THE METHOD OF THOUGHT EXPERIMENTS: PROBABILITY AND COUNTERFACTUALS [*]

The method of thought experiments pops up in *Counterfactuals*, section 3.2. Lewis considers the idea that the assertable counterfactuals are those whose consequent follows from the supposition of the antecedent, together with further unstated premises thought to be cotenable with it:

> Imagine that you somehow came to know the antecedent for certain and reorganized your system of beliefs under the impact of this new knowledge: the beliefs you would retain are the ones you regard as cotenable with the antecedent. The problem of cotenability then reduces, as Mackie observes, to the familiar problem of induction: how should one's system of beliefs change under the impact of an exogenous piece of new knowledge?
>
> But the method of thought experiments is wrong. [...] There is no reason at all why my most probable antecedent-worlds should be the same as the antecedent-worlds closest to my most probable worlds. The method of thought experiments gives me the character of the former worlds, but the assertability of counterfactuals depends on the character of the latter worlds.
>
> Perhaps I have considered the wrong thought experiment; the right one is to add your antecedent to your system of beliefs not as if it were an item of new knowledge, but simply *as a counterfactual supposition.* That is the right thing to do, I agree, but it is unhelpful to say so. For what is the thought experiment of adding φ to your beliefs as a counterfactual supposition? I suppose it is nothing else than the exercise of deciding which counterfactuals with the antecedent φ you believe. (Lewis 1973, 70–2)

Indeed, Lewis starts by considering the wrong kind of thought experiment for the assessment of counterfactuals: the right one involves taking the antecedent φ as a counterfactual supposition. But, contrary to what he claims, this is not quite the same as the exercise of deciding which counterfactuals with antecedent φ one believes or accepts. Such an exercise is given by counterfactual supposition, *plus* one fur-

[*]. This work has been presented at the Philosophy Research Seminar, University of Nottingham (May 2024), and at the LOGOS Seminar, University of Barcelona (June 2024). We thank the audiences, as well as two anonymous reviewers of this Journal, for their valuable comments.

ther ingredient which, once added, will shed light on the acceptability of counterfactuals.[1]

A *simple* counterfactual '$\varphi > \psi$' ('If it were/had been the case that φ, then it would be/have been the case that ψ') has no counterfactuals nested in φ or in ψ. We are after their assessment and acceptability conditions.[2] What is acceptability? One may take it just as believability (Douven 2016, ch. 4). However, we will consider situations where these may come apart. For we'll make comparisons with the acceptability of indicatives ('$\varphi \rightarrow \psi$', 'If it is the case that φ, then it is the case that ψ'). And if, as Adams (1975), Edgington (1995), Bennett (2003) think, these do not express propositions and lack truth values, then one should not speak of believability for them: to believe something is to believe it is true.

Why acceptability rather than assertability? We take acceptance as a mental state, assertion as a linguistic act expressing acceptance or belief, and we focus on the mental state: we are after a cognitive phenomenon, not the pragmatics of its communication. We accept in the privacy of our head. So acceptability is not subject to social norms the way assertability is: one may find something quite acceptable or believable, but inappropriate to assert in a conversational context because it would be weird, an insensitive thing to say, or so. We will also refer, to draw parallels with counterfactuals, to some experiments concerning indicatives which have been carefully designed in terms of acceptability, not assertability.

Why only simple counterfactuals? One reason is practical: below, we come up with a probabilistic logic to reason about the (un)acceptability of counterfactuals. Doing that for simple counterfactuals was complex enough: we resort to a probabilistic belief update different from Bayesians' favourite i.e. conditionalization, and we endow our models with an algebraic structure. This is new logical territory: while the interactions of conditionals with probabilities have been formally investigated at least since the groundbreaking work of Adams (1966, 1975, 1998), most accounts only make use of conditional probabilities. They

[1] So we borrow the expression 'method of thought experiments' from Lewis, to use his label for candidate procedures for the assessment of counterfactuals. But of course there are broad connections betweens counterfactuals and thought experiments in science and philosophy, explored e.g. in (Williamson 2007; Shaffer 2017). We'll get back to this.

[2] We are actually after something more restricted: our *primary* way of assessing – via counterfactual supposition. We sometimes accept a counterfactual just by taking on board someone's testimony. But we take this to be a secondary way, parasitic on the primary: Williamson (2020), ch. 2, has a similar distinction, and dependence, between primary (via a suppositional procedure) and secondary (via testimony) ways of assessing indicatives.

also do not embed an algebraic component. So we made our life easier by not working with nested counterfactuals.

Another reason is methodological: one defeasible test for an account of the acceptability conditions of counterfactuals is, we submit, how well it matches various intuitive acceptability judgments. But with nested counterfactuals such judgments may fail to provide clear verdicts. This is not unexpected: to the extent that we assess counterfactuals suppositionally, nested counterfactuals may have us engage in suppositions within suppositions; these may induce more cognitive strain and divergent verdicts than plain suppositions. But with simple counterfactuals, at least for certain relevant cases we will discuss, intuitions speak with a uniform voice, thus being harder to explain away via some error theory – or so we argue below.

We do *not* stick with simple counterfactuals because we believe they are not freely embeddable, as some think, as a consequence of Lewis' and others' celebrated triviality results concerning their probabilities (Lewis 1976; Hájek 1989). Analogous triviality results for counterfactuals have been produced (Leitgeb 2012; Williams 2012). But our account will be triviality-proof.

I. SUPPOSITIONS

Suppositional accounts of indicatives are popular in philosophy (Edgington 1995) and psychology (Evans and Over 2004; Oaksford and Chater 2010). They are inspired by Ramsey's footnote:

> If two people are arguing 'If p will q' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; so that in a sense 'If p, q' and 'If p, $\neg q$' are contradictories. We can say that they are fixing their degrees of belief in q given p. (Ramsey 1990, 155n)

(What we now call) the Ramsey Test links our assessment of conditionals to the update of our prior beliefs in the light of new information – except that instead of actually getting the news online, we just imagine getting them, in 'offline mode' (Williamson 2016). Suppositional thinking, then, must work as a kind of simulated belief revision governed, just as its real counterpart, by a maxim of minimal alteration: we change our beliefs as little as possible, compatibly with the need to accommodate the supposition.

Matching the distinction between the two kinds of conditionals, it is common in the literature (Joyce 1999; Leitgeb 2012, 2017) to distinguish between indicative supposition (imagining how things are like if φ is the case) and subjunctive or counterfactual supposition (imagining what things would be or have been like if φ was or had been the case).

Lewis' discussion of the method of thought experiments in *Counterfactuals* is aimed as showing that a key difference between the two lies in which beliefs are cotenable in the two modes. The first method gives the wrong kind of thought experiment for the assessment of counterfactuals precisely because it delivers the wrong verdict on cotenability. Section 3.2 repeats the example famously given at the start of the book:

1. If Oswald did not kill Kennedy, someone else did.
2. If Oswald had not killed Kennedy, someone else would have.

We can assess both by supposing the antecedent and wondering about the consequent. But we accept (1) and reject (2) because, when we suppose indicatively that Oswald has not killed Kennedy we retain as cotenable our belief that Kennedy was actually killed – and so it must have been someone else, so (1) must be ok. When we suppose the same thing counterfactually, we relinquish that belief and find it plausible that nobody else kills Kennedy in the counterfactual scenario – so (2) must not be ok.

Suppositional accounts take indicative supposition as governed by conditionalization (Adams 1975; Edgington 1995; Bennett 2003; Evans and Over 2004). But if indicative and counterfactual supposition differ, what is the latter governed by? Lewis gave the answer three years after the publication of *Counterfactuals*, in the very same paper in which he came up with the triviality results for the probabilities of indicatives: it is governed by a procedure from which conditionalization differs just as looking at 'the most probable antecedent-worlds' differs from looking at 'the antecedent-worlds closest to the most probable worlds'.

II. IMAGING

Take a finite set of worlds W on which a total closeness ordering is defined, as per the conditional logics of Stalnaker (1968) and Lewis (1973). When Lewis (1976) introduced the procedure at issue, ***imaging***, he followed Stalnaker in assuming that for each w and φ there is a single closest φ-world, w_φ. A selection function $f : W \times \mathcal{P}(W) \mapsto W$ outputs, for each $w \in W$ and $|\varphi| \subseteq W$ ($|\varphi|$ being the truth set of φ), the relevant w_φ. Given a probability distribution π over W, the ***image*** π^φ of π under φ is defined: for all $w_1 \in W$, $\pi^\varphi(w_1) := \sum_{w:w_\varphi = w_1} \pi(w)$. The probability of each w is transferred to its closest φ-world w_φ. Because w_φ may be the closest world to more than one world, one adds up the probabilities of all of those worlds. Each φ-world keeps the probability it had before, and may gain probabilities transferred from non-φ-worlds. Probabilities are only moved around but not created or destroyed, so π^φ is a probability distribution when π is. We then define, as usual, the probability of a sentence ψ as the sum of the probabilities of the worlds where ψ is true, so $\pi^{\varphi}(\psi) := \sum_{w_1 \in |\psi|} \pi^{\varphi}(w_1)$.

Both imaging and conditionalization comply with the idea of minimal change governing suppositions – but they are minimal in different ways:

> Imaging π on φ gives a minimal revision in this sense: unlike all other revisions of π to make φ certain, it involves no gratuitous movement of probability from worlds to dissimilar worlds. Conditionalizing π on φ gives a minimal revision in this different sense: unlike all other revisions of π to make φ certain, it does not distort the profile or probability ratios, equalities, and inequalities among sentences that imply φ. (311, notation adjusted)

The simple Lewisian example illustrating how imaging $\pi^{\varphi}(\psi)$ differs from conditionalization $\pi(\psi|\varphi)$ goes thus: say we have three equiprobable worlds w, w_1, w_2 so the probability of each one is $1/3$. w_1 and w_2 make φ true while w does not. w_1 is closer to w than w_2. When we revise by conditionalizing on φ, we kick out w and renormalize, distributing the probabilities uniformly so $\pi(w_1|\varphi) = \pi(w_2|\varphi) = 1/2$. Instead, imaging makes use of closeness: all of w's probability is transferred to w_1, thus $\pi^{\varphi}(w_1) = 2/3$ while $\pi^{\varphi}(w_2) = 1/3$.

Here is how the two differ in cotenability: Gärdenfors (1982) proved that conditionalization, unlike imaging, has a property, *conservativity*, which, phrased in terms of supposition, goes thus: when one supposes indicatively that φ and one is certain of χ for a cotenable χ, i.e., $\pi(\chi) = 1$, then also $\pi(\chi|\varphi) = 1$. So when one wonders what is the case if Oswald did not kill Kennedy and one is certain that Kennedy has been killed, one will retain that certainty under the supposition – thus, indicative-Oswald (1) sounds ok. Imaging is not conservative: χ may become uncertain when one supposes that φ counterfactually. So when one wonders what would have been the case if Oswald had not killed Kennedy, one may relinquish one's certainty that Kennedy was killed – thus, counterfactual-Oswald (2) does not sound ok.

Thought experiments and counterfactuals are involved in the assessment and revision of scientific theories. For interesting works on the subject, one can look at Shaffer (2001, 2012), which connect the idealisation of scientific theories to counterfactuals dealt with in Lewisian-Stanakerian fashion: Shaffer proposes that counterfactuals capture the simplyfing assumptions occurring in standard scientific theory-building. Now theory revision can be broadly understood as minimal revision of a theory given new evidence or information. And Shaffer (2001) rightly remarks that, if we regiment theories via counterfactuals, Bayesian probabilistic revision via conditionalization may have troubles in accounting for the prior probabilities of theories in counterfactual form. We remark that imaging captures exactly a form of minimal revision which could be applied to theories, alternative to standard Bayesian conditionalization, and proven free from the conservativity feature of the latter, thanks to the Gärdenfors result.

III. ADAMS' THESIS, LEWIS' THESIS, STALNAKER'S HYPOTHESIS, LEWIS' PROOF

A popular conjecture, subscribed to by McGee (1989), Jackson (1987), and others, has it that the (degree of) acceptability of indicatives equals the subjective probability of the consequent conditional on the antecedent. One may think that the (degree of) acceptability of counterfactuals equals the subjective probability of the consequent under the image of the antecedent:

(AT) $Acc(\varphi \to \psi) = \pi(\psi|\varphi)$

(LT) $Acc(\varphi > \psi) = \pi^{\varphi}(\psi)$

(LT) is Lewis' Thesis (as one may call it, given *Counterfactuals* 3.2): deciding which counterfactuals with antecedent φ one accepts is nothing else than assessing the status of the consequent under the counterfactual supposition of φ – once the latter is understood as imaging on φ. (AT) is Adams' Thesis, named after Adams (1966, 1975). This is sometimes, but should not be, confused with Stalnaker's Hypothesis, named after Stalnaker (1975):

(SH) $\pi(\varphi \to \psi) = \pi(\psi|\varphi)$

Two differences between (AT) and (SH): (1) (AT) applies only to simple indicatives (no indicatives embedded in the antecedent or consequent); (2) (AT) is phrased in terms of acceptability, not probability. Both seem to be required: (SH) (in spite of robust empirical evidence confirming it: see Evans, Handley, and Over (2003), Evans and Over (2004), and Douven and Verbrugge (2010)) is widely considered false precisely because of the Lewis (1976) triviality results. On the other hand, non-propositionalists like Adams, Edgington (1995), and Bennett (2003) have claimed that (AT) can hold insofar as indicatives do not express propositions (thus, they cannot be freely embedded: hence the restriction to simple conditionals) and lack truth values. Because of this, they cannot have probabilities of truth properly so called, as per (SH); but they can have acceptability conditions, as per (AT).

