, in the context of deliberation, the difference between the three sentences:

1. "If I take action a the outcome will be x",
2. "If I were to take action a the outcome would be x",
3. "If I had taken action a the outcome would have been x"

is one of *timing*: the first is uttered at the pre-choice stage, the second at the post-choice, but pre-action, stage and the third (assuming no surprises) at the post-action stage. However, all three statements are expressions of the *same* beliefs, which at the deliberation stage can be modeled as material conditionals.

In the example of Figure 8, we have that $L \rightsquigarrow H$ and $R \rightsquigarrow C$ are true at every world and thus believed at date 1, that is, $\omega \models B_1(L \rightsquigarrow H) \wedge B_1(R \rightsquigarrow C)$ for every $\omega \in \{\alpha, \beta, \gamma, \delta\}$.

[23] We put quotation marks around the word 'counterfactual' to stress that we are not thinking of such formulas as counterfactuals in the sense that is normally understood, namely in the objective sense of the Stalnaker-Lewis theory.

[24] Note, however, that this would still be a *subjective* counterfactual: it might very well be that both faucets were connected to hot water and thus the stated counterfactual would be objectively false, but still subjectively true and thus assertable by Bob.

References

[1] R. Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, 55:1–18, 1987.

[2] R. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6–19, 1995.

[3] P. Battigalli and G. Bonanno. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53:149–225, 1999.

[4] K. DeRose. The conditionals of deliberation. *Mind*, 119:1–42, 2010.

[5] I. Douven. *The epistemology of indicative conditional*. Cambridge University Press, Cambridge, 2016.

[6] D. Edgington. On conditionals. In D. M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, 2nd edition*, volume 14, pages 127–222. Springer, 2007.

[7] D. Edgington. Conditionals, causation and decision. *Analytic Philosophy*, 52(2):75–87, 2011.

[8] A. Egan. Some counterexamples to causal decision theory. *The Philosophical Review*, 116(1):93–114, 2007.

[9] A. Gibbard. Two recent theories of conditionals. In W. L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs: conditionals, belief, decision, chance, and time*, pages 211–247. Springer Netherlands, Dordrecht, 1981.

[10] A. Gibbard and W. L. Harper. Counterfactuals and two kinds of expected utility. In W. L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs: conditionals, belief, decision, chance, and time*, pages 153–190. D. Reidel, 1978.

[11] I. Gilboa. Can free choice be known? In C. Bicchieri, R. Jeffrey, and B. Skyrms, editors, *The logic of strategy*, pages 163–174. Oxford University Press, 1999.

[12] C. Ginet. Can the will be caused? *The Philosophical Review*, 71:49–55, 1962.

[13] A. Goldman. *A theory of human action*. Princeton University Press, 1970.

[14] J. Joyce. Regret and instability in causal decision theory. *Synthese*, 187:123–145, 2012.

[15] K. Krzyżanowska. Deliberationally useless conditionals. *Episteme*, 17(1):1–27, 2020.

[16] K. Krzyżanowska, S. Wenmackers, and I. Douven. Rethinking Gibbard's riverboat argument. *Studia Logica*, 102(4):771–792, 2014.

[17] H. E. Kyburg. Powers. In W. L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics: Proceedings of the Irvine Conference on Probability and Causation*, pages 71–82. Springer Netherlands, Dordrecht, 1988.

[18] M. Ledwig. The no probabilities for acts-principle. *Synthese*, 144:171–180, 2005.

[19] I. Levi. *Hard choices*. Cambridge University Press, 1986.

[20] I. Levi. *The covenant of reason: rationality and the commitments of thought*. Cambridge University Press, 1997.

[21] D. Lewis. *Counterfactuals*. Harvard University Press, 1973.

[22] R. Nozick. Newcomb's problem and two principles of choice. In N. Rescher, editor, *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of his Sixty-Fifth Birthday*, pages 114–146. Springer Netherlands, Dordrecht, 1969.

[23] M. Peterson, editor. *The Prisoner's Dilemma*. Classic Philosophical Arguments. Cambridge University Press, Cambridge, 2015.

[24] K. Reuter and P. Brössel. No knowledge required. *Episteme*, 16(3):303–321, 2019.

[25] W. Spohn. Where Luce and Krantz do really generalize Savage's decision model. *Erkenntnis*, 11:113–134, 1977.

[26] W. Spohn. *Strategic Rationality*, volume 24 of *Forschungsberichte der DFG-Forschergruppe Logik in der Philosophie*. Konstanz University, 1999.

[27] R. Stalnaker. A theory of conditionals. In N. Rescher, editor, *Studies in logical theory*, pages 98–112. Blackwell, 1968.

[28] T. L. Williamson. Causal decision theory is safe from psychopaths. *Erkenntnis*, Apr 2019.