Impossible Worlds and Partial Belief

Edward Elliott

University of Leeds

One response to the problems of logical omniscience in epistemic logic is to extend the space of classically possible worlds to include impossible worlds. It is natural to think that essentially the same basic line of response can be applied to our standard probabilistic models of partial belief, for which the parallel problem of probabilistic coherence (which implies a kind of logical omniscience) also arises. In this paper, I note that there is a problem with the inclusion of impossible worlds into our probabilistic models. Most of the propositions which can be constructed from possible and impossible worlds are in an important sense *inexpressible*; whereas the probabilistic model seems committed to saying that agents in general have at least as many attitudes towards inexpressible propositions as they do towards expressible propositions. Since it is reasonable to think that (at least most of) our attitudes are expressible, a model with such commitments looks problematic.

Suppose we wish to model the total doxastic state of some thinking subject, who may or may not be ideally rational. By 'total doxastic state' I mean the sum total of facts about a subject's doxastic attitudes broadly construed, i.e., their full beliefs, partial beliefs, comparative degrees of confidence, and so on—those aspects of a subject's mental life which characterise how she takes the world to be.

We'll need two main ingredients. One, a way to represent potential objects of thought; i.e., the kinds of things fit to serve as the contents of some cognitive mental state. Important here is the capacity to adequately represent the apparent hyperintensionality of thought. On the face of it, belief cuts more finely than necessary equivalence, and this appearance needs to be accounted for. And two, we'll need some way of representing which of these are the contents of the subject's doxastic attitudes, for each different kind of attitude that we wish to represent.

In this paper, I want to argue that one common approach to modelling the hyperintensional objects of thought (*viz.*, as sets of possible and impossible worlds) does not sit nicely with another very common approach to modelling total doxastic states (*viz.*, as a probability function defined on a Boolean algebra of propositions). Roughly, the problem is that most of the propositions which can be constructed from possible and impossible worlds are in a certain strong sense *inexpressible*, while the probabilistic model will end up saying that agents *in general* have *at least as many* attitudes towards inexpressible propositions as they do towards expressible propositions. Since it is reasonable to think that at least most of our attitudes are expressible, a model with such commitments looks problematic.

In §1, I will outline a background assumption about the linguistic expressibility of thought which I will use to set up my main argument. Then, in §2, I will outline the problems of logical omniscience as they apply to a possible worlds model of full belief, and note how the introduction of impossible worlds is supposed to help solve these problems. §3 then introduces the probabilistic analogues to the problems of logical omniscience, for which an analogous solution seems to apply, and §4 presents the central argument of the paper. Finally, §5 considers in more depth the basis for (and importance of) the linguistic expressibility assumption, while §6 looks at the consequences of altering those aspects of the probabilistic model which give rise to the problem—specifically, the assumption of Booleanism.

Before moving on, it's worth flagging some things that I do *not* take myself to be arguing. First, I do not think that the mere *existence* of inexpressible propositions is problematic for the impossible worlds model—nor indeed should it be considered especially problematic for the possible worlds model. I doubt that it would be a significant problem if our models treated some of inexpressible propositions as potential objects of belief for some believers. I do, however, think that there is a problem when our models commit us to saying that subjects systematically believe a very large number of them, and it is this problem that I intend to highlight here. (See §5 for more discussion.) And second, my argument is not against the intelligibility of impossible worlds, nor do I want to claim that there are no benefits to be had by including them within our ontology.

1. The Expressibility Assumption

In setting up my argument, I will presuppose the existence of a language, $\mathcal{L}$, about which I will need to make some assumptions. $\mathcal{L}$ can be thought of as a class of declarative sentences, each a (possibly infinite) string of symbols taken from a (possibly infinite) alphabet, with a corresponding interpretation. The interpretation of every sentence in $\mathcal{L}$ should be non-ambiguous, precise, and context-independent. I'll stick to characterising $\mathcal{L}$ at the sentential level, since it is here that the issues we will be interested in arise. Nothing in what follows should be taken to suggest that there can be no quantifiers, modal operators, and so on, in $\mathcal{L}$.

I will assume that if α is the thinking subject whom we wish to model, then $\mathcal{L}$ is *at least* expressive enough to be capable of representing everything towards which α might have beliefs (or partial beliefs).[1] Important here is that $\mathcal{L}$ has the appropriate degree of granularity to adequately distinguish between distinct belief contents. So, if α can have a belief with such-and-such a content, then it had better be possible to express *precisely* that content, and *just* that content, in $\mathcal{L}$. Given this basic assumption, it's reasonable to expect that $\mathcal{L}$ will contain at least the negation (¬), conjunction (∧), and disjunction (∨) sentential operators, and that it will be closed under at least finite iterations thereof. After all, if I can have any beliefs at all, then I can presumably have conjunctive beliefs, disjunctive beliefs, and (in the relevant sense) negative beliefs.

Whatever $\mathcal{L}$ is (if it exists), it's not English. You might think that a suitably regimented and perhaps significantly extended version of English might do the trick. More likely, I suspect, is that an appropriately constructed *Lagadonian* language will suit our purposes. A Lagadonian language is one wherein particulars are taken to be names of themselves, and properties and relations are taken to be predicates for themselves. Thus, for instance, the content *Frank is taller than Mary* may be some construction out of Frank, Mary, and the relation of *taller than*, e.g., <*taller than*, *Frank*, *Mary*>.[2] In a series of recent works, Mark Jago has defended the thesis that a Lagadonian language might be rich enough to adequately express

[1] Note that I am here ignoring any (partial) beliefs which might be, as Perry (1979) calls them, *essentially indexical*. The capacity to express irreducibly indexical beliefs in a language whose interpretation is stipulated as being context-independent may rightly be doubted. If need be, the arguments that follow could be naturally adapted without any significant changes to a centred worlds framework (see Lewis 1979), thus permitting the presence of context-dependent sentences in $\mathcal{L}$. However, for simplicity's sake I will ignore these complications in what follows, and pretend that there are no essentially indexical doxastic attitudes.

[2] The Lagadonian sentence representing the thought content which we would express in *English* with the sentence 'Frank is taller than Mary' *need not* include either Frank or Mary. A neo-descriptivist approach to the content of names will want to reject the translation of the English sentence into the Lagadonian <*taller than, Frank, Mary*> in favour of a more complicated construction which replaces Frank and Mary with Lagadonian definite descriptions which more accurately reflect their view of the semantics of those names. Likewise, everything said here is compatible with a two-dimensionalist approach to thought content. To say that thought contents can be expressed within a broadly Lagadonian language does not commit one to a particular kind of response to Frege puzzles, despite some initial appearances to the contrary.

each of our beliefs (see esp. his 2012; 2015a). Indeed, the expressive richness of Jago's language is a key component of his use of ersatz possible and impossible worlds to model hyperintensional contents, in more or less the manner described in the next section.