But Lewis (1976) proved that when $\varphi > \psi$ is the Stalnaker conditional, true iff the closest antecedent-world makes the consequent true, then the following *does* hold in general for all probability distributions π, with no triviality ensuing – let us call it Lewis' Proof:

(LP) $\pi(\varphi > \psi) = \pi^{\varphi}(\psi)$

A Stalnaker counterfactual $\varphi > \psi$ is true at w iff ψ is true at w_φ. So its truth set is $\{w : w_\varphi \Vdash \psi\}$ ('$\Vdash$' is *makes true*). Now the probability of this set is $\pi(\{w : w_\varphi \Vdash \psi\}) = \sum_{w:w_\varphi \Vdash \psi} \pi(w)$. But this *is* the probability of ψ under the image of φ, $\sum_{w_1 \in |\psi|} \sum_{w:w_\varphi = w_1} \pi(w)$. If we assume that the degree of acceptability of a counterfactual equals its subjective probability, we get (LT): we assess a counterfactual by checking how the consequent fares once we have shifted the due probabilistic mass to the closest antecedent-world. Then the situation with counterfactuals (insofar as the Stalnaker conditional is taken as a good enough approximation to the natural language counterfactual – we'll get back to this) is somewhat streamlined. Lewis' positive result (LP) reassures us that, unlike what happens with indicatives, we need not divorce acceptability from probability of truth.

(LT) may then be taken as describing in full generality our primary way of assessing a (simple) counterfactual suppositionally: we accept $\varphi > \psi$ to a degree equal to the probability we assign to the consequent ψ under the image of the antecedent φ. In terms of supposition: we accept it to the extent that we judge the consequent likely in a counterfactually imagined situation in which the antecedent is true.

IV. RELEVANCE

But (LT) is not quite right. We often find a counterfactual unacceptable although we judge the probability of the consequent high under the counterfactual supposition of the antecedent. That is because the consequent was already deemed very likely outside of the supposition, and the antecedent we image our probabilities on is irrelevant to that:

3. If this Banbury house were in Oxford, then Melbourne would be in Australia.

4. If Caesar had had firearms in 58 BC, then Saturn would have been a planet.

5. If we offered Midori a pay rise, then there would be some heads in the first 100 tosses of this fair coin.

Can one cordon off the anomalies by claiming that these happen just with propositions one is already fully certain of, as may be the case with the consequent of (3) or (4)? One would have to explain why: the Gärdenfors result tells us that certainties are not guaranteed to be preserved under imaging. Anyway, (5) is different: knowing the coin is fair, we think it quite likely that it will land heads sometimes if tossed 100 times, but we are not certain.

What will end up being acceptable at the end of the exercise is a concessive: 'Even if this Banbury house were in Oxford, Melbourne would still be in Australia'; 'There would be some heads in the first 100 tosses of this fair coin, whether or not we offered Midori a pay rise'. Unlike ordinary counterfactuals, concessives can take 'even' or 'whether or not' in the antecedent; they cannot take 'then' in the consequent, as it is there precisely to exclude the irrelevance of the antecedent for the consequent (Iatridou (1993) makes the point for indicatives). Perhaps concessives and ordinary counterfactuals differ in truth conditions, perhaps not. But we're only after acceptability; and it seems clear that 'If it were the case that φ, then it would be the case that ψ' is normally acceptable precisely when 'Even if it was the case that φ, it would still be the case that ψ' and 'Whether or not φ, it would be the case that ψ' are not. Compare:

6. If this Banbury house were in Oxford, then it would be very expensive.
7. If Caesar had had firearms in 58 BC, then he would have used them against the Gauls.
8. If there were some heads in the first 10 tosses of this (fair) coin, then there would be some heads in its first 100 tosses.

We accept these because, besides finding the consequent likely on the counterfactual supposition of the antecedent, we find the latter relevant, or on-topic, or pertinent, with respect to the former. The relevance effect is detectable even in cases where the antecedent is necessarily false, and known to be:

9. If there was a largest prime, then some number would be both prime and composite.
10. If there was a largest prime, then bachelors would be married.

We can accept (9) while in the business of proving the infinity of primes. We already regard the antecedent as impossible: we are just carrying out a proof by *reductio* of what we already know to be a necessary truth (perhaps because we trust the textbook, or the teacher, telling us this much). (10) would not fare so well in the context. The difference between (9) and (10) gives evidence that the acceptability conditions for counterfactuals are hyperintensional: we can have different attitudes towards counterfactuals whose antecedent and consequent are, respectively, co-intensional – and known to be such.

To counter the conjecture that (LT) *describes* what we primarily do when we assess a counterfactual, we need the counterexamples to be pervasive. People do all sorts of things: isolated exceptions will not count against a descriptive conjecture. So, how widespread are relevance-based counterexamples? We are not aware of experiments carried out to test (LT). But we are, of experiments carried out to test its counterpart for indicatives, (AT). Douven and Verbrugge (2010) gave to a group of subjects contexts (short stories) C_i, $1 \leq i \leq 30$, and asked them to rate the acceptability of indicatives $\varphi_i \to \psi_i$ in C_i. They gave to another group the same contexts C_i and asked them to judge the probability of ψ_i in C_i on the supposition that φ_i. It turned out that people's patterns (of degrees) of acceptance for indicatives do not even approximate the corresponding conditional probabilities: the acceptability ratings are often significantly lower than the conditional probabilities. This 'manifestly refute[s] Adams' Thesis, both in its strict form AT and in its approximate form' (Douven 2016, 99). Douven's own favourite *inferentialist* approach to indicatives (a label for a family of accounts, including Skovgaard-Olsen, Singmann, and Klauer (2016), Krzyżanowska (2015), Krzyżanowska, Collins, and Hahn (2017), Rott (2022)) explains this in terms of the lack of a relevant connection between antecedent and consequent in conditionals with corresponding high conditional probability.[3]

We do not know how straightforwardly such negative empirical results concerning (AT) may carry over to (LT). But they point at an obvious pervasive feature of suppositional thinking in general, whether

[3] Unlike (AT), Stalnaker's Hypothesis (SH) enjoys big empirical support. But Douven and Verbrugge were the first, as far as we know, to phrase an experiment specifically in terms of acceptability. They even considered the idea that their subjects may have confused acceptability with assertability. So they also came up with control experiments, where they (1) compared answers to questions phrased in terms of acceptability with answers phrased directly in terms of reasonableness to believe; and (2) explicitly asked the participants of the main experiment how they themselves had interpreted acceptability:

> The answers do suggest that the notion of acceptability was interpreted in an epistemic sense rather than in some other sense; things that seem logical, or self-evident, or that can be taken to be true, are probably things that are reasonable to believe, though not obviously also things that it would be appropriate to contribute to a conversation. Indeed, there was no indication that any of the participants had understood 'acceptable' as meaning something like 'conforming to broadly social norms governing good conversational practice.' (Douven and Verbrugge 2010, 311)

of the indicative or of the counterfactual kind: such thinking has a *focus.* We usually engage in suppositional thinking with an issue to address: supposing φ, would it be that ψ? Lots of things can then turn out to be irrelevant to the issue, even when they are otherwise perfectly cotenable. Lewis (1973) claimed that necessary or logical truths should be cotenable with any supposition. But when we counterfactually suppose that the house is moved to Oxford in the business of estimating its value increase, we do not imagine that $2+2=4$, or that either Melbourne is in Australia or not, just because these are necessary truths. We will not imagine that Melbourne is in Australia at all, though that is a true belief we hold, and perfectly compatible with the supposition. Just as rational thinkers with finite resources should not 'clutter their minds' with pointless albeit valid inferences (as argued by Harman (1986) and Cherniak (1986), and others), so they should not clutter it with pointless albeit cotenable propositions.

This suggest a fixing for Lewis' Thesis: we accept a counterfactual $\varphi > \psi$ to the extent that (1) $\pi^{\varphi}(\psi)$ is high as per (LT), *provided* that (2) the antecedent φ is relevant for the issue addressed via ψ. We propose to capture (2) by using recent ideas on *topics* or *subject matters*: items suitable for the purpose of capturing relevance or topicality in discourse and thought.[4]

V. ABOUTNESS

Aboutness is 'the relation that meaningful items bear to whatever it is that they are *on* or *of* or that they *address* or *concern*' (Yablo 2014, p. 1): this is their *subject matter* or, as we shall also call it, their *topic.*[5] Work on topics has been burgeoning among philosophers (Lewis 1988a; Gemes 1994; Hawke 2016, 2018; Plebani 2020; Plebani and Spolaore 2021), linguists (Roberts 2011; Moltmann 2018), logicians (Fine 1986; Humberstone 2008; Fine 2017). We use declarative sentences to say true things about all kinds of conversational topics. One says: 'Midori is a professor'. One thereby addresses the topic of *Midori's profession*, *what Midori does* or, more generally, *Midori.* What one says is true just in case Midori's profession is or includes being a professor. One addresses certain topics and says that things are such-and-so with respect to them.

[4] Compare the story told in (Berto and Özgün 2021), where a probabilistic logic for the acceptability of indicatives is introduced, mirroring the one presented below for the acceptability of counterfactuals. In that paper, we use *conditional probability functions* (aka Popper functions) whereas in the present one we resort to imaging. And that paper proposes to fix Adams' Thesis via a relevance constraint essentially like the one proposed below to fix Lewis' Thesis.

[5] The quick introduction to aboutness and topics in this section piggy-backs on the one proposed in (Berto 2022). We used the same story in our aforementioned paper on indicatives (Berto and Özgün 2021).

Topics are often linked to questions or issues under discussion (Lewis 1988a; Roberts 2012): ‘Our topic is whether Oxford is too expensive for its lecturers’ maps to ‘Is Oxford too expensive for its lecturers?’. Thus Lewis (1988a, 1988b) took topics as partitions of modal space. We talk about *the number of stars*: there comes the partition determined by the question, ‘How many stars are there?’. It puts worlds in the same cell when they agree on the answer: all zero-star worlds in one cell, all one-star worlds in another, etc.

However, any old sort of thing can also serve as a conversational topic: ‘The topic of this module is deep neural networks’; ‘Our topic today is Rishi, not his wife’; ‘Let us talk about deportations to Rwanda’. Thus some approaches to subject matters are more object- or state-of-affairs- oriented (Hawke 2018, provides an excellent overview). Prominent ones take topics as sets or fusions of a sentence’s (exact) truthmakers / falsemakers (Fine 2017; Fine and Jago 2019), understood in their turn as close to states or situations in the style of Barwise and Perry (1983).

We do not need to take a stance on the nature of topics. We just need them to obey three structural constraints, on which there is some agreement in the literature:

(1) Topics tend to come with hyperintensional accounts of what sentences say, because co-intensional sentences can be about different things: only one of ‘2 + 2 = 4’ and ‘Equilateral triangles are equiangular’ is about equilateral triangles, and made true by what these are like.

(2) The space of topics has a mereological structure (Yablo 2014; Fine 2017): topics can have proper parts; distinct topics may have common parts; etc. *Mathematics* includes *arithmetic*. *Mathematics* and *philosophy* overlap, having (certain parts of) *logic* as a common part.

(3) The Boolean operators add no subject matter of their own: they are ‘topic-transparent’ (Hawke 2018; Fine 2020). The topic of $\neg\varphi$ is the same as that of φ. (‘Midori is not a professor’ is exactly about what ‘Midori is a professor’ is about: say, *Midori’s profession*, or *what Midori does*, or simply *Midori*. It is not about *not*.) Conjunction and disjunction merge topics: ‘Simon is rich or beautiful’, ‘Simon is rich and beautiful’ are both about, say, *Simon’s wealth and looks*.

In the next section we introduce a formal language including a counterfactual conditional, for which we give acceptability conditions in terms of imaging and topics. We then provide a probabilistic logic to reason about the (un)acceptability of counterfactuals.

VI. ON-TOPIC COUNTERFACTUALS

Let $\mathcal{L}_{PL}$ be the language of classical propositional logic on a countable set of propositional variables $\mathsf{Prop} = \{p, q, \dots\}$ with connectives $\neg$ and $\wedge$. The well-formed formulas are the elements of Prop, $\neg\varphi$, and $(\varphi \wedge \psi)$ whenever φ and ψ are formulas. We identify $\mathcal{L}_{PL}$ with the set of its well-formed formulas and employ the usual abbreviations for the connectives $\vee, \supset, \equiv$ as $\varphi \vee \psi := \neg(\neg\varphi \wedge \neg\psi)$, $\varphi \supset \psi := (\neg\varphi \vee \psi)$, and $\varphi \equiv \psi := (\varphi \supset \psi) \wedge (\psi \supset \varphi)$. So, $\supset$ is the material conditional and $\equiv$ is the material biconditional. As for $\top$ and $\bot$, we set $\top := p \vee \neg p$ and $\bot := \neg\top$. We call the elements of $\mathcal{L}_{PL}$ *Boolean sentences.* For any $\varphi \in \mathcal{L}_{PL}$, P_φ denotes the set of propositional variables occurring in φ.

Let $\mathcal{L}_{PL}$ be interpreted in possible worlds semantics the usual way. Following Stalnaker (1968), we add to our models an *absurd world*, λ, in which every proposition is true. Given a tuple $\mathcal{M} = (W, \lambda, V)$, where W is a nonempty set of worlds plus λ (i.e., $\lambda \in W$), and $V : \mathsf{Prop} \to \mathcal{P}(W)$ is a valuation function such that $\lambda \in V(p)$ for all $p \in \mathsf{Prop}$ and otherwise standard, $|\varphi|^{\mathcal{M}}$ denotes the truth set of φ in $\mathcal{M}$: the set of worlds that make φ true. Note that $\lambda \in |\varphi|^{\mathcal{M}}$ for all $\varphi \in \mathcal{L}_{PL}$. When $|\varphi|^{\mathcal{M}} = \{\lambda\}$ we call φ *impossible* in $\mathcal{M}$, and *possible* otherwise. We omit the superscript and write $|\varphi|$ when the model is contextually clear. '$\models_{PL}$' stands for classical logical truth/consequence.

The language $\mathcal{L}$ of simple counterfactual conditionals extends $\mathcal{L}_{PL}$ by a counterfactual, '$>$', connecting only the elements of $\mathcal{L}_{PL}$ so as to avoid nesting: the well-formed formulas in $\mathcal{L}$ are the elements of $\mathcal{L}_{PL}$, plus $(\varphi > \psi)$ whenever φ and ψ are in $\mathcal{L}_{PL}$. Now a few definitions:

Definition 1 (Stalnakerian selection function). Given a tuple $\mathcal{M} = (W, \lambda, V)$ as described above, a *Stalnakearian selection function* $f : (W \times \mathcal{L}_{PL}) \to W$ assigns a possible world to each pair of a world in W and a sentence in $\mathcal{L}_{PL}$, and satisfies the following properties:

1. For all $\varphi \in \mathcal{L}_{PL}$ and $w \in W$, $f(w, \varphi) \in |\varphi|$,
2. For all $\varphi \in \mathcal{L}_{PL}$ and $w \in W$, $f(w, \varphi) = \lambda$ iff $|\varphi| = \{\lambda\}$,
3. For all $\varphi \in \mathcal{L}_{PL}$ and $w \in W$, if $w \in |\varphi|$ then $f(w, \varphi) = w$.
4. For all $\varphi, \psi \in \mathcal{L}_{PL}$ and $w \in W$, if $f(w, \varphi) \in |\psi|$ and $f(w, \psi) \in |\varphi|$ then $f(w, \varphi) = f(w, \psi)$.

We call the tuple (W, f, λ, V) a *Stalnaker model.* The above conditions on selection functions are nothing new with respect to Stalnaker (1968): (1) has it that the φ-selected world is a φ-world. This makes sense for supposition in particular: when we suppose that φ, we only look at a world where φ to begin with (when we suppose that the Banbury house is in Oxford, we look at a world where it *is* in Oxford). In (2) we follow Stalnaker in having a λ, the absurd world, where everything is just stipulated to be true. This is the value of f iff at any world it takes as input that is not true in any *possible* world, that is, an input with the impossible truth set $\{\lambda\}$. (3) says that w is the single world closest to itself whenever it is a φ-world already. (4) is needed to make sure f selects based on closeness as comparative similarity.

Definition 2 (Topic models with operators). A *topic model with operators (*in short, *topic model)* $\mathcal{T}$ is a tuple $\langle T, \oplus, t, k\rangle$ where

1. T is a non-empty set of *possible topics.* We use variables a, b, c $(a_1, a_2, \dots)$ ranging over possible topics.
2. $\oplus : T \times T \to T$ is a binary idempotent, commutative, associative operation: *topic fusion*, making of topics part of larger topics. We assume unrestricted fusion, that is, $\oplus$ is always defined on T: $\forall a, b \in T\ \exists c \in T(c = a \oplus b)$.
3. t : $\mathsf{Prop} \to T$ is a *topic function* assigning a topic to each element in Prop. t extends to $\mathcal{L}_{PL}$ by taking the topic of a sentence φ as the fusion of the elements in $\mathsf{P}_\varphi = \{p_1, \dots, p_k\}$, i.e., the atoms showing up in it:
$$t(\varphi) = \oplus \mathsf{P}_\varphi = t(p_1) \oplus \cdots \oplus t(p_k).$$
We abbreviate $t(\varphi)$ as t_φ.
4. $k : T \to T$ is a function on T that satisfies for all $a, b \in T$:
   (a) $a \sqsubseteq k(a)$ (Inclusion);
   (b) $k(a) = k(k(a))$ (Idempotence);
   (c) $k(a \oplus b) = k(a) \oplus k(b)$ (Additivity).

Out of fusion as per (2), we can define *topic parthood*, $\sqsubseteq$, what it means that a topic is included in another, standardly:

$$\forall a, b(a \sqsubseteq b \text{ iff } a \oplus b = b).$$

This makes $\sqsubseteq$ a partial order on T. That the topic of a φ is the fusion of those of its atoms, as per (3), secures topic-transparency for connectives.

(4) has k as a *Kuratowski closure operator* on the poset $\langle \mathcal{T}, \sqsubseteq\rangle$. It will come handy when we give acceptability conditions. Inclusion (4a) guarantees that the closure $k(a)$ of a given topic a will always be an expansion: it will enlarge the original topic, but never take us far away from it; k expands a topic a in a minimal way. Idempotence (4b) says that one cannot repeat the expansion unless the topic changes. Additivity (4c) guarantees that closure on a whole never outstrips closure on its parts. Thus in particular closing t_φ is the same as closing the topics of its atoms, then fusing them.

Definition 3 (Stalnakerian Discrete Probability Space). A *Stalnakerian discrete probability space* is a tuple (W, λ, π) such that W is a nonempty and finite set of possible worlds and λ (i.e., $\lambda \in W$), the *sample space*, and $\pi : W \to [0, 1]$ is a probability mass function such that either $\pi(\lambda) = 1$ or $\sum_{w \in W \setminus \{\lambda\}} \pi(w) = 1$.

We call the tuple (W, λ, π) *Stalnakerian* to emphasise the occurrence of the absurd world λ. It will be useful in formalising acceptability conditions of counterfactuals with impossible antecedents. Throughout the paper we employ only Stalnakerian discrete probability spaces, so we'll simply call them *probability spaces.* All our probability mass functions are to satisfy the conditions given in Definition 3.

Definition 4 (Probabilistic Model). A *probabilistic model* is a tuple $\mathcal{N} = (W, f, \lambda, V, \pi)$ where (W, f, λ, V) is a Stalnakerian model and $\pi : W \to [0, 1]$ is a probability mass function as described in Definition 3. Then, for all $\varphi \in \mathcal{L}_{PL}$, we have $\pi(\varphi) = \sum_{w \in |\varphi|} \pi(w)$.

Definition 5 (Imaging). Given a probabilistic model (W, f, λ, V, π) and $\varphi \in \mathcal{L}_{PL}$, the *image* π^{φ} of π under φ is defined as: for all $w' \in W$, $\pi^{\varphi}(w') := \sum_{w : f(w,\varphi) = w'} \pi(w)$.

Easily, π^{φ} is a probability mass function over W when π is (see Lemma 3.4 for the proof). Given π^{φ} and $\psi \in \mathcal{L}_{PL}$, $\pi^{\varphi}(\psi) = \sum_{w \in |\psi|} \pi^{\varphi}(w)$.

Now comes the heart of the story:

Definition 6 (Degrees of (Un)Acceptability). For any probabilistic model $\mathcal{N} = (W, f, \lambda, V, \pi)$ and topic model $\mathcal{T} = \langle T, \oplus, t, k \rangle$ defined on $\mathcal{L}_{PL}$, the degree of acceptability $\mathcal{A}_{\mathcal{N},\mathcal{T}} : \mathcal{L} \to [0, 1]$ of an element in $\mathcal{L}$ is defined as:

1. for all $\varphi \in \mathcal{L}_{PL}$, $\mathcal{A}_{\mathcal{N},\mathcal{T}}(\varphi) = \pi(\varphi)$; and
2. $\mathcal{A}_{\mathcal{N},\mathcal{T}}(\varphi > \psi) = \begin{cases} \pi^{\varphi}(\psi), & \text{if } t_{\psi} \sqsubseteq k(t_{\varphi}) \\ 0 & \text{otherwise.} \end{cases}$

For any $\varphi \in \mathcal{L}$, the degree of unacceptability $\mathcal{U}_{\mathcal{N},\mathcal{T}}(\varphi)$ is then given by $\mathcal{U}_{\mathcal{N},\mathcal{T}}(\varphi) = 1 - \mathcal{A}_{\mathcal{N},\mathcal{T}}(\varphi)$.

(When it is clear which probability and topic model are used, we omit the subscripts and simply write $\mathcal{A}$ and $\mathcal{U}$.)[6]

[6] Easily, given a probabilistic model $\mathcal{N} = (W, f, \lambda, V, \pi)$ and a topic model $\mathcal{T} = \langle T, \oplus, t, k \rangle$, we have:

1. for all $\varphi \in \mathcal{L}_{PL}$, $\mathcal{U}(\varphi) = 1 - \pi(\varphi)$; and
2. $\mathcal{U}(\varphi > \psi) = \begin{cases} 1 - \pi^{\varphi}(\psi), & \text{if } t_{\psi} \sqsubseteq k(t_{\varphi}) \\ 1 & \text{otherwise.} \end{cases}$

(1) just says that the degree of acceptability of a Boolean sentence φ is its plain probability. (2) is the core of our proposal. It says that the degree of acceptability of a simple counterfactual $\varphi > \psi$ is given by (i) the probability of its consequent under the counterfactual supposition, i.e., the image, of the antecedent, so long as (ii) it is an *on-topic counterfactual* (with respect to $\mathcal{T}$): the topic of its consequent, t_ψ, is included in the closure of the topic of the antecedent $k(t_\varphi)$. Otherwise, $\varphi > \psi$ is plainly unacceptable. Let us unpack.

First, the plain unacceptability of off-topic counterfactuals can be taken as an idealisation of the proposed formalism: being off-topic may sometimes actually lower the acceptability of a counterfactual without necessarily taking it down to 0, especially if one endorses a graded notion of topic relevance. A treatment of the latter is beyond the scope of this (already long) initial paper. We anticipate that allowing for low-but-non-zero acceptability of off-topic conditionals should not affect the *acceptability-preserving* (in-)validities of the corresponding logic, and the intuitions they reflect: acceptability should in any case not be preserved from on-topic conditionals to off-topic ones. So our plain unacceptability for the off-topics does not bear on our substantive philosophical claims and the principles of interest in a logic of on-topic counterfactuals. (We thank an anonymous reviewer for pressing us on this issue.)

Second: why did we not require plain inclusion of the topic of ψ in that of φ, $t_\psi \sqsubseteq t_\varphi$? Because we often accept a counterfactual in contexts where, on an intuitive way of understanding topicality, there is no plain topic-inclusion between what the consequent and what the antecedent are about:

11. If we had stopped burning fossil fuels twenty years ago, the polar ice would not have been melting so quickly.
12. If Brexit was *that* bad, the Tories would have lost the majority in parliament by now.
13. If you had pushed that button, the plane would not have stalled.

In cases like (11)-(13), the antecedent φ is relevant for the consequent although it does not, on its own, address an issue with respect to which the consequent is fully on-topic. Rather, the counterfactual supposition of the antecedent is carried out in a context where one is tackling a question or issue, triggering a bunch of background assumptions, say BA_φ, with respect to which the consequent is fully on topic. Surely these assumptions are connected with the antecedent (that is what the subscript in 'BA_φ' is there to remind you of; in particular, plausibly,

$\varphi \in BA_\varphi$). E.g., in (11), the issue of polar ice melting can make contextually relevant topics connected to fossil fuel burning, such as the emission of CO_2, raising global temperatures, etc. In (12), the issue of Tory electoral success can make contextually relevant topics connected to the badness of Brexit, such as electoral reactions to socioeconomic decline.

Here the Kuratowski operator earns its keep. For suppose (with a small abuse of notation) $k(t_\varphi) = t(BA_\varphi)$, that is, we think of k exactly as mapping the topic of the antecedent φ to that of the relevant background assumptions BA_φ contextually determined by φ and the tackled issue. This makes precise what it means that suppositional thinking has a *focus*. When we suppose that φ, wondering about the issue whether ψ, we can move, in a way dictated by context/tackled issues, beyond the topic of the suppositional input φ. But our expansion will be *regimented*: it will expand to distinct, but *connected* topics. The Kuratowski is a topological closure operator, giving to *connectedness* a precise topological meaning: Inclusion guarantees that $t_\varphi \sqsubseteq t(BA_\varphi)$, i.e., the topic of the relevant background assumptions BA_φ possibly expands, but always includes that of the antecedent φ. Idempotence has it that $t(BA_\varphi)$ is complete: contemplating on the background assumptions does not lead to new topics unless given additional inputs. Additivity ensures that $t(BA_\varphi)$ is the same as the fusion of the topics determined by its simpler components.