Still, if the reader finds it hard to swallow the idea that such a richly expressive $\mathcal{L}$ can even possibly exist, I ask that they hold off their objections for now—I will return to discuss the matter in some detail in §5, when the relative importance of this assumption in the context of my argument can be made clear. (Spoiler alert: it may well turn out that there are inexpressible objects of thought, but unless we have reasons to think that these are very frequently believed, then the main thrust of my discussion is unchanged.)

Two fiddly points to note about $\mathcal{L}$. First, I haven't assumed that any of ¬, ∧ or ∨ are primitives. Some find it useful to define some connectives in terms of others; e.g., by letting $S \vee S^*$ be a mere shorthand for $\neg(\neg S \wedge \neg S^*)$.[3] If we can legitimately take ¬ and ∧ as the only primitive connectives in $\mathcal{L}$, then in some respects the argument of §4 becomes more straightforward. However, in general I think it would be a mistake to suppose that ∨ should be reduced to ¬ and ∧, or that ∧ should be reduced to ¬ and ∨—or worse: that all of the above should be reduced to the Sheffer stroke! Part of the point of $\mathcal{L}$ is to capture distinctions in thought wherever they exist, and a language which treats disjunctions (for example) as mere shorthands for negated conjunctions won't cut finely enough for this purpose. The fact that two sentences with the forms $S \vee S^*$ and $\neg(\neg S \wedge \neg S^*)$ are (classically) logically equivalent is no guarantee that having a belief adequately expressed by a sentence of the form $S \vee S^*$ goes hand-in-hand with having a belief adequately expressed by a sentence of the form $\neg(\neg S \wedge \neg S^*)$: it's plausible that one may accept the former but not the latter, and if so then we will need a way to mark that distinction in $\mathcal{L}$.

Second, I have only assumed that $\mathcal{L}$ contains *at least* ¬, ∧, and ∨, *not* that these are the only connectives in $\mathcal{L}$. But should we also include, for example, the material conditional → (where $S \rightarrow S^*$ isn't just a shorthand for $\neg S \vee S^*$), or an exclusive disjunction $\vee_x$? Here, matters are much less clear to me. It's not obvious whether one can have a belief that's adequately expressed by a sentence of the form $S \vee_x S^*$ which is not equally well expressed by a sentence of the form $(S \vee S^*) \wedge \neg(S \wedge S^*)$. To the best of my awareness, this question is not discussed anywhere in the philosophical literature, and I will not try to settle it here—ultimately, it won't make a great difference to my argument.

2. The Problems of Logical Omniscience and Impossible Worlds

Let Ω be a non-empty space of possible worlds. I want to remain as neutral as possible as to what worlds *are*; what's important is that they are the kinds of things of which it makes sense to speak of the truth or falsity of a sentence at a world. In calling Ω a set of *possible* worlds, I'm specifically making the following assumptions about every $\omega \in \Omega$ and $S, S_1, S_2, \ldots \in \mathcal{L}$:

Non-Contradiction
At most one of S or $\neg S$ is true at ω

Maximal Specificity
At least one of S or $\neg S$ is true at ω

Closure under Implication
If $S_1, S_2, \ldots$ are true at ω and jointly imply S, then S is true at ω

What happens at these worlds with respect to sentences *not* in $\mathcal{L}$ won't be important, and in the sequel it should be assumed that the sentences S, S^*, etc., that I quantify over are always members of $\mathcal{L}$. I'll also assume that the relevant notion of implication (here and throughout) is

[3] See, e.g., (Bjerring 2013), where ∧ and ∨ are defined in terms of ¬ and →; or (Jago 2012), where ∨ is defined in terms of ¬ and ∧ (themselves defined in terms of mereological operations on 'facts').

at least as strong as that of classical propositional logic. If need be, we can throw some conceptual or analytic implications in there as well, so as to rule out worlds with, e.g., married bachelors, male vixens, four-sided triangles, and the like.

Call any set of (possible and/or impossible) worlds a *proposition*. The powerset of Ω, $\wp(\Omega)$, contains each of the propositions which can be formed from out of the worlds in Ω. Every sentence S can be mapped to some (perhaps empty) proposition $P_S \in \wp(\Omega)$, the set of all worlds in Ω where what S says is true.[4] Given the three assumptions about Ω above, logically equivalent sentences will always be mapped to the same propositions. Moreover, ¬, ∧, and ∨ will correspond to basic set operations of complementation, intersection, and union respectively, in the following way:[5]

(i) $P_{\neg S} = P_S{}^{C}$
(ii) $P_{S \wedge S^*} = P_S \cap P_{S^*}$
(iii) $P_{S \vee S^*} = P_S \cup P_{S^*}$

Where Ω is populated richly enough, it's reasonable to think that many if not all of the propositions in $\wp(\Omega)$ either *are*, or otherwise directly *correspond to*, genuine objects of thought.

With this in place, we might make a start on modelling some doxastic states. Begin with a model that originates with (Hintikka 1962), which focuses solely on full belief. Let α be our subject. Exactly one member of $\wp(\Omega)$, which we'll label P^α, is supposed to represent the way the world must be given all of α's beliefs. The worlds in P^α will be called α's *doxastically possible worlds*. To obtain a complete representation of what α believes, we might on first thought suppose that any proposition in $\wp(\Omega)$ of which P^α is a subset represents some specific content that α believes; so α believes that S if and only if $P^\alpha \subseteq P_S$. The upshot is a neat and relatively compact model of a *total belief state*: fix the relevant space of worlds Ω and the set P^α, and the rest of the work is done automatically by the subset relation.

But that's a little too quick. For all I've said, it may well be the case that $\wp(\Omega)$ contains many propositions that correspond to no proper object of belief. Modelling the objects of thought as sets of worlds does not commit one to assuming that every set of worlds models an object of belief, and it shouldn't be taken for granted that every way the world might (not) be corresponds to something that α might believe.[6] So let's make a very minor adjustment: suppose that $\mathcal{B} \subseteq \wp(\Omega)$ contains those propositions which represent genuine objects of thought, and say now that α believes that S if and only if $P^\alpha \subseteq P_S$ *and* $P_S \in \mathcal{B}$.