Third: connectedness is, admittedly, dealt with in a rather abstract fashion. If one asked, 'But exactly *which* are the connected topics?', one would not find very informative replies in the formal setting itself. This is, we submit, unavoidably so. What the relevant background assumptions in BA_φ are, is a volatile, fuzzy, focus-dependent matter. In some cases, the connection between antecedent and consequent will be so obvious that little or no context or focus is required to acknowledge it. Sometimes, only a lengthy story will tell whether the topicality constraint is satisfied or not.[7]

[7] A nice example provided by a helpful anonymous referee: 'If you spent more time reading Hegel, your cholesterol level would not improve'. Where is the topicality connection? Well, say the context fixing the relevant BA_φ is one where the conversational focus is on my sedentary attitude as a reader of philosophy, and how it affects my psychophysical health. I claim: 'If I spent more time reading Hegel, I would feel more peaceful at the end of the day.' You retort: 'If you spent more time reading Hegel, your cholesterol would not improve'. This seems perfectly on-topic as stressing that what I need is physical exercise, rather than more Hegel and peace of mind. But of course, one can come up with several different contexts, where 'If you spent more time reading Hegel, your cholesterol would not improve' just turns out to be an irrelevant conditional – e.g., pick one where we are addressing the topic of my competence in classical German philosophy, and my psychophysical health is not at issue at all.

For a hopefully helpful analogy: any modal account of the truth conditions of counterfactuals, of a broadly Kratzerian (see Kratzer 2012) or Lewisian-Stalnakerian kind, will involve some apparatus for focusing on the contextually relevant worlds for the interpretation of the counterfactual at hand. Take the Lewisian-Stalnakerian story: '$\varphi > \psi$' is true at w iff ψ is true at the φ-world(s) closest to w. Intuitively, the closest worlds are those where the background assumptions in BA_φ hold (Priest 2001, ch. 5). If one asked, 'But exactly *which* are the closest worlds?', one would not find the Lewis-Stalnaker semantics very informative. And of course, the question *has* been asked – starting with Fine (1973)'s critical notice of Lewis' *Counterfactuals* book, which has generated a large literature on how to specify the relevant similarity respects (see Bennett (2003) for a masterful reconstruction). Now most of such literature is about informally *glossing* on the formal semantics from the outside. That is because, as Lewis puts it in *Counterfactuals*: 'The truth conditions for counterfactuals [...] are a highly volatile matter, varying with every shift of context and interest' (Lewis 1973, 92). The Lewis-Stalnaker semantics *presupposes* that we have some (vague, fuzzy, context-dependent) intuitions on what counts as the most similar situations, and piggy-backs on that, giving us a precise but merely formal account via a total ordering of worlds by closeness. We claim that the same holds for the topics of counterfactuals. Our formal setting *presupposes* that we have some (vague, fuzzy, context-dependent) intuitions on what counts as topic-connectedness, and piggy-backs on that, giving us a precise but merely formal account via a topological closure operation.

Fourth: topics make acceptability hyperintensional. To see that, look back at (9) and (10) above. In both cases, we assign probability zero to the antecedent and to the consequent. But there *is* a world where they are all true, namely λ. That is where we look when we suppose counterfactually that there is a largest prime. The expansion of the discrete probability space so that it includes λ (Definition 3), and of the definition of imaging so that it comes out well-defined for impossible counterfactual antecedents, have it that the probability of all no-largest-prime worlds, which is all possible worlds in W, is shifted to λ in our counterfactual supposition. The probability of both consequents under the image of such an antecedent is 1. But, topicality tells them apart: by our acceptability conditions, (9) is fully acceptable while (10) is not, insofar as we assign to the consequent of the former, not of the latter, a topic which is included in the closure of the antecedent's topic.

VII. THE LOGIC OF ON-TOPIC COUNTERFACTUALS

We take the closure principles of our logic as premise-conclusion rules of the form '$\Gamma \vdash \Delta$' where $\Gamma, \Delta \subseteq \mathcal{L}$ with $\Gamma = \emptyset$ for zero-premise rules.

Following Adams (1998), we define validity probabilistically in terms of degrees of unacceptability:

Definition 7 (Validity). A principle of the form $\Gamma \vdash \Delta$ is *valid* if and only if for any probabilistic model $\mathcal{N} = (W, f, \lambda, V, \pi)$ and topic model $\mathcal{T} = \langle T, \oplus, t, k\rangle$,

$$\sum_{\varphi \in \Gamma} \mathcal{U}(\varphi) \geq \mathcal{U}(\psi),$$

for all $\psi \in \Delta$. When $\Gamma = \emptyset$, we say $\vdash \Delta$ is *valid* if and only if $\mathcal{U}(\psi) = 0$ for all $\psi \in \Delta$. $\Gamma \vdash \Delta$ is *invalid* otherwise. [8]

Our notion of validity depends on both probability and topicality. Besides investigating valid closure principles, we thus want to check that the invalid ones fail for the right *reason*. So we also consider *probabilistic (in)validity* and *topical (in)validity* as distinct sources of invalidity.

We say that $\Gamma \vdash \Delta$ is *probabilistically valid* (*p*-valid) iff for any probabilistic model (W, f, λ, V, π) and *singleton* topic model $\mathcal{T}$ (i.e., when T is a singleton), $\sum_{\varphi \in \Gamma} \mathcal{U}(\varphi) \geq \mathcal{U}(\psi)$, for all $\psi \in \Delta$. When $\Gamma = \emptyset$, we say $\vdash \Delta$ is *p-valid* if and only if $\mathcal{U}(\psi) = 0$ for all $\psi \in \Delta$; $\Gamma \vdash \Delta$ is *p-invalid* otherwise: *p*-validity ignores topicality by focusing on trivial singletons, and just checks how a putative closure principle fares probabilistically.

We say $\Gamma \vdash \Delta$ is *topically valid* (*t*-valid) iff for any topic model $\mathcal{T} = \langle T, \oplus, t, k\rangle$, if every conditional in Γ is an on-topic conditional wrt $\mathcal{T}$ then every conditional in Δ is also an on-topic conditional wrt $\mathcal{T}$; $\Gamma \vdash \Delta$ is *t-invalid* otherwise: *t*-validity ignores probabilities and just checks how a putative closure principle fares topically.

We now focus on the principles in Table 1. We label them sticking to popular names or acronyms from the literature on conditional logics.

Lemma 1. If $\Gamma \vdash \Delta$ is valid then it is *p*-valid but not necessarily *t*-valid. If $\Gamma \vdash \Delta$ is both *p*- and *t*-valid, then it is valid.

Proof. See Appendix A.[9] □

[8] We formulate the relevant closure principles as premise-conclusion rules, plainly following (Douven 2016, Chapter 5). In this logical framework, a premise-conclusion rule is interpreted as: 'Whenever all elements of Γ are acceptable, every element of Δ is also acceptable'. Our notion of validity is probabilistic, preserving degrees of (un)acceptability (as opposed to preserving truth, as validity in standard non-probabilistic logic does); and our language is restricted to simple counterfactuals. So we cannot restate our consequence relation between Γ and Δ by simply replacing (a finite) Δ with the conjunction of its elements. And our way of formulating the principles of interest is more economical: instead of stating a closure principle for every conclusion we can derive from a set Γ of premises, we package all its conclusions of interest in a set Δ of conclusions. See, e.g., (p. 128-130), for further elaboration on how the principles in Table 1 should be interpreted.

[9] Aside from details of wording, the *t*-(in)validity proofs in the appendix are essentially the same as the ones supplied in our paper (Berto and Özgün 2021) on indicatives. Moreover, when one leaves nested conditionals aside, as we do in both papers, the logics are perfectly aligned.

(REF)	$\vdash \varphi > \varphi$
(ANT)	$\varphi > \psi \vdash \varphi > (\varphi \wedge \psi)$
(CM)	$\varphi > (\psi \wedge \chi) \vdash \varphi > \psi, \varphi > \chi$
(CC)	$\varphi > \psi, \varphi > \chi \vdash \varphi > (\psi \wedge \chi)$
(CSO)	$\varphi > \psi, \psi > \varphi, \varphi > \chi \vdash \psi > \chi$
(CT)	$\varphi > \psi, (\varphi \wedge \psi) > \chi \vdash \varphi > \chi$
(CMon)	$\varphi > \psi, \varphi > \chi \vdash (\varphi \wedge \psi) > \chi$
(OR)	$\varphi > \psi, \chi > \psi \vdash (\varphi \vee \chi) > \psi$
(M. Ponens)	$\varphi, \varphi > \psi \vdash \psi$
(Trans)	$\varphi > \psi, \psi > \chi \vdash \varphi > \chi$
(SA)	$\varphi > \psi \vdash (\varphi \wedge \chi) > \psi$
(MOD)	$\neg\varphi > \varphi \vdash \psi > \varphi$
(RCE)	If $\varphi \vdash_{PL} \psi$, then $\vdash \varphi > \psi$
(RCEA)	If $\vdash_{PL} \varphi \equiv \psi$, then $\varphi > \chi \dashv\vdash \psi > \chi$
(RCEC)	If $\vdash_{PL} \varphi \equiv \psi$, then $\chi > \varphi \dashv\vdash \chi > \psi$
(RCK)	If $\vdash_{PL} (\varphi_1 \wedge \cdots \wedge \varphi_n) \supset \psi$, then $\chi > \varphi_1, \ldots, \chi > \varphi_n \vdash \chi > \psi$
(RCM)	$\vdash_{PL} \varphi \supset \psi$, then $\chi > \varphi \vdash \chi > \psi$
(And-to-If)	$\varphi \wedge \psi \vdash \varphi > \psi$
(Or-to-If)	$\varphi \vee \psi \vdash \neg\varphi > \psi$
(Contr.)	$\varphi > \neg\psi \vdash \psi > \neg\varphi$
(SDA)	$(\varphi \vee \psi) > \chi \vdash \varphi > \chi, \psi > \chi$

Table 1. Closure principles of interest

Theorem 2.

1. REF, ANT, CM, CC, CSO, CT, CMon, OR, and Modus Ponens are both *p*- and *t*-valid. Therefore, they all are valid.
2. MOD, RCE, RCEA, RCEC, RCK, RCM, and And-to-If are *p*-valid but *t*-invalid.
3. Trans and SA are *p*-invalid but *t*-valid.
4. Or-to-if, Contraposition, and SDA are both *p*-invalid and *t*-invalid.
5. MOD, RCE, RCEA, RCEC, RCK, RCM, And-to-If, Trans, SA, Or-to-If, Contraposition, and SDA are invalid.

Proof. Supplied in Appendix B. □

We comment on some validities and invalidities. Looking at the former, in group 1: REF (Reflexivity), ANT, CM, and Modus Ponens appear obviously desirable. Segerberg (1989) claimed that CC (Conjunction in the Consequent) should hold in any reasonable conditional logic.[10]

[10] One may take issue, due to Lottery Paradox cases (Kyburg 1961) if one has a qualitative idea of acceptability, whereby something becomes (plainly) acceptable by passing an intermediate probabilistic threshold θ. Ours, however, is a quantitative setting with

	valid	p-valid	t-valid
REF	✓	✓	✓
ANT	✓	✓	✓
CM	✓	✓	✓
CC	✓	✓	✓
CSO	✓	✓	✓
CT	✓	✓	✓
CMon	✓	✓	✓
OR	✓	✓	✓
Modus Ponens	✓	✓	✓
MOD	X	✓	X
RCE	X	✓	X
RCEA	X	✓	X
RCEC	X	✓	X
RCK	X	✓	X
RCM	X	✓	X
And-to-If	X	✓	X
Trans	X	X	✓
SA	X	X	✓
Or-to-If	X	X	X
Contraposition	X	X	X
SDA	X	X	X

Table 2. Validities (✓) and invalidities (X): summary of the results in Theorem 2.

CT (Cautious Transitivity), CMon (Cautious Monotonicity) and OR hold in most conditional logics and theories of non-monotonic entailment (Nute 1984). In particular, CT and CMon feature in Chellas (1975)'s basic conditional logic and are put by Gabbay (1985) among the minimal requirements for a logic of non-monotonic entailment. They hold in the system C of Kraus, Lehmann, and Magidor (1990); their popular nonmonotonic logic P has them, too.

The invalidities in group 2 are all probabilistically kosher; their failure is due to topicality. E.g., look at MOD: we can accept that a necessary truth is counterfactually implied by its own negation (we could even *define* necessity thus, following (Lewis 1973, 22): $\Box\varphi := \neg\varphi > \varphi$); but we do not accept the counterfactual implication from whatnot to such a truth ('If the moon was made of green cheese, then there would be no largest prime').