Some work will be needed to characterise exactly what propositions get into $\mathcal{B}$, but here's a start. Say that a proposition P is *expressible* (in $\mathcal{L}$) just in case there is a sentence S in $\mathcal{L}$ such

[4] In general, I will use P_S to designate the set of worlds (within some space of worlds Ω or Ω^+ as determined by context) at which S is true. For later discussion, it will be helpful to distinguish between arbitrary sets of worlds and the sentences that they (sometimes) correspond to—hence the relatively cumbersome notation.

[5] That logically equivalent sentences map to the same proposition follows from Closure under Implication. Proof of (i): Non-Contradiction implies that $P_S \cap P_{\neg S} = \emptyset$, and Maximal Specificity implies that $P_S \cup P_{\neg S} = \Omega$. So $P_{\neg S} = \Omega \backslash P_S$. Proof of (ii): Closure under Implication implies that if $S \wedge S^*$ is true at ω, then S and S^* are true at ω. Hence $P_{S \wedge S^*} \subseteq P_S \cap P_{S^*}$. In the other direction, Closure under Implication also implies that if S, S^* are true at ω, then $S \wedge S^*$ is true at ω. So $P_S \cap P_{S^*} \subseteq P_{S \wedge S^*}$. Proof of (iii): Closure under Implication implies that if either of S or S^* are true at ω, then $S \vee S^*$ is too. So $P_S \cup P_{S^*} \subseteq P_{S \vee S^*}$. In the other direction, if $S \vee S^*$ is true at ω, then by Closure under Implication, $\neg(\neg S \wedge \neg S^*)$ is true at ω. By Maximal Specificity, for any $\omega \in P_{S \vee S^*}$, either S is true at ω or $\neg S$ is. If S is true at ω, then $\omega \in P_S$ ($\cup P_{S^*}$). If $\neg S$ is true at ω, then S^* must be true at ω (otherwise $\neg S$, $\neg S^*$, and so $\neg S \wedge \neg S^*$ would be true at ω, which would contradict Non-Contradiction). Hence, for any $\omega \in P_{S \vee S^*}$, ω is either in P_S or P_{S^*}; and in either case $P_{S \vee S^*} \subseteq P_S \cup P_{S^*}$.

[6] For instance, in response to the Russell-Kaplan paradox (see Davies 1981, p. 262; Kaplan 1995), Lewis (1986, pp. 104-7) argues that there are many more ways the world might be than there are possible functional roles, and hence more than there are possible belief contents (at least $\beth_3$ for the former, and probably no more than $\beth_0$ for the latter).

that $P_S = P$—that is, a proposition P is expressible just in case there's a sentence S which holds at all and only the worlds in P. In §1 it was assumed that for every belief α has, $\mathcal{L}$ includes a sentence S which expresses exactly that belief. So, if you think this assumption is reasonable, then it's only natural to suppose that a proposition should be found in $\mathcal{B}$ only if it is expressible in $\mathcal{L}$. After all, what could it mean to represent α as believing a proposition P, where P is not characterised by any sentence in a language which, *ex hypothesi*, is capable of expressing every one of α's beliefs?

Now it's well known that this basic model suffers from a cluster of issues that usually come under the heading of the *problems of logical omniscience*, each of which is due in part to the basic assumptions we've made about what kinds of worlds are in Ω. Let me highlight four illustrative examples, which hold for all S, S*:

(i) If S implies S* and P_S, $P_{S*} \in \mathcal{B}$, then α believes S only if she also believes S*
(ii) If S and S* imply each other, then α believes S if and only if she also believes S*
(iii) If S is a tautology and $P_S \in \mathcal{B}$, then α believes S
(iv) α's beliefs are inconsistent only if $P^\alpha = \emptyset$ (in which case α believes everything in $\mathcal{B}$)

The first problem is a result of Closure under Implication, which ensures that if S implies S*, then $P_S \subseteq P_{S*}$. Corollary: if S and S* are logically equivalent, then $P_S = P_{S*}$; this generates the second problem. Maximal Specificity and Closure under Implication together imply that if S is a tautology, then $P_S = \Omega$. Since Ω is a superset of any proposition in $\mathcal{B}$, this gives rise to our third problem. With the addition of Non-Contradiction we also get that if S is a contradiction, then $P_S = \emptyset$, which ultimately leads to the fourth problem. Indeed, Non-Contradiction alone says that P_S and $P_{\neg S}$ are disjoint, so α can believe both S and $\neg S$ only if $P^\alpha = \emptyset$.

There are a number of responses to logical omniscience that we might opt for here. Perhaps the error is in thinking that we can adequately model belief sets using unstructured sets of possible worlds and subset relations. Or, perhaps it is in thinking that we can use a *single* set of worlds to encode an agent's total doxastic state, which may be better represented using multiple 'fragments'. Or perhaps there isn't really a problem here after all, our beliefs really *are* closed under classical logic and it is only the complexities of belief attribution and our imperfect access to our own beliefs which makes it seem otherwise. I think that each of these captures part of the truth, but my intention for this paper is not to provide a solution to the problems of logical omniscience. Instead, I wish to focus on one common response which begins with the thought that perhaps there are *not enough* propositions in $\wp(\Omega)$: we need to make our space of worlds bigger, to accommodate more fine-grained divisions amongst the objects of thought.

Suppose we make an extension to Ω, call it Ω^+, such that Ω^+ contains not only the original set of possible worlds, but also worlds where various kinds of impossible affairs obtain.[7] To make Ω^+ rich enough, we will want worlds which are blatantly inconsistent (where both S and $\neg S$ are true), as well as worlds which are not closed under classical consequence. Indeed, we will plausibly need to ensure that our worlds are not closed under any non-trivial consequence relation. It would not be very helpful to remove closure under classical consequence but retain closure under, e.g., intuitionistic consequence—otherwise, we're just swapping one sort of logical omniscience for another.

[7] The use of impossible worlds to help solve the problem of logical omniscience and related problems in epistemic logic was explicitly introduced in (Rantala 1982), although the idea can also be in (Hintikka 1975) and (Creswell 1973). Numerous authors have since made use of the idea, and a recent defence can be found in a series of works by Mark Jago (2009; 2013; 2014a; 2015a; 2015b) and Francesco Berto (2010). See also Nolan (1997; 2013), though Nolan's general focus is on using impossible worlds to give a Lewisian semantics for counterpossible conditionals.