Look at RCE: that φ logically entails ψ is insufficient for the acceptability of the corresponding counterfactual: 'This Banbury house is in Oxford' classically entails 'Either Melbourne is in Australia, or not',

degrees of acceptability. One can rephrase it qualitatively, if one likes: take $\varphi > \psi$ as acceptable when (i) $\pi^{\varphi}(\psi) \geq \theta$ and (ii) $t_{\psi} \sqsubseteq k(t_{\varphi})$; re-define validity accordingly, as threshold-preservation: CC becomes invalid for threshold values $\theta \in (\frac{1}{2}, 1)$.

but we do not accept the patently off-topic ‘If this Banbury house were in Oxford, then either Melbourne would be in Australia or not’. Similarly for RCEA and RCEC: that φ and ψ are logical equivalents (i.e., such that their material equivalence is a theorem) does not guarantee that replacing either with the other as a counterfactual antecedent or consequent preserves acceptability.

We left CSO behind when talking of validities. Now it is time to mention it: for it limits the hyperintensional anarchy of acceptability in our topic-sensitive setting. Replacement of logical equivalents can fail to preserve acceptability due to topicality constraints. However, CSO tells us that replacement of *counterfactual* equivalents works just fine.

Groups 3 and 4 include inferences generally agreed to be invalid for any *ceteris paribus* conditional, whether in the indicative or in the counterfactual mood: Trans(itivity), SA (Strengthening the Antecedent), Or-to-If, Contraposition, SDA (Simplification of Disjunctive Antecedents), all fail both in the Adams (1998) probabilistic logic for indicatives, and in the standard possible worlds semantics for counterfactuals by Stalnaker (1968) and Lewis (1973).

And-to-If, the inference from a conjunction to the corresponding counterfactual, deserves more attention. It is often called ‘Centering’ for in the Lewis semantics it is valid when one assumes that the world of evaluation is always the single world most similar to itself (in the ‘spheres’ setting of Lewis (1973): it is the unique world at the centre of the nested spheres of worlds arranged around it). Discussing And-to-If on pp. 26-29 of *Counterfactuals*, Lewis finds Weak Centering, i.e., the assumption that nothing is more similar to a world w than w itself, ‘perfectly safe’ (as it is required to validate Modus Ponens), but Centering ‘not quite such a safe assumption’ (29). His argument for And-to-If from the previous page is to the effect that it is necessarily *truth*-preserving, while he grants the oddness of asserting a counterfactual just because antecedent and consequent are both true:

> [The argument] is evidence for my truth conditions. What can be said against them? So far as I know, only this: it would seem very odd to pick two completely unrelated truths φ and ψ and, on the strength of their truth, to deny the counterfactual $\varphi > \neg\psi$; and even odder, to assert the counterfactual $\varphi > \psi$. What would we make of someone who saw it fit to deny that if the sky were blue then grass would not be green, or to assert that if the sky were blue then grass would be green? It would be doubly odd. First, because he is using the counterfactual construction with an antecedent he takes to be true, though this construction is customarily reserved for antecedents taken to be false; second, because his assertions could serve no likely conversational purpose that would not be better served by separate assertions of φ and ψ. But oddity is not falsity; not everything true is a good thing to say. (28; notation adjusted)

Insofar as *acceptability* is concerned, And-to-If should go: it may be truth-preserving, but it is not acceptability-preserving due to considerations of relevance. So it fails in the right way in our logic: it is probabilistically kosher; but while for a conjunction $\varphi \wedge \psi$ to be found acceptable nothing more may generally be required than believing the truth of both conjuncts, these may be completely unrelated claims; and this makes the corresponding counterfactual off-topic, albeit *p*-valid.

Off-topic counterfactuals with true antecedent and consequent are *triply* odd. Besides the twofold pragmatic oddness remarked by Lewis, there is additional oddness, independent from communication, contemplated in the privacy of our heads. Not only do we not want to assert 'If the sky were blue, then grass would be green': we find it unacceptable, for we judge the colour of the sky irrelevant for that of grass. Seeing relevant counterfactual connections where there are none would be superstition.

VIII. FURTHER WORK

Expanding the logic to nested counterfactuals would be interesting but intuitions of acceptability might be all over the place. Some experimental work would then be dearly needed. Next, the original Lewisian definition of imaging relied on Stalnaker's Assumption: for each w and φ there is a unique closest antecedent-world w_{φ}. This has been famously criticized by Lewis (1973) and others. (It has been defended: see e.g. Williams (2010).) After taking what we have called Lewis' Proof (LP), $\pi(\varphi > \psi) = \pi^{\varphi}(\psi)$, as a good candidate for giving the probabilities of counterfactuals, Schulz (2017) claims that 'the only problem with imaging is that it presupposes uniqueness' (81). One can try and generalize Stalnakerian imaging to a setting with set-selection functions outputting, for each world and available antecedent, a set of maximally close antecedent-worlds which may feature more than one world. Then one variously distributes probabilities among the worlds in such a set (Gärdenfors 1982; Joyce 1999; Leitgeb 2017).

But while (LP) protects our setting with Stalnakerian imaging from triviality, the situation when one generalizes is more cumbersome. Chapter 8 of Schulz (2017) has an elaborate discussion, where he tries to show how to protect a generalized set-up he has proposed in previous chapters from the triviality results of Williams (2012) and Leitgeb (2012). What would settle the issue is a mathematical result that, for all we know, is still wanting.

REFERENCES

Adams, Ernest. 1966. “Probability and the Logic of Conditionals.” In *Aspects of Inductive Logic,* edited by Jaakko Hintikka and Patrick Suppes, 43:265–316. Studies in Logic and the Foundations of Mathematics. Elsevier. https://doi.org/https://doi.org/10.1016/S0049-237X(08)71673-2.

———. 1975. *The Logic of Conditionals.* Dordrecht: Reidel. https://doi.org/https://doi.org/10.1007/978-94-015-7622-2.

———. 1998. *A Primer of Probability Logic.* Stanford: CSLI Publications.

Barwise, J., and J. Perry. 1983. *Situations and Attitudes.* Stanford: CSLI Publications.

Bennett, Jonathan. 2003. *A Philosophical Guide to Conditionals.* Oxford: Oxford University Press. https://doi.org/https://doi.org/10.1093/0199258872.001.0001.

Berto, Francesco. 2022. *Topics of Thought. The Logic of Knowledge, Belief, Imagination.* Oxford: Oxford University Press.

Berto, Franz, and Aybüke Özgün. 2021. “Indicative Conditionals: Probabilities and Relevance.” *Philosophical Studies* 178:3697–3730. https://doi.org/10.1007/s11098-021-01622-3.

Chellas, Brian F. 1975. “Basic Conditional Logic.” *Journal of Philosophical Logic* 4:133–153. https://doi.org/https://doi.org/10.1007/BF00693270.

Cherniak, Christopher. 1986. *Minimal Rationality.* Bradford Books. Cambridge, MA: MIT Press.

Douven, Igor. 2016. *The Epistemology of Indicative Conditionals: Formal and Empirical Approaches.* Cambridge: Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9781316275962.

Douven, Igor, and Sara Verbrugge. 2010. “The Adams Family.” *Cognition* 117:302–318. https://doi.org/https://doi.org/10.1016/j.cognition.2010.08.015.

Edgington, Dorothy. 1995. “On Conditionals.” *Mind* 104:235–329. https://doi.org/https://doi.org/10.1093/mind/104.414.235.

Evans, Jonathan St. B. T., Simon J. Handley, and David E. Over. 2003. “Conditionals and Conditional Probability.” *Journal of experimental psychology. Learning, memory, and cognition* 29:321–335. https://doi.org/https://doi.org/10.1037/0278-7393.29.2.321.

Evans, Jonathan St. B. T., and David E. Over. 2004. *If.* Oxford: Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780198525134.001.0001.

Fine, Kit. 1973. "Critical Notice of *Counterfactuals.*" *Mind* 84:451–458. https://doi.org/https://doi.org/10.1093/mind/LXXXIV.1.451.

———. 1986. "Analytic Implication." *Notre Dame Journal of Formal Logic* 27 (2): 169–179. https://doi.org/10.1305/ndjfl/1093636609.

———. 2017. "A Theory of Truthmaker Content I: Conjunction, Disjunction and Negation." *Journal of Philosophical Logic* 46 (6): 625–674. https://doi.org/10.1007/s10992-016-9413-y.

———. 2020. "Yablo on Subject-Matter." *Philosophical Studies* 177 (1): 129–171. https://doi.org/10.1007/s11098-018-1183-7.

Fine, Kit, and Mark Jago. 2019. "Logic for Exact Entailment." *The Review of Symbolic Logic* 12 (3): 536–gabb556. https://doi.org/10.1017/S1755020318000151.

Gabbay, Dov. 1985. "Theoretical foundations for non-monotonic reasoning in expert systems." In *Logics and Models of Concurrent Systems,* 439?457. Berlin, Heidelberg: Springer-Verlag.

Gärdenfors, Peter. 1982. "Imaging and Conditionalization." *Journal of Philosophy* 79 (12): 747–760. https://doi.org/10.2307/2026039.

Gemes, Ken. 1994. "A New Theory of Content I: Basic Content." *Journal of Philosophical Logic* 23 (6): 595–620. https://doi.org/10.1007/bf01052779.

Hájek, Alan. 1989. "Probabilities of Conditionals – Revisited." *Journal of Philosophical Logic* 18 (4): 423–428. https://doi.org/10.1007/bf00262944.

Harman, Gilbert. 1986. *Change In View.* Cambridge, MA: MIT Press. https://doi.org/https://doi.org/10.1016/0004-3702(87)90007-5.

Hawke, Peter. 2016. "Questions, Topics and Restricted Closure." *Philosophical Studies* 173 (10): 2759–2784. https://doi.org/10.1007/s11098-016-0632-4.

———. 2018. "Theories of Aboutness." *Australasian Journal of Philosophy* 96 (4): 697–723. https://doi.org/10.1080/00048402.2017.1388826.

Humberstone, Lloyd. 2008. "Parts and Partitions." *Theoria* 66:41–82. https://doi.org/10.1111/j.1755-2567.2000.tb01144.x.

Iatridou, Sabine. 1993. “On the Contribution of Conditional ‘Then’.” *Natural Language Semantics* 2:171–99. https://doi.org/10.1007/bf01256742.

Jackson, Frank. 1987. *Conditionals.* Oxford: Blackwell.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory.* Cambridge: Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511498497.

Kratzer, Angelika. 2012. *Modals and Conditionals.* Oxford: Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780199234684.001.0001.

Kraus, Sarit, Daniel Lehmann, and Menachem Magidor. 1990. “Nonmonotonic Reasoning, Preferential Models and Cumulative Logics.” *Artificial Intelligence* 44:167–207. https://doi.org/https://doi.org/10.1016/0004-3702(90)90101-5.

Krzyżanowska, Karolina. 2015. “Between If and Then.” PhD diss., University of Groningen.

Krzyżanowska, KArolina, Peter J. Collins, and Ulrike Hahn. 2017. “Between a Conditional’s Antecedent and Its Consequent: Discourse Coherence vs. Probabilistic Relevance.” *Cognition* 164:199–205. https://doi.org/https://doi.org/10.1016/j.cognition.2017.03.009.

Kyburg, Henry Ely. 1961. *Probability and the Logic of Rational Belief.* Middletown CT: Wesleyan University Press.

Leitgeb, Hannes. 2012. “A Probabilistic Semantics for Counterfactuals - Part A.” *The Review of Symbolic Logic* 5:26–84. https://doi.org/10.1017/S1755020311000153.

———. 2017. “Imaging All the People.” *Episteme* 14 (4): 463–479. https://doi.org/10.1017/epi.2016.14.

Lewis, David. 1973. *Counterfactuals.* Oxford: Blackwell.

———. 1976. “Probabilities of Conditionals and Conditional Probabilities.” *Philosophical Review* 85 (3): 297–315. https://doi.org/10.2307/2184279.

———. 1988a. “Relevant Implication.” *Theoria* 54 (3): 161–174. https://doi.org/10.1111/j.1755-2567.1988.tb00716.x.

———. 1988b. “Statements Partly About Observation.” *Philosophical Papers* 17 (1): 1–31. https://doi.org/10.1080/05568648809506282.

McGee, Vann. 1989. "Conditional Probabilities and Compounds of Conditionals." *Philosophical Review* 98 (4): 485–541. https://doi.org/10.2307/2185116.

Moltmann, Friederike. 2018. "An Object-Based Truthmaker Semantics for Modals." *Philosophical Issues* 28 (1): 255–288. https://doi.org/10.1111/phis.12124.

Nute, Donald. 1984. "Conditional Logic." In *Handbook of Philosophical Logic: Volume II: Extensions of Classical Logic,* edited by D. Gabbay and F. Guenthner. Dordrecht: Springer. https://doi.org/https://doi.org/10.1007/978-94-009-6259-0_8.

Oaksford, Mike, and Nick Chater. 2010. *Cognition and Conditionals. Probability and Logic in Human Thinking.* Oxford: Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780199233298.001.0001.

Plebani, Matteo. 2020. "Why Aboutness Matters: Meta-Fictionalism as a Case Study." *Philosophia* 49 (3): 1177–1186. https://doi.org/10.1007/s11406-020-00272-9.

Plebani, Matteo, and Giuseppe Spolaore. 2021. "Subject Matter: A Modest Proposal." *The Philosophical Quarterly* 71 (3): 605–622. https://doi.org/10.1093/pq/pqaa054.