To really free up the model, proponents of impossible worlds will typically posit an *unrestricted comprehension* principle, along the following lines:[8]

> **Unrestricted Comprehension**
> For any maximal set of sentences $\mathcal{S} \subseteq \mathcal{L}$, there will be worlds in Ω^+ where every $S \in \mathcal{S}$ is true and no $S \in \mathcal{L} \backslash \mathcal{S}$ is true

Now Ω^+ contains every logically possible world, plus every kind of maximal impossible world. (We can hold onto the Maximal Specificity assumption, though some proponents of impossible worlds choose to drop even this condition to allow for *incomplete* worlds as well (e.g., Jago 2012; 2014a; 2014b). Whether we include incomplete worlds in Ω^+ won't make a difference to my arguments in this section—if what I say applies to Ω^+ as characterised by Unrestricted Comprehension, then it will apply to any superset of Ω^+ as well.)

By building our model around this expanded space of worlds, we can easily block all four of the unwelcome 'omniscience' results noted earlier. Indeed, we can say more than this. Let $\{S^*, S^{**}, \ldots\}$ be any consistent or inconsistent subset of $\mathcal{L}$, and let P^α be the intersection of P_{S^*}, $P_{S^{**}}$, …. Now P^α will be non-empty (there will be worlds where all of S^*, S^{**}, … are true), and for any S that's not in $\{S^*, S^{**}, \ldots\}$, there will be maximally specific worlds in P^α where S is not true. So regardless of what we take α's beliefs set of beliefs $\{S^*, S^{**}, \ldots\}$ to be, we will be able to find some P^α such that $P^\alpha \subseteq P_S$ if and only if α believes S. That looks like a nice property for our model to have, and all we had to do was load Ω^+ up with enough impossible worlds.

But note a consequence of Unrestricted Comprehension: there is no sentence S—at least, no sentence in $\mathcal{L}$—such that S is true at all and only the worlds in P^α. The set of expressible propositions $\{P_S: S \in \mathcal{L}\}$ is an antichain of the partially ordered set $\langle \wp(\Omega^+), \subseteq \rangle$: for *any* two distinct sentences S, S^*, there will be worlds in Ω^+ where S is true and S^* isn't true; so, P_S will *never* be a subset of P_{S^*}. Suppose that α believes that S. Now whatever P^α ends up being, it will have to be a subset of P_S. Hence, there's no S such that $P_S = P^\alpha$. So P^α is *inexpressible* (in $\mathcal{L}$).[9]

I'm inclined to think that the inexpressibility of P^α is not by itself especially problematic. It would be immediately problematic if we were to assume that P^α (whatever it is) ought to itself represent something that α believes, and hence that it should always be included within $\mathcal{B}$. However, nothing internal to the model I've described suggests that this ought to be the case. That P^α should itself be a proposition that α believes was never a commitment of the model, even when we were working with just possible worlds. What's needed for the representational system to work is that (a) if $\mathcal{P}_\alpha \subseteq \wp(\Omega^+)$ is the set of all and only the propositions towards which some agent α has beliefs, then $\mathcal{P}_\alpha$ has some lower bound with respect to $\subseteq$ which we can designate as P^α; and (b) if $\mathcal{P}_\alpha \neq \mathcal{P}_\beta$, then $P^\alpha \neq P^\beta$. That is, every distinct total belief state can be uniquely represented by (at least one) set of doxastically possible worlds. We can satisfy this by letting P^α be the intersection of each proposition that α believes.

(None of this is to say that the impossible worlds model of belief just developed is without problems—just that it is not committed to saying that α believes something she cannot possibly believe. It is worth noting that if we can only believe expressible propositions, and

[8] I am here referring to proponents of the so-called "American stance" on impossible worlds; my arguments are not intended to touch upon the "Australasian" use of impossible worlds as the basis for an interpretation of some non-classical logic. See, e.g., (Nolan 1997), (Jago 2012), (Berto 2013, §4.1). By 'maximal set of sentences' $\mathcal{S} \subseteq \mathcal{L}$, I mean a set such that for any $S \in \mathcal{L}$, at least one of S or $\neg S$ is in $\mathcal{S}$.

[9] Note that P^α can be inexpressible even if we have a name a_i for each of the worlds within P^α and $\mathcal{L}$ contains a way of saying "The actual world is a_1 or a_2 or …" (or something to that effect). Assuming that such a sentence exists in $\mathcal{L}$, if an unrestricted comprehension principle holds then the sentence will be true at *some* of the worlds in P^α, but it will also be false at some of those worlds (and true at some worlds outside of P^α).

no expressible proposition is a subset of any other expressible proposition, then there is a genuine question as to the *point* of using this kind of set-theoretic model to represent our beliefs in the first place.[10] The machinery of set theory only comes into play at a single step, linking the (non-believed) proposition P^{α} to the set of expressible propositions that α believes, the latter of which has no interesting set-theoretic structure. The only thing which unites the worlds in the proposition P^{α} is that they are those worlds where each member of a set of sentences S, S', S'', … is true; and characterising that proposition amounts to just listing out all and only those sentences which express something α believes. What we've done with P^{α} and $\subseteq$, we could have done more perspicuously with a simple list; we gain nothing in economy by the addition of P^{α}.)

So much for full belief. If you're like me and you think that beliefs generally come in degrees (so that full belief is ultimately just a species of partial belief), then you will likely want your model of α's doxastic states *in general* to represent all of her partial beliefs, not just those that qualify as full beliefs. Luckily enough, there are natural ways to generalise the basic model outlined in the previous section. As Lewis' puts it,

> [W]e must also provide for partial belief. Being a [doxastically possible world] is not an all or nothing matter, rather it must admit of degree. The simplest picture, idealised to be sure, replaces the sharp-edged class of [doxastically possible worlds] by a subjective probability distribution. … We can say that a [doxastically possible world] simpliciter is a possible [world which] gets a non-zero (though perhaps infinitesimal) share of probability, but the non-zero shares are not all equal. (1986, p. 30)

In the rest of this paper, I want to focus on these probabilistic models, and the problems of inexpressibility that come with them.

3. The Problem of Probabilistic Coherence

For the sake of concreteness, here's one way generalise the full belief model to partial beliefs, along the lines suggested by Lewis.[11] Suppose again that Ω is a non-empty space of possible worlds. A subjective probability distribution over Ω could be understood as a function $\mathcal{D}$ which assigns 0 to all but countably many ω in Ω, and a real value between 0 and 1 to the rest of the worlds such that those values sum to unity. We might interpret $\mathcal{D}$ as representation of α's degree of belief that the actual world is ω, for each ω in Ω.[12]

We can now use $\mathcal{D}$ to induce a function $\mathcal{C}r$ on $\boldsymbol{\mathcal{B}}$, by stipulating that for each $P \in \boldsymbol{\mathcal{B}}$,

$$\mathcal{C}r(P) = \sum_{\omega \in P} \mathcal{D}(\omega)$$

Independent of any assumptions we make about what worlds get into Ω, $\mathcal{C}r$ will satisfy:

Nonnegativity
If $\emptyset$ is in $\boldsymbol{\mathcal{B}}$, then $\mathcal{C}r(\emptyset) = 0$

Normalisation
If Ω is in $\boldsymbol{\mathcal{B}}$, then $\mathcal{C}r(\Omega) = 1$

[10] My complaint in this paragraph parallels one made by Bjerring and Schwarz in their (2017, §3).

[11] My choice to treat the probability distribution $\mathcal{D}$ as discrete and real-valued probability mass function is primarily for expositional simplicity, and inessential to my main argument, which depends more on the assumption of Booleanism than on the details of any particular model of partial belief which adopts the Boolean assumption. See §6 for further discussion.

[12] If it's implausible that any agent assigns a positive degree of belief to (singleton sets of) worlds, we could let $\mathcal{D}$ be defined over an arbitrarily coarse-grained partition of Ω instead, and interpret $\mathcal{D}$ accordingly. Alternatively, we could drop the assumption that $\mathcal{D}$ is discrete (although this would also require some re-interpretation), and treat it more along the lines of a continuous probability distribution.

Monotonicity
For all pairs P, P^* in $\mathcal{B}$, if $P \subseteq P^*$, then $\mathcal{C}r(P) \leq \mathcal{C}r(P^*)$

Σ-Additivity
If $\mathcal{B}$ is any countable set of disjoint propositions in $\mathcal{B}$ whose union ($\bigcup \mathcal{P}$) is also in $\mathcal{B}$, then $\mathcal{C}r(\bigcup \mathcal{P}) = \sum_{P \in \mathcal{P}} \mathcal{C}r(P)$

If $\mathcal{B}$ contains $\emptyset$, then $\mathcal{C}r$ is a measure on $\mathcal{B}$. If we want $\mathcal{C}r$ to be a probability function as usually understood, then we need to make the stronger assumption that $\mathcal{B}$ constitutes some Boolean sub-algebra of $\wp(\Omega)$. That is, we need to assume something at least as strong as Booleanism:

Booleanism
For all $P \in \wp(\Omega)$,
(i) If $P \in \mathcal{B}$, then $P^C \in \mathcal{B}$
(ii) If $P, P^* \in \mathcal{B}$, then $P \cup P^* \in \mathcal{B}$
(iii) If $P, P^* \in \mathcal{B}$, then $P \cap P^* \in \mathcal{B}$

We'll follow suit and assume for now that $\mathcal{B}$ satisfies these conditions.

Where $\mathcal{B}$ includes all and only those propositions towards which α has partial beliefs, a very natural way to read $\mathcal{C}r$ is as a representation of α's *total degree of belief state*: α believes that S to degree x if and only if $\mathcal{C}r(P_S) = x$. This generalises our model of full belief nicely. In the simplest case, say full belief equates to degree of belief 1. Then, we will be able to characterise P^α as just those worlds assigned some positive value by $\mathcal{D}$, and $\mathcal{C}r(P_S) = 1$ for every $P_S \supseteq P^\alpha$ that finds its way into $\mathcal{B}$. But our new model also lets us represent each of the many grades of belief that α can have towards any proposition in $\mathcal{B}$, removing the "sharp edges" between belief and non-belief.

However, because we've built $\mathcal{D}$ (and $\mathcal{C}r$) on top of a space of *possible* worlds Ω, it's easy to see that our model of α's partial beliefs will have its very own issues with logical omniscience. Indeed, the new model takes the idealisation a step further, as $\mathcal{C}r$ will impose *probabilistically coherence* upon α's degrees of belief more generally. For all $S, S^* \in \mathcal{L}$,

(i) If S is a contradiction then $P_S \in \mathcal{B}$, and $\mathcal{C}r(P_S) = 0$
(ii) If S is a tautology then $P_S \in \mathcal{B}$, and $\mathcal{C}r(P_S) = 1$
(iii) If S implies S^* and $P_S, P_{S^*} \in \mathcal{B}$, then $\mathcal{C}r(P_S) \leq \mathcal{C}r(P_{S^*})$
(iv) If $P_S, P_{S^*} \in \mathcal{B}$, then $1 = \mathcal{C}r(P_S) + \mathcal{C}r(P_{S^*}) + \mathcal{C}r(P_{S \vee S^*}) - \mathcal{C}r(P_{S \wedge S^*})$

Additionally, if full belief is degree of belief 1, the new model implies that it's not even possible for α to have inconsistent beliefs—the distribution $\mathcal{D}$ will assign a positive value to at least one possible world ω, and the set of propositions P such that $\mathcal{C}r(P) = 1$ will be consistent (or empty). (On an alternative account, full belief might be characterised in terms of exceeding some threshold degree T, for $T < 1$. In that case, there may be no P^α such that α fully believes P if and only if $P^\alpha \subseteq P$, and it may be possible for α's beliefs to be inconsistent. However, if $T > 0.5$, then it will be impossible for α to believe S and $\neg S$ simultaneously; and as long as $T > 0$, it will be impossible for α to believe any contradictions.)

If logical omniscience is bad, then probabilistic coherence seems much worse. But never fear—impossible worlds to the rescue! If we define $\mathcal{D}$ on a space of worlds Ω^+ that satisfies Unrestricted Comprehension, then $\mathcal{C}r$ need not satisfy any of the coherence conditions (i) through (iv). Indeed $\mathcal{C}r$ can be *almost* as wild and wacky as we want it to be. A simple example: suppose that $\mathcal{D}$ assigns a positive value only to worlds where S and $S \wedge \neg S$ are both true, and neither $\neg S$ nor $S \vee \neg S$ is true. Now $\mathcal{C}r(S) = \mathcal{C}r(S \wedge \neg S) = 1$, and $\mathcal{C}r(\neg S) = \mathcal{C}r(S \vee \neg S) = 0$.