Priest, Graham. 2001. *An Introduction to Non-Classical Logic, 2nd ed. 2008.* Cambridge: Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511801174.

Ramsey, Frank P. 1990. "General Propositions and Causality." In *Philosophical Papers,* edited by D.H. Mellor, 145–163. Cambridge: Cambridge University Press.

Roberts, Craige. 2011. "Topics." In *Semantics: An International Handbook of Natural Language Meaning,* edited by Maienborn Heusinger and Portner, 2:1908–1934. De Gruyter Mouton.

———. 2012. "Information Structure: Towards an Integrated Formal Theory of Pragmatics." *Semantics and Pragmatics* 5:1–69. https://doi.org/http://dx.doi.org/10.3765/sp.5.6.

Rott, Hans. 2022. "Difference-Making Conditionals and the Relevant Ramsey Test." *The Review of Symbolic Logic* 15 (1): 133–164. https://doi.org/10.1017/s1755020319000674.

Schulz, Moritz. 2017. *Counterfactuals and Probability.* Oxford: Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780198785958.001.0001.

Segerberg, Krister. 1989. "Notes on conditional logic." *Studia Logica* 48:157–168. https://doi.org/10.1007/bf02770509.

Shaffer, Michael J. 2001. "Bayesian Confirmation of Theories That Incorporate Idealizations." *Philosophy of Science* 68:36–52. https://doi.org/10.1086/392865.

———. 2012. *Counterfactuals and Scientific Realism.* Dordrecht: Springer. https://doi.org/https://doi.org/10.1057/9781137271587.

———. 2017. ""Filling In", Thought Experiments and Intuitions." *Episteme* 14:255–262. https://doi.org/10.1017/epi.2016.15.

Skovgaard-Olsen, Niels, Henrik Singmann, and Karl Christoph Klauer. 2016. "The Relevance Effect and Conditionals." *Cognition* 150:26–36. https://doi.org/10.1016/j.cognition.2015.12.017.

Stalnaker, Robert. 1968. "A Theory of Conditionals." In *Studies in Logical Theory (American Philosophical Quarterly Monographs 2),* edited by Nicholas Rescher, 98–112. Oxford: Blackwell.

———. 1975. "Indicative Conditionals." *Philosophia* 5 (3): 269–286. https://doi.org/10.1007/bf02379021.

Williams, J. Robert G. 2010. "Defending Conditional Excluded Middle." *Noûs* 44 (4): 650–668. https://doi.org/10.1111/j.1468-0068.2010.00766.x.

———. 2012. "Counterfactual Triviality: A Lewis-Impossibility Argument for Counterfactuals." *Philosophy and Phenomenological Research* 85 (3): 648–670. https://doi.org/10.1111/j.1933-1592.2012.00636.x.

Williamson, Timothy. 2007. *The Philosophy of Philosophy.* Oxford: Blackwell.

———. 2016. "Knowing by Imagining." In *Knowledge Through Imagination,* edited by A. Kind and P. Kung, 113–23. Oxford: Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780198716808.003.0005.

———. 2020. *Suppose and Tell. The Semantics and Heuristics of Conditionals.* Oxford: Oxford University Press. https://doi.org/https://doi.org/10.1080/01445340.2021.1958648.

Yablo, Stephen. 2014. *Aboutness.* Princeton: Princeton University Press. https://doi.org/https://doi.org/10.1515/9781400845989.

APPENDIX A. PROOF OF LEMMA 1

For the first part, easily: validity implies p-validity by definition; the latter is a special case of the former obtained by restricting validity to the class of singleton topic models. Consider a sample valid but t-invalid inference: $p \wedge \neg p \vdash p > q$. To show its validity, let $\mathcal{N} = (W, f, \lambda, V, \pi)$ be a probabilistic model and $\mathcal{T} = \langle T, \oplus, t, k \rangle$ a topic model. By Definition 6, we have $\mathcal{A}(p \wedge \neg p) = \pi(p \wedge \neg p) = 0$, thus, $\mathcal{U}(p \wedge \neg p) = 1$. As $\mathcal{U}(p > q) \in [0, 1]$ by the definition of $\mathcal{U}$, we obtain that $\mathcal{U}(p \wedge \neg p) \geq \mathcal{U}(p > q)$. To show its t-invalidity, take the topic model $\langle \{a, b\}, \oplus, k, t \rangle$ such that $\oplus$ is idempotent and $a \oplus b = a$, thus, $b \sqsubset a$. Moreover, k is a constant function and $t_p = b$ and $t_q = a$. Therefore, $b = t_p = k(t_p)$ but $a = t_q \not\sqsubseteq k(t_p) = b$ (see Figure 1).

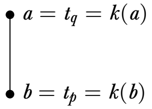

Figure 1. Topic model $\langle \{a, b\}, \oplus, t, k \rangle$

For the second part, suppose that $\Gamma \vdash \Delta$ is both p- and t-valid. Let $\mathcal{N} = (W, f, \lambda, V, \pi)$ be a probabilistic model and $\mathcal{T} = \langle T, \oplus, t, k \rangle$ a topic model. Since $\Gamma \vdash \Delta$ is t-valid, we have two cases:

Case 1: Every conditional in $\Gamma \cup \Delta$ is an on-topic conditional wrt $\mathcal{T}$. Then validity and p-validity coincide, thus $\Gamma \vdash \Delta$ is valid.

Case 2: There is a conditional in Γ that is not an on-topic conditional wrt $\mathcal{T}$.
Wlog, suppose that $\varphi \in \Gamma$ is not an on-topic conditional wrt $\mathcal{T}$. Thus, $\mathcal{U}(\varphi) = 1$ (by Definition 6.2). Recall that $\mathcal{U}(\chi) \in [0, 1]$ for all $\chi \in \mathcal{L}$. Therefore, we conclude that $\sum_{\varphi \in \Gamma} \mathcal{U}(\varphi) \geq \mathcal{U}(\chi)$ for all $\chi \in \Delta$.

APPENDIX B. PROOF OF THEOREM 2

The following lemmas will be useful in proving Theorem 2.

Lemma 3. Given a probabilistic model (W, f, λ, V, π) and $\varphi, \psi \in \mathcal{L}_{PL}$,

1. if $\pi(\lambda) = 1$, then $\pi(\varphi) = 1$ and $\pi^{\varphi}(\lambda) = 1$, therefore, $\pi^{\varphi}(\psi) = 1$;
2. $\pi^{\varphi}(\psi) = 1$ if $|\varphi| = \{\lambda\}$;
3. $\pi^{\varphi}(\varphi) = 1$;
4. π^{φ} is a probability mass function as described in Definition 3.

Proof.

1. Assume that $\pi(\lambda) = 1$. Then, since $\lambda \in |\varphi|$ for all $\varphi \in \mathcal{L}_{PL}$, we have $\pi(\varphi) = \sum_{w \in |\varphi|} \pi(w) = 1$. For the second part: by Definition 5, we have $\pi^{\varphi}(\lambda) = \sum_{w: f(w,\varphi)=\lambda} \pi(w)$. By the fact that $\lambda \in |\varphi|$ and Definition 1.3, we know that $f(\lambda, \varphi) = \lambda$. Therefore, since $\pi(\lambda) = 1$, we obtain that $\pi^{\varphi}(\lambda) = \sum_{w: f(w,\varphi)=\lambda} \pi(w) = 1$. As $\lambda \in |\psi|$, we also have $\pi^{\varphi}(\psi) = \sum_{w \in |\psi|} \pi^{\varphi}(w) = 1$.

2. Let $\varphi \in \mathcal{L}_{PL}$ such that $|\varphi| = \{\lambda\}$. Then, by Definition 1.2, we know that $f(w, \varphi) = \lambda$ for all $w \in W$. Therefore, $\pi^{\varphi}(\lambda) = \sum_{w: f(w,\varphi)=\lambda} \pi(w) = \sum_{w \in W} \pi(w) = 1$ (since π is a probability mass function). Since $\lambda \in |\psi|$ for all $\psi \in \mathcal{L}_{PL}$, we obtain that $\pi^{\varphi}(\psi) = \sum_{w \in |\psi|} \pi^{\varphi}(w) = 1$.

3. $\pi^{\varphi}(\varphi) = \sum_{w' \in |\varphi|} \pi^{\varphi}(w') = \sum_{w' \in |\varphi|} (\sum_{w: f(w,\varphi)=w'} \pi(w)) = \sum_{w \in W} \pi(w) = 1$ (the last step follows from Definition 1.1).

4. We only need to show that either $\pi^{\varphi}(\lambda) = 1$ or $\sum_{w \in W \setminus \{\lambda\}} \pi^{\varphi}(w) = 1$. Suppose that $\pi^{\varphi}(\lambda) \neq 1$. By item 1, we obtain that $\pi(\lambda) \neq 1$. Therefore, by the conditions of probabilistic mass functions given in Definition 3, that $\sum_{w \in W \setminus \{\lambda\}} \pi(w) = 1$. Then the result follows from the following equation: $\sum_{w \in W \setminus \{\lambda\}} \pi^{\varphi}(w) = \sum_{w \in W \setminus \{\lambda\}} (\sum_{w': f(w',\varphi)=w} \pi(w')) = \sum_{w \in W \setminus \{\lambda\}} \pi(w)$.

□

Lemma 4. Given a Stalnaker model (W, f, λ, V), $\varphi, \psi \in \mathcal{L}_{PL}$ and $w \in W$,

1. if $f(w, \varphi) \in |\psi|$, then $f(w, \varphi \wedge \psi) = f(w, \varphi)$,

2. $f(w, \varphi \vee \psi) = f(w, \varphi)$ or $f(w, \varphi \vee \psi) = f(w, \psi)$.

Proof.

1. Suppose that $f(w, \varphi) \in |\psi|$. Then, by Definition 1.1, we have $f(w, \varphi) \in |\varphi \wedge \psi|$. Since $f(w, \varphi \wedge \psi) \in |\varphi|$ as well, by Definition 1.4, we obtain that $f(w, \varphi \wedge \psi) = f(w, \varphi)$.

2. By Definition 1.1, we have that (1) $f(w, \varphi \vee \psi) \in |\varphi \vee \psi|$, (2) $f(w, \varphi) \in |\varphi \vee \psi|$, and (3) $f(w, \psi) \in |\varphi \vee \psi|$. (1) implies that $f(w, \varphi \vee \psi) \in |\varphi|$ or $f(w, \varphi \vee \psi) \in |\psi|$.
   If $f(w, \varphi \vee \psi) \in |\varphi|$, by (2) and Definition 1.4, we obtain $f(w, \varphi \vee \psi) = f(w, \varphi)$.
   If $f(w, \varphi \vee \psi) \in |\psi|$, by (3) and Definition 1.4, we obtain $f(w, \varphi \vee \psi) = f(w, \psi)$.

□

B.1. Proof of Theorem 2.1. Let $\mathcal{N} = (W, f, \lambda, V, \pi)$ be a probabilistic model and $\mathcal{T} = \langle T, \oplus, t, k \rangle$ a topic model. By Lemma 1, we only need to show that REF, ANT, CM, CC, CSO, CT, CMon, OR, and Modus Ponens are both p- and t-valid.

The proofs of t-validity follow straightforwardly from Definition 2, so we skip the details (see also Berto and Özgün (2021, Appendix) for similar proofs).

For p-validity, assume T is a singleton. Every conditional is on-topic with respect to a topic model $\mathcal{T} = \langle T, \oplus, t, k \rangle$ with a singleton T. Moreover, we present the proofs only for π's such that $\pi(\lambda) = 0$ because the degrees of unacceptability of the conclusions in Δ equal to 0 with respect to a π such that $\pi(\lambda) = 1$ (see Lemma 3.1). Finally, given a premise-conclusion rule $\Gamma \vdash \Delta$, we only consider the cases where the antecedents of the conditionals in Δ are possible with respect to the given Stalnaker model (W, f, λ, V) since otherwise the degrees of unacceptability of the conclusions again equal to 0 (see Lemma 3.2). Due to the structure of the rules in Theorem 2.1, we cannot have that the antecedents of all the premises are impossible but the antecedent of the conclusion is possible.

REF: $\vdash \varphi > \varphi$
p-valid: By Lemma 3.3, we have $\pi^{\varphi}(\varphi) = 1$. Moreover, we also have that $t_{\varphi} \sqsubseteq k(t_{\varphi})$. Therefore, $\mathcal{U}(\varphi > \varphi) = 1 - \pi^{\varphi}(\varphi) = 0$.

ANT: $\varphi > \psi \vdash \varphi > (\varphi \wedge \psi)$
p-valid: As T is a singleton, we have $\mathcal{U}(\varphi > \psi) = 1 - \pi^{\varphi}(\psi)$ and $\mathcal{U}(\varphi > (\varphi \wedge \psi)) = 1 - \pi^{\varphi}(\varphi \wedge \psi)$. Observe that $\pi^{\varphi}(\varphi \wedge \psi) = \pi^{\varphi}(\varphi) + \pi^{\varphi}(\psi) - \pi^{\varphi}(\varphi \vee \psi)$ since π is a probability function. Moreover, by Lemma 3.3 and the fact that π is a probability function, we have $\pi^{\varphi}(\varphi) = 1$ and $\pi^{\varphi}(\varphi \vee \psi) = 1$. Therefore, $\mathcal{U}(\varphi > \psi) = \mathcal{U}(\varphi > (\varphi \wedge \psi))$.