Proviso: if Maximal Specificity holds and P_S, $P_{\neg S} \in \mathcal{B}$, then $\mathcal{C}r(P_S) + \mathcal{C}r(P_{\neg S}) \geq 1$.[13] So Unrestricted Comprehension by itself does not give us complete freedom to let $\mathcal{C}r$ assign values to expressible propositions however we might like. But we can fix this by expanding Ω^+ even further to allow for non-maximal, or *incomplete*, worlds:

> **Really Unrestricted Comprehension**
> For any set of sentences $\mathcal{S} \subseteq \mathcal{L}$, there will be worlds in Ω^+ where every $S \in \mathcal{S}$ is true and no $S \in \mathcal{L}\backslash\mathcal{S}$ is true

Now if you want $\mathcal{C}r$ to assign 0 to both P_S and $P_{\neg S}$, you just need to make sure that $\mathcal{D}$ assigns positive values only to worlds where neither S nor $\neg S$ is true. More generally, for pretty much *any* way you want $\mathcal{C}r$ to distribute values across the expressible propositions in any set $\mathcal{B}$, we'll be able to find a $\mathcal{D}$ which generates exactly that distribution.[14]

The idea to use a probability function over a space of possible and impossible worlds as a way of modelling the probabilistic incoherence of non-ideal agents is common in conversation, but also shows up at several points in the literature. Mikaël Cozic (2006) has recently advocated the strategy, and Halpern and Pucella (2011, §4) make similar points. In his (1997) and (1999), Lipman attempts to deal with logical non-omniscience by deriving a probabilistic expected utility representation from an agent's preferences, where the probability function in question is defined over a state-space involving both possibilities and impossibilities. Easwaran (2014, esp. pp. 1-2, 29) also suggests using impossible worlds in our probabilistic models of agents' doxastic states, albeit in a slightly different context.

4. Booleanism with Impossible Worlds

In this section, I will argue that if Ω^+ satisfies a very weak (and very plausible) richness assumption, then either Booleanism is false, or our model won't allow us to adequately represent logically non-ideal agents—which, of course, is the central motivation for introducing impossible worlds into the model in the first place. My argument is premised on the claim that whatever $\mathcal{B}$ is, it should contain only propositions which are expressible in $\mathcal{L}$.

For any S, take the set of all worlds where S is true, and consider its complement $P_S{}^C$. If Unrestricted Comprehension holds, then there is no S^* such that $P_{S^*} = P_S{}^C$, as for any pair of sentences S and S^* there will be worlds where both S and S^* are true. If Really Unrestricted Comprehension also holds, then there will be also be worlds where neither S nor S^* is true. In either case, P_S and P_{S^*} are not complements of one another. Hence, if P_S is expressible, then $P_S{}^C$ is *in*expressible. And since we've assumed that $\mathcal{B}$ is closed under complementation, it follows that there must be at least as many inexpressible propositions in $\mathcal{C}r$'s domain as there are expressible propositions. And that's not a nice result: $\mathcal{L}$ is supposed to include a sentence capable of expressing every object of thought towards which we might have partial beliefs, and yet the model we've now developed is assigning nonsensical values to propositions expressed by no sentences of $\mathcal{L}$.

[13] Maximal Specificity says that $P_S \cup P_{\neg S} = \Omega$. Normalisation plus Σ-Additivity then imply that $\mathcal{C}r(P_S - P_{\neg S}) + \mathcal{C}r(P_{\neg S} - P_S) + \mathcal{C}r(P_S \cap P_{\neg S}) = 1$. Since $\mathcal{C}r(P_S \cap P_{\neg S}) \geq 0$, $\mathcal{C}r(P_S) \geq \mathcal{C}r(P_S - P_{\neg S})$ and $\mathcal{C}r(P_{\neg S}) \geq \mathcal{C}r(P_{\neg S} - P_S)$, it follows that $\mathcal{C}r(P_S) + \mathcal{C}r(P_{\neg S}) \geq 1$.

[14] A quick example to demonstrate the point. Let S, S^* and S^{**} be any three distinct sentences whatsoever. Suppose we want a $\mathcal{C}r$ that assigns x to P_S, y to P_{S^*}, z to $P_{S^{**}}$ (where $1 \geq x > y > z \geq 0$), and 0 to everything else. We let $\mathcal{D}$ be as follows. Where ω^{**} is the world where only S, S^* and S^{**} are true, $(\omega^{**}) = z$. Where ω^* is the world where only S and S^* are true, $(\omega^*) = y - z$. Where ω is the world where just S is true, $(\omega) = x - y$. The 'empty world' (where no sentences are true) is then assigned $1 - x$, and every other world in Ω assigned 0. It follows that $\mathcal{C}r(P_S) = x$, $\mathcal{C}r(P_{S^*}) = y$, and $\mathcal{C}r(P_{S^{**}}) = z$. Give my assumptions about $\mathcal{D}$, the same basic trick can be extended for any $\mathcal{C}r$ that assigns a positive value to at most countably many expressible propositions.

We could get around the foregoing argument if (and only if) we impose the following restrictions on the worlds in Ω^+:

> **Restriction R1**
> For every S such that $P_S \in \mathcal{B}$, there is an S^* such that exactly one of S or S^* is true at ω

I'll have more to say about R1 in a moment, but first, note that merely imposing R1 on Ω^+ won't solve all our problems.

We've also supposed that $\mathcal{B}$ is closed under (at least finite) intersections and unions, and with no further restrictions on Ω^+ the set of expressible propositions (in $\mathcal{B}$) will still be an antichain of $<\wp(\Omega^+), \subseteq>$. (The only difference from before is that it's now closed under complementation.) So take any two sentences S and S^* such that $P_S \neq P_{S^*}$: there is then no S^{**} such that $P_{S^{**}} = P_S \cap P_{S^*}$. After all, nothing about R1 implies that there must be any sentences in $\mathcal{L}$ which are true at a world if and only if two other sentences are true at that world. Likewise, there is no S^{**} such that $P_{S^{**}} = P_S \cup P_{S^*}$. Consequence: even with R1 in place, there will *still* be at least as many inexpressible propositions in $\mathcal{C}r$'s domain as there are expressible propositions.

The following is necessary and sufficient for ensuring that the intersection of any two expressible propositions (in $\mathcal{B}$) is itself expressible:

> **Restriction R2**
> For every pair S, S^* such that P_S, $P_{S^*} \in \mathcal{B}$, there is an S^{**} such that both S and S^* are true at ω if and only if S^{**} is true at ω

Given R1, R2 also implies that the union of any two expressible propositions (in $\mathcal{B}$) is expressible—that is, for any pair of expressible propositions P_S, P_{S^*} in $\mathcal{B}$, there is some sentence S^{**} such that at least one of S or S^* are true at ω if and only if S^{**} is.[15]

Exactly how restrictive R1 and R2 end up being depends on which expressible propositions should be included in $\mathcal{B}$. We can safely assume that whatever $\mathcal{B}$ is, it will be richly populated with plenty of expressible propositions. So R1 and R2 are never trivially satisfied. On the other hand, if there are sentences whose characteristic propositions are not in $\mathcal{B}$, then R1 and R2 are consistent with a degree of freedom in relation to those propositions. But this is not especially interesting: since $\mathcal{B}$ contains all of the propositions in $\mathcal{C}r$'s domain, whatever is true of the expressible propositions *not* in $\mathcal{B}$ will be irrelevant to the representation of α's degrees of belief that we are left with. Hence, I'll simplify the following discussion and pretend that for every sentence S of $\mathcal{L}$, $P_S \in \mathcal{B}$.

The key point in what follows will be that how these two restrictions are implemented is constrained by what kinds of worlds we want to *keep* in Ω^+. For example, if R1 holds, then at any *possible* world $\omega \in \Omega^+$, the required sentence S^* can be true at ω if and only if $\neg S$ is true. So if we were to require that every world satisfying Maximal Specificity, Non-Contradiction, and Closure under Implication remains in Ω^+, then S^* must be *at least* logically equivalent to $\neg S$. Indeed, if we wanted Ω^+ to contain every world that's maximally specific and non-contradictory, then S^* *must* be $\neg S$—there are no other sentences which consistently hold at every maximally specific and non-contradictory world where S isn't true.

I will not assume that Ω^+ contains every possible world, though I think that something in the vicinity must be true if we hope to use $\mathcal{C}r$ to model arbitrary *ideal* agents as well as non-ideal agents. Instead, I will assume something much weaker. Say that a sentence S is *blatantly*

[15] R1 says that the complement of every expressible proposition in $\mathcal{B}$ is expressible. R2 says that the intersection of every pair of expressible propositions in $\mathcal{B}$ is expressible. Suppose P_S and P_{S^*} are two expressible propositions in $\mathcal{B}$. So, $P_S{}^C$ and $P_{S^*}{}^C$ are both expressible, and both are in $\mathcal{B}$. Thus, $P_S{}^C \cap P_{S^*}{}^C$ is an expressible proposition in $\mathcal{B}$, and so is $(P_S{}^C \cap P_{S^*}{}^C)^C = P_S \cup P_{S^*}$.

inconsistent with another sentence S^* just in case either $S = \neg S^*$ or $S^* = \neg S$. Then my assumption can be put as follows:

> **Minimal Richness**
> For any consistent triple S_1, S_2, S_3, there is at least one world $\omega \in \Omega^+$ such that:
> (i) S_1, S_2, and S_3 are all true at ω, and
> (ii) If S_4 is blatantly inconsistent with any of S_1, S_2, or S_3, then S_4 is not true at ω

Minimal Richness should be uncontroversial, *especially* since it can be motivated by precisely the same considerations which led us to insert a rich space of impossible worlds into our models in the first place. If S_1, S_2, and S_3 are jointly consistent in classical propositional logic, then it is surely *possible* to fully believe each, and on the kinds of models we've been considering this is only possible if there is a world in Ω^+ where each is true. Furthermore, it's surely possible to fully believe a consistent triple S_1, S_2, and S_3, while fully *dis*believing any S_4 that's blatantly inconsistent with S_1, S_2, or S_3—even ordinary agents can be a little bit rational, sometimes!

So let's consider R1, which states that every S can be paired with another sentence S^* which is true at a world if and only if S is not true. If Minimal Richness is true, then whatever S^* ends up being, it must be logically equivalent to $\neg S$. For suppose that S^* is not logically equivalent to $\neg S$. Then either S^* does not imply $\neg S$, or $\neg S$ does not imply S^* (or both). If S^* does not imply $\neg S$, then $\{S^*, S\}$ (and so $\{S^*, S^*, S\}$) is consistent, and there will be at least one world where S^* and S are both true, which contradicts R1. On the other hand, if $\neg S$ does not imply S^*, then $\{\neg S, \neg S^*\}$ is consistent and there will be worlds where $\neg S$ and $\neg S^*$ are both true. Since S and S^* are blatantly inconsistent with $\neg S$ and $\neg S^*$ respectively, this would have to be a world where neither S nor S^* is true, which also contradicts R1. Hence, any sentence S^* that satisfies R1 must be logically equivalent to $\neg S$, if Minimal Richness is true.

One very straightforward way to implement R1 would be to let the required sentence S^* just *be* $\neg S$. In effect, this is just to assume that every world in Ω^+ must satisfy Non-Contradiction and Maximal Specificity. And it's easy enough to think of some plausible motivations for assuming Non-Contradiction: one *could* argue that no model of a minimally rational agent's doxastic state should represent her as having any degree of belief that both S and $\neg S$ could be true simultaneously (cf. Lewis 2004; Jago 2014b; Bjerring 2013). To the extent that we make errors of logical reasoning, they tend to be more subtle—e.g., a failure to deduce a downstream consequence of what we believe, rather than blatant inconsistencies.

Motivating Maximal Specificity is a little more difficult, as it amounts to removing the incomplete worlds from Ω^+. Some are independently happy to do this (e.g., Bjerring 2014; Bjerring and Schwarz 2017, p. 28; cf. Stalnaker 1996); for others, incomplete worlds are a crucial aspect of the model (Jago 2014a; 2014b). Furthermore, it'll be a consequence of assuming Non-Contradiction and Maximal Specificity together that the worlds we are left with are closed under the rules of double negation introduction and elimination, and $\mathcal{C}r$ satisfies $\mathcal{C}r(P_S) = \mathcal{C}r(P_{\neg\neg S})$ for all P_S in $\boldsymbol{\mathcal{B}}$. This is already quite a strong restriction, and it's probably not satisfied by all thinkers.