CM: $\varphi > (\psi \wedge \chi) \vdash \varphi > \psi, \varphi > \chi$
p-valid: As T is a singleton, we have $\mathcal{U}(\varphi > (\psi \wedge \chi)) = 1 - \pi^{\varphi}(\psi \wedge \chi)$, $\mathcal{U}(\varphi > \psi) = 1 - \pi^{\varphi}(\psi)$, and $\mathcal{U}(\varphi > \chi) = 1 - \pi^{\varphi}(\chi)$. Since $\psi \wedge \chi \models_{PL} \psi$ and $\psi \wedge \chi \models_{PL} \chi$, and π^{φ} is a probability function, we know that $\pi^{\varphi}(\psi \wedge \chi) \leq \pi^{\varphi}(\psi)$ and $\pi^{\varphi}(\psi \wedge \chi) \leq \pi^{\varphi}(\chi)$. Therefore, $\mathcal{U}(\varphi > (\psi \wedge \chi)) \geq \mathcal{U}(\varphi > \psi)$ and $\mathcal{U}(\varphi > (\psi \wedge \chi)) \geq \mathcal{U}(\varphi > \chi)$.

CC: $\varphi > \psi, \varphi > \chi \vdash \varphi > (\psi \wedge \chi)$
p-valid: As T is a singleton, we have:

$\mathcal{U}(\varphi > \psi) + \mathcal{U}(\varphi > \chi) = 1 - \pi^{\varphi}(\psi) + 1 - \pi^{\varphi}(\chi)$

(since $\pi^{\varphi}(\psi \vee \chi) \leq 1$)

$$\geq 1 - \pi^{\varphi}(\psi) - \pi^{\varphi}(\chi) + \pi^{\varphi}(\psi \vee \chi)$$

(since π^{φ} is a probability function)

$$= 1 - \pi^{\varphi}(\psi \wedge \chi)$$

(by the defn. of $\mathcal{U}$)

$$= \mathcal{U}(\varphi > (\psi \wedge \chi))$$

CSO: $\varphi > \psi, \psi > \varphi, \varphi > \chi \vdash \psi > \chi$
p-valid: We need to show that:

$$\mathcal{U}(\varphi > \psi) + \mathcal{U}(\psi > \varphi) + \mathcal{U}(\varphi > \chi) \geq \mathcal{U}(\psi > \chi)$$

i.e., that

$$\pi^{\varphi}(\psi) + \pi^{\psi}(\varphi) + \pi^{\varphi}(\chi) - \pi^{\psi}(\chi) \leq 2$$

Recall that:
$\pi^{\varphi}(\psi) = \sum_{w' \in |\psi|}(\sum_{w:f(w,\varphi)=w'} \pi(w))$
$\pi^{\psi}(\varphi) = \sum_{w' \in |\varphi|}(\sum_{w:f(w,\psi)=w'} \pi(w))$
$\pi^{\varphi}(\chi) = \sum_{w' \in |\chi|}(\sum_{w:f(w,\varphi)=w'} \pi(w))$
$\pi^{\psi}(\chi) = \sum_{w' \in |\chi|}(\sum_{w:f(w,\psi)=w'} \pi(w))$

We explain the main idea behind the proof and leave the details to the reader. For each $w \in W$, $\pi(w)$ is added to the total sum $\pi^{\varphi}(\psi) + \pi^{\psi}(\varphi) + \pi^{\varphi}(\chi) - \pi^{\psi}(\chi)$ at most twice. So the whole sum adds up to at most 2. In particular, for any w such that $\pi(w) > 0$, if $\pi(w)$ is added to $\pi^{\varphi}(\psi)$, $\pi^{\psi}(\varphi)$ and $\pi^{\varphi}(\chi)$, it is also added to $\pi^{\psi}(\chi)$.

Let $w \in W$ such that $\pi(w) > 0$ and suppose that $\pi(w)$ is added to $\pi^{\varphi}(\psi)$, $\pi^{\psi}(\varphi)$, and $\pi^{\varphi}(\chi)$. If $\pi(w)$ is added to both $\pi^{\varphi}(\psi)$ and $\pi^{\psi}(\varphi)$, we have $f(w, \varphi) \in |\psi|$ and $f(w, \psi) \in |\varphi|$. This implies, by Definition 1.4, that $f(w, \varphi) = f(w, \psi)$. Now suppose further that $\pi(w)$ is added to $\pi^{\varphi}(\chi)$ but not to $\pi^{\varphi}(\chi)$. The former means that $f(w, \varphi) \in |\chi|$ and the latter that $f(w, \psi) \notin |\chi|$, contradicting $f(w, \varphi) = f(w, \psi)$. Therefore, for each $w \in W$, $\pi(w)$ is added to the total sum $\pi^{\varphi}(\psi) + \pi^{\psi}(\varphi) + \pi^{\varphi}(\chi) - \pi^{\psi}(\chi)$ at most twice. This means that

$$\pi^{\varphi}(\psi) + \pi^{\psi}(\varphi) + \pi^{\varphi}(\chi) - \pi^{\psi}(\chi) \leq 2 \cdot \sum_{w \in W} \pi(w) = 2.$$

CT: $\varphi > \psi, (\varphi \wedge \psi) > \chi \vdash \varphi > \chi$

p-valid: The proof is similar to the proof for CSO and uses Lemma 4.1. We need to show that

$$\mathcal{U}(\varphi > \psi) + \mathcal{U}((\varphi \wedge \psi) > \chi) \geq \mathcal{U}(\varphi > \chi)$$

i.e., that

$$\pi^{\varphi}(\psi) + \pi^{\varphi \wedge \psi}(\chi) - \pi^{\varphi}(\chi) \leq 1$$

Recall that:
$\pi^{\varphi}(\psi) = \sum_{w' \in |\psi|} (\sum_{w: f(w,\varphi)=w'} \pi(w))$
$\pi^{\varphi \wedge \psi}(\chi) = \sum_{w' \in |\chi|} (\sum_{w: f(w,\varphi \wedge \psi)=w'} \pi(w))$
$\pi^{\varphi}(\chi) = \sum_{w' \in |\chi|} (\sum_{w: f(w,\varphi)=w'} \pi(w))$

For each $w \in W$, $\pi(w)$ is added to the total sum $\pi^{\varphi}(\psi) + \pi^{\varphi \wedge \psi}(\chi) - \pi^{\varphi}(\chi)$ at most once, therefore, the whole sum adds up to at most 1. In particular, for any w such that $\pi(w) > 0$, if $\pi(w)$ is added to $\pi^{\varphi}(\psi)$ and $\pi^{\varphi \wedge \psi}(\chi)$, it is also added to $\pi^{\varphi}(\chi)$.

Let $w \in W$ such that $\pi(w) > 0$ and suppose that $\pi(w)$ is added to $\pi^{\varphi}(\psi)$ and $\pi^{\varphi \wedge \psi}(\chi)$, but not to $\pi^{\varphi}(\chi)$. This implies that $f(w,\varphi) \in |\psi|$, $f(w, \varphi \wedge \psi) \in |\chi|$ but $f(w,\varphi) \notin |\chi|$. However, by Lemma 4.1, we have that $f(w,\varphi) = f(w, \varphi \wedge \psi)$, contradicting $f(w, \varphi \wedge \psi) \in |\chi|$ but $f(w,\varphi) \notin |\chi|$. Therefore, for each $w \in W$, $\pi(w)$ is added to the total sum $\pi^{\varphi}(\psi) + \pi^{\varphi \wedge \psi}(\chi) - \pi^{\varphi}(\chi)$ at most once. This means that

$$\pi^{\varphi}(\psi) + \pi^{\varphi \wedge \psi}(\chi) - \pi^{\varphi}(\chi) \leq \sum_{w \in W} \pi(w) = 1.$$

CMon: $\varphi > \psi, \varphi > \chi \vdash (\varphi \wedge \psi) > \chi$
p-valid: The proof is similar to the proof for CT. We now need to show that

$$\pi^{\varphi}(\psi) + \pi^{\varphi}(\chi) - \pi^{\varphi \wedge \psi}(\chi) \leq 1$$

Let $w \in W$ such that $\pi(w) > 0$ and suppose that $\pi(w)$ is added to $\pi^{\varphi}(\psi)$ and $\pi^{\varphi}(\chi)$, but not to $\pi^{\varphi \wedge \psi}(\chi)$. This implies that $f(w,\varphi) \in |\psi|$, $f(w,\varphi) \in |\chi|$ but $f(w, \varphi \wedge \psi) \notin |\chi|$. However, by Lemma 4.1, we have that $f(w,\varphi) = f(w, \varphi \wedge \psi)$, contradicting $f(w, \varphi \wedge \psi) \notin |\chi|$ but $f(w,\varphi) \in |\chi|$. Therefore, as in the previous case, we obtain that

$$\pi^{\varphi}(\psi) + \pi^{\varphi}(\chi) - \pi^{\varphi \wedge \psi}(\chi) \leq \sum_{w \in W} \pi(w) = 1.$$

OR: $\varphi > \psi, \chi > \psi \vdash (\varphi \vee \chi) > \psi$
p-valid: The proof is similar to the proof for the above cases and uses Lemma 4.2. We need to show that

$$\mathcal{U}(\varphi > \psi) + \mathcal{U}(\chi > \psi) \geq \mathcal{U}((\varphi \vee \chi) > \psi)$$

i.e., that

$$\pi^{\varphi}(\psi) + \pi^{\chi}(\psi) - \pi^{\varphi\vee\chi}(\psi) \leq 1.$$

Let $w \in W$ such that $\pi(w) > 0$ and suppose that $\pi(w)$ is added to $\pi^{\varphi}(\psi)$ and $\pi^{\chi}(\psi)$, but not to $\pi^{\varphi\vee\chi}(\psi)$. This implies that $f(w,\varphi) \in |\psi|$, $f(w,\chi) \in |\psi|$ but $f(w,\varphi\vee\chi) \notin |\psi|$. However, by Lemma 4.2, we have that $f(w,\varphi) = f(w,\varphi\vee\chi)$ or $f(w,\chi) = f(w,\varphi\vee\chi)$, contradicting $f(w,\varphi) \in |\psi|$, $f(w,\chi) \in |\psi|$ but $f(w,\varphi\vee\chi) \notin |\psi|$. Therefore, for each $w \in W$, $\pi(w)$ is added to the total sum $\pi^{\varphi}(\psi) + \pi^{\chi}(\psi) - \pi^{\varphi\vee\chi}(\psi)$ at most once. This means that

$$\pi^{\varphi}(\psi) + \pi^{\chi}(\psi) - \pi^{\varphi\vee\chi}(\psi) \leq \sum_{w\in W} \pi(w) = 1.$$

Modus Ponens: $\varphi, \varphi > \psi \vdash \psi$
p-valid: We are only interested in simple counterfactuals so φ, ψ in the formulation of Modus Ponens are Booleans, that is, $\varphi, \psi \in \mathcal{L}_{PL}$. We need to show that

$$\mathcal{U}(\varphi) + \mathcal{U}(\varphi > \psi) \geq \mathcal{U}(\psi)$$

i.e., that

$$\pi(\varphi) + \pi^{\varphi}(\psi) - \pi(\psi) \leq 1.$$

Let $w \in W$ such that $\pi(w) > 0$ and suppose that $\pi(w)$ is added to $\pi(\varphi)$ and $\pi^{\varphi}(\psi)$, but not to $\pi(\psi)$. This implies that $w \in |\varphi|$, $f(w,\varphi) \in |\psi|$, but $w \notin |\psi|$. However, since $w \in |\varphi|$, by Definition 1.3, we have that $f(w,\varphi) = w$. This contradicts with $f(w,\varphi) \in |\psi|$ and $w \notin |\psi|$. Therefore, for each $w \in W$, $\pi(w)$ is added to the total sum $\pi(\varphi) + \pi^{\varphi}(\psi) - \pi(\psi)$ at most once. This means that

$$\pi(\varphi) + \pi^{\varphi}(\psi) - \pi(\psi) \leq \sum_{w\in W} \pi(w) = 1.$$

B.2. Proof of Theorem 2.2. *t*-invalidity proofs follow exactly as in the corresponding *t*-invalidity proofs in (Berto and Özgün 2021). We here repeat those proofs for the convenience of the reader. For *p*-validity proofs, let $\mathcal{N} = (W, f, \lambda, V, \pi)$ be a probabilistic model and $\mathcal{T} = \langle T, \oplus, t, k\rangle$ a topic model with singleton T.

MOD: $\neg\varphi > \varphi \vdash \psi > \varphi$
We want to show that $\mathcal{U}(\neg\varphi > \varphi) \geq \mathcal{U}(\psi > \varphi)$.

Case 1: $\neg\varphi$ is possible
Then, $\pi^{\neg\varphi}(\varphi) = 0$. Since $t_{\varphi} \sqsubseteq k(t_{\neg\varphi})$, we have $\mathcal{U}(\neg\varphi > \varphi) = 1 - \pi^{\neg\varphi}(\varphi) = 1$. Therefore, since $\mathcal{U}(\psi > \varphi) \in [0,1]$, we obtain the result.

Case 2: $\neg\varphi$ is impossible
Then, $\pi^{\neg\varphi}(\varphi) = 1$ (by Lemma 3.2), thus, $\mathcal{U}(\neg\varphi > \varphi) = 0$. Since $\mathcal{T}$ is

a singleton model, we have $\mathcal{U}(\psi > \varphi) = 1 - \pi^{\psi}(\varphi)$. If ψ is impossible, then $\pi^{\psi}(\varphi) = 1$, thus, $\mathcal{U}(\psi > \varphi) = 0$. If ψ is possible, then $\pi^{\psi}(\varphi) = 1$ since $\neg\varphi$ is impossible. Therefore, $\mathcal{U}(\psi > \varphi) = 0$.

t-invalidity: Consider the instance $\neg p \to p \vdash q \to p$ and the topic model $\langle \{a, b\}, \oplus, t, k\rangle$ such that $\oplus$ is idempotent and $a \oplus b = a$, thus, $b \sqsubset a$. Moreover, k is a constant function and $t_q = b$ and $t_p = a$. Therefore, $a = t_p = k(t_{\neg p})$ but $a = t_p \not\sqsubseteq k(t_q) = b$ (see Figure 2).

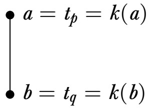

Figure 2. Topic model $\langle \{a, b\}, \oplus, t, k\rangle$

RCE: If $\varphi \vdash_{PL} \psi$, then $\vdash \varphi > \psi$
p-validity: Follows immediately from Lemma 3.3 and the fact that π^{φ} is a probability function.

t-invalidity: Consider the counterexample given in Figure 2, where $\varphi :=$ q and $\psi : p \vee \neg p$.

RCEA: If $\vdash_{PL} \varphi \equiv \psi$, then $\varphi > \chi \dashv\vdash \psi > \chi$
p-validity: It is easy to see by Definition 1 that if $|\varphi| = |\psi|$, then $f(w, \varphi) = f(w, \psi)$ for all $w \in W$. Therefore, whenever $|\varphi| = |\psi|$, we have $\pi^{\varphi} = \pi^{\psi}$. Then the p-validity follows.

t-invalidity: Consider the counterexample given in Figure 2 and take $\varphi := p \vee \neg p$, $\psi := q \vee \neg q$, $\chi := r \vee \neg r$ such that $t_r = a$. Then, $\models_{PL} \varphi \equiv \psi$, $(p \vee \neg p) \to (r \vee \neg r)$ is an on-topic conditional wrt $\mathcal{T}$ but $(q \vee \neg q) \to (r \vee \neg r)$ is not.

RCEC: If $\vdash_{PL} \varphi \equiv \psi$, then $\chi > \varphi \dashv\vdash \chi > \psi$
p-validity: Follows immediately from Definition 5.

t-invalidity: Consider the counterexample above but take $t_p = t_r = b$ and $t_q = a$.

RCK: If $\vdash_{PL} (\varphi_1 \wedge \dots \wedge \varphi_n) \supset \psi$, then $\chi > \varphi_1, \dots, \chi > \varphi_n \vdash \chi > \psi$
p-validity: Suppose that $\vdash_{PL} (\varphi_1 \wedge \dots \wedge \varphi_n) \supset \psi$. We want to show that $\mathcal{U}(\chi > \varphi_1) + \dots \mathcal{U}(\chi > \varphi_n) \geq \mathcal{U}(\chi > \psi)$, i.e., that $1 - \pi^{\chi}(\varphi_1) + \dots + 1 - \pi^{\chi}(\varphi_n) \geq 1 - \pi^{\chi}(\psi)$, i.e., that $\pi^{\chi}(\psi) \geq \pi^{\chi}(\varphi_1) + \dots \pi^{\chi}(\varphi_n) - n + 1$.

It is easy to see that

(by the assumption and π^{χ} is a prob. func.)
$\pi^{\chi}(\psi) \geq \pi^{\chi}(\varphi_1 \wedge \cdots \wedge \varphi_n)$
(π^{χ} is a probability function)
$\geq \pi^{\chi}(\varphi_1) + \ldots \pi^{\chi}(\varphi_n) - n + 1$

t-invalidity: Same as the proof of RCEC.

RCM: $\vdash_{PL} \varphi \supset \psi$, then $\chi > \varphi \vdash \chi > \psi$
p-validity: Follows from the fact that π^{χ} is a probability function.

t-invalidity: Same as the proof of RCEC.

And-to-If: $\varphi \wedge \psi \vdash \varphi > \psi$
p-validity: We want to show that $\pi^{\varphi}(\psi) \geq \pi(\varphi \wedge \psi)$. By Definition 1.3, for all $w \in |\varphi \wedge \psi|$, $f(w, \varphi) = w$. Therefore, for all $w \in |\varphi \wedge \psi|$, $\pi^{\varphi}(w) \geq \pi(w)$. Thus,

$$\pi^{\varphi}(\psi) = \sum_{w' \in |\psi|} \pi^{\varphi}(w') \geq \sum_{w' \in |\varphi \wedge \psi|} \pi^{\varphi}(w') \geq \sum_{w' \in |\varphi \wedge \psi|} \pi(w') = \pi(\varphi \wedge \psi).$$

t-invalidity: See the counterexample given in Figure 2 and take $\varphi := q$ and $\psi := p$.

B.3. Proof of Theorem 2.3. The proofs of *t*-validity are straightforward, so we skip the details (see also Berto and Özgün (2021, Appendix) for similar proofs).

Trans: $\varphi > \psi, \psi > \chi \vdash \varphi > \chi$
p-invalidity: Consider the probabilistic model $\mathcal{N} = (W, f, \lambda, V, \pi)$ such that $W = \{w_1, w_2, \lambda\}$, $V(p) = \{w_1\}$, $V(q) = \{w_1, w_2\}$, and $V(r) = \{w_2\}$, f satisfies Definition 1 such that $f(w_2, p) = \{w_1\}$, and $\pi(w_1) = \pi(w_2) = 1/2$. Then, $\pi^{p}(q) = 1, \pi^{q}(r) = 1/2$, and $\pi^{p}(r) = 0$. We then obtain that $\mathcal{U}(p > q) + \mathcal{U}(q > r) = 0 + 1/2 < \mathcal{U}(p > r) = 1$.

SA: $\varphi > \psi \vdash (\varphi \wedge \chi) > \psi$
p-invalidity: Consider the probabilistic model $\mathcal{N} = (W, f, \lambda, V, \pi)$ such that $W = \{w_1, w_2, w_3, \lambda\}$, $V(p) = \{w_1, w_2, w_3\}$, $V(q) = \{w_1, w_2\}$, and $V(r) = \{w_2, w_3\}$, f satisfies Definition 1 such that $f(w_1, p \wedge r) = \{w_3\}$, and $\pi(w_1) = \pi(w_2) = \pi(w_3) = 1/3$. Then, $\pi^{p}(q) = 2/3$ and $\pi^{p \wedge r}(q) = 1/3$. We then obtain that $\mathcal{U}(p > q) = 1/3 < \mathcal{U}((p \wedge r) > q) = 2/3$.

B.4. Proof of Theorem 2.4.
Or-to-if: $\varphi \vee \psi \vdash \neg\varphi > \psi$
p-invalidity: Consider the probabilistic model $\mathcal{N} = (W, f, \lambda, V, \pi)$ such that $W = \{w_1, w_2, w_3, w_4, \lambda\}$, $V(p) = \{w_1, w_2\}$ and $V(q) = \{w_2, w_3\}$, f satisfies Definition 1 such that $f(w_1, \neg p) = f(w_2, \neg p) = w_4$, and $\pi(w_1) = \pi(w_2) = 1/3$, $\pi(w_3) = \pi(w_4) = 1/6$. Then, $\pi(p \vee q) = 5/6$ and $\pi^{\neg p}(q) = 1/6$. We then obtain that $\mathcal{U}(p \vee q) = 1/6 < \mathcal{U}(\neg p > q) = 5/6$.

Contraposition: $\varphi > \neg\psi \vdash \psi > \neg\varphi$
p-invalidity: Consider the probabilistic model $\mathcal{N} = (W, f, \lambda, V, \pi)$ such that $W = \{w_1, w_2, w_3, \lambda\}$, $V(p) = \{w_1, w_2\}$ and $V(q) = \{w_2, w_3\}$, f satisfies Definition 1 such that $f(w_1, q) = w_2$ and $f(w_3, p) = w_1$, and $\pi(w_1) = \pi(w_2) = \pi(w_3) = 1/3$. Then, $\pi^{p}(\neg q) = 2/3$ and $\pi^{q}(\neg p) = 1/3$. We then obtain that $\mathcal{U}(p > \neg q) = 1/3 < \mathcal{U}(q > \neg p) = 2/3$.

SDA: $(\varphi \vee \psi) > \chi \vdash \varphi > \chi, \psi > \chi$
p-invalidity: Consider the probabilistic model $\mathcal{N} = (W, f, \lambda, V, \pi)$ such that $W = \{w_1, w_2, w_3, \lambda\}$, $V(p) = \{w_1\}$, $V(q) = \{w_2\}$, and $V(r) = \{w_2, w_3\}$, f satisfies Definition 1 such that $f(w_2, p) = f(w_3, p) = w_1$ and $f(w_3, p \vee q) = w_2$, and $\pi(w_1) = 1/6, \pi(w_2) = 1/3$, $\pi(w_3) = 1/2$. Then, $\pi^{p \vee q}(r) = 5/6$ and $\pi^{p}(r) = 0$. We then obtain that $\mathcal{U}((p \vee q) > r) = 1/6 < \mathcal{U}(p > r) = 1$.

t-invalidity:
t-invalidity proofs follow exactly as in the corresponding *t*-invalidity proofs in (Berto and Özgün 2021). We here repeat those proofs for the convenience of the reader. For Or-to-If and Contraposition, consider the topic model given in Figure 2. This model *t*-invalidates Or-to-If since $q \vee p$ is not a conditional and $\neg q \rightarrow p$ is not an on-topic conditional wrt $\mathcal{T}$: $a = t_p \not\sqsubseteq k(t_{\neg q}) = b$. It also *t*-invalidates Contraposition since $p \rightarrow \neg q$ is an on-topic conditional wrt $\mathcal{T}$ (since $b = t_{\neg q} \sqsubseteq k(t_p) = a$) but $q \rightarrow \neg p$ is not (since $a = t_{\neg p} \not\sqsubseteq k(t_q) = b$). For SDA, consider the topic model $\mathcal{T}' = \langle \{a, b, c, d\}, \oplus', k', t' \rangle$ where $\oplus'$ is as depicted in Figure 3, k' is a constant function, and $t'_p = b$, $t'_q = c$, and $t'_r = a$. It is then easy to see that $(p \vee q) \rightarrow r$ is an on-topic conditional wrt $\mathcal{T}$, however, neither $p \rightarrow r$ nor $q \rightarrow r$ is.

B.5. Proof of Theorem 2.5. The invalidity of Trans, SA, Or-to-If, Contraposition, and SDA follows from Lemma 1, Theorems 2.3, and 2.4. For RCE, RCEA, RCEC, RCK, and RCM, take the counter-topic models given in the proof of Theorem 2.2 together with any arbitrary probability mass π: they constitute counterexamples for the respective validity claims, since in each case the degree of unacceptability of the elements in Γ is 0 and the degree of unacceptability of the elements

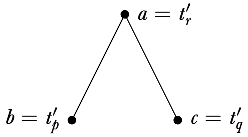

Figure 3. Counterexample for the *t*-invalidity of SDA

in Δ is 1 (since the conditionals in Δ are off-topic conditionals with respect to the corresponding topic models). For And-to-If, consider the instance $q \wedge p \vdash q \rightarrow p$. This is invalidated by the topic model given in Figure 2 plus a probability mass π such that $\pi(p \wedge q) \neq 0$ (as in the corresponding case in Berto and Özgün (2021)). Finally, MOD is invalid only when $\neg\varphi$ is impossible. Consider the instance $\neg(p \vee \neg p) \rightarrow (p \vee \neg p) \vdash q \rightarrow (p \vee \neg p)$. Observe that, for any probability mass π, we have (by Lemma 3.2), $\pi^{\neg(p \vee \neg p)}(p \vee \neg p) = 1$. Consider also the topic model given in Figure 2. Then, we have that $\mathcal{U}(\neg(p \vee \neg p) \rightarrow (p \vee \neg p)) = 1 - \pi^{\neg(p \vee \neg p)}(p \vee \neg p) = 0$. However, since $t_{p \vee \neg p} \not\sqsubseteq k(t_q)$, we have $\mathcal{U}(q \rightarrow (p \vee \neg p)) = 1$. Therefore, $\mathcal{U}(\neg(p \vee \neg p) \rightarrow (p \vee \neg p)) < \mathcal{U}(q \rightarrow (p \vee \neg p))$.

FRANCESCO BERTO[1],[2]
AYBÜKE ÖZGÜN[2]

[1] Arché Research Centre, University of St Andrews
[2] ILLC, Univeristy of Amsterdam