Nevertheless, there are good reasons to think that if the imposition of R1 is to be even remotely well-motivated, then S^* certainly shouldn't be anything *other* than $\neg S$. Suppose that S^* is any sentence that's logically equivalent to $\neg S$ other than $\neg S$ itself—say, $\neg\neg\neg S$. We might then preserve the presence of some non-maximally specific and/or contradictory worlds in Ω^+, but now our worlds will be closed under the rules of *sextuple negation introduction* and *elimination*: S is true at ω if and only if $\neg\neg\neg\neg\neg\neg S$ is true at ω. Any reasons we might have had to avoid closing worlds under the (relatively simple) rules of double negation would apply with all the more force here: to the extent that ordinary agents might generally accept something like sextuple negation introduction and elimination, it's *because* they accept that S is true if

and only if $\neg S$ is not true. Given Minimal Richness, the very best case we can make for implementing R1 involves letting S^* be $\neg S$.

But it is in combination with R2 that R1 most worrisome. R2 states that every pair of sentences S, S^* can be paired with a third sentence S^{**} such that S^{**} is true at a world if and only if both S and S^* are true at that world. Given Minimal Richness, we know that S^{**} must be logically equivalent to $S \wedge S^*$. The argument here is similar to the one we just earlier with R1. Suppose that S^{**} is not logically equivalent to $S \wedge S^*$. Then S^{**} does not imply $S \wedge S^*$, or $S \wedge S^*$ does not imply S^*. If S^{**} does not imply $S \wedge S^*$, then at least one of the following sets of sentences is consistent: $\{S^{**}, \neg S, \neg S^*\}$, $\{S^{**}, \neg S, S^*\}$, $\{S^{**}, S, \neg S^*\}$. In each case, there will be at least one world in Ω^+ where S^{**} is true and at least one of S or S^* is not true, which would contradict R2. If $S \wedge S^*$ does not imply S^{**}, then S and S^* do not jointly imply S^{**}, so $\{S, S^*, \neg S^{**}\}$ is consistent and there is at least one world in Ω^+ where S and S^* are both true and S^{**} is not. This would also contradict R2. So, S^{**} must be logically equivalent to $S \wedge S^*$.

An argument analogous to that given for R1 then immediately suggests how we ought to implement the restriction, if at all: require that all worlds in Ω^+ satisfy ∧-Consistency:

> **∧-Consistency**
> For all S, $S^* \in \mathcal{L}$, S and S^* are both true at ω if and only if $S \wedge S^*$ is true at ω

Certainly, it would be absurd to suppose that R2 is not satisfied by $S \wedge S^*$, but rather some other sentence equivalent to $S \wedge S^*$. For suppose that R2 was satisfied by, say, $\neg(\neg S \vee \neg S^*)$. Then our models would end up representing an agent who, without fail, always infers back and forth between S, S^* and $\neg(\neg S \vee \neg S^*)$, while potentially skipping over the much more natural and direct inferences between S, S^* and $S \wedge S^*$. But anyone who doesn't reliably follow the rules of conjunction introduction and elimination is not going to be unfailingly adhere to inference rules which link S, S^* and $\neg(\neg S \vee \neg S^*)$ to one another. (To be sure, one *could* in principle describe a consequence relation such that the former inferences are admitted but the latter are not, but why would we think that closing the worlds in Ω^+ under that relation makes for a good model any doxastic agent, let alone an ordinary believer?)

In conjunction with Non-Contradiction and Maximal Specificity, ∧-Consistency guarantees that $\neg(\neg S \wedge \neg S^*)$ is true at any world where at least one of S or S^* are true: for any P_S and P_{S^*}, $P_S{}^C = P_{\neg S}$ and $P_S \cap P_{S^*} = P_{S \wedge S^*}$, hence $(P_S{}^C \cap P_{S^*}{}^C)^C = P_{\neg(\neg S \wedge \neg S^*)} = P_S \cup P_{S^*}$. Moreover, they imply that (i) every Boolean combination of expressible propositions will be expressible by some sentence involving $\neg$ and/or $\wedge$, and (ii) every world in Ω^+ will be closed under the $\{\neg, \wedge\}$ fragment of classical propositional logic. We're fast running out of impossibilities—and with them, our capacity to represent logically non-ideal subjects.

Now since we've not stipulated that other basic connectives should be defined in terms of $\neg$ and $\wedge$, nothing yet requires that the worlds in Ω^+ are closed under classical propositional logic *simpliciter*. For example, disjunctive sentences may still behave erratically for all we've said so far, with (e.g.) $S \vee S^*$ not being true at all and only the worlds where at least one of S or S^* is true. Likewise, if $\mathcal{L}$ contains a material conditional $\rightarrow$, where $S \rightarrow S^*$ is *not* simply defined as $\neg(S \wedge \neg S^*)$, then we've not said anything yet to guarantee that the worlds in Ω^+ must validate even basic inference rules for conditionals.

There may thus still be plenty of logically impossible worlds in Ω^+. Nevertheless, with Non-Contradiction and Maximal Specificity, ∧-Consistency alone we've managed to close the worlds in Ω^+ under a very strong consequence relation. Indeed, Ω^+ is already only apt for modelling agents who are very highly idealised: for *every* classically valid inference $S_1, S_2, \ldots \Rightarrow S$, the worlds in Ω^+ will be closed under an corresponding inference which replaces each of $S_1, S_2, \ldots$ and S with an equivalent sentence expressed using only $\neg$ and $\wedge$. For instance, while Ω^+ might not be closed under modus ponens, we do know that at every world where S and $\neg(S \wedge \neg S^*)$ are both true, so too will S^* be true. What we have, in effect, is a model that represents

an agent capable of infallibly performing extraordinarily complex inferences expressed using certain kinds of sentences. That the agent might *also* be represented as logically incompetent with respect to other basic inferences hardly makes her seem more realistic.

So it seems that giving up on Really Unrestricted Comprehension in order to accommodate Booleanism is the wrong tact. Minimal Richness looks to be on a firm footing, and if it holds then if $\mathcal{B}$ is going to satisfy these strong algebraic requirements, then the worlds in Ω^+ are going to have to be closed under at least some classically valid inference rules. We have a degree of choice as to what these rules might be (e.g., *double negation elimination* versus *sextuple negation elimination*), but closing under the most simple and natural rules (*viz*., those which ordinary agents are most likely to consistently follow) leads us directly into closing Ω^+ under a huge fragment of classical logic—and thus misrepresenting the logical capacities of ordinary, non-ideal subjects.

The impossible worlds theorist has two main options for response. First, she could go after the assumption that there exists an $\mathcal{L}$ such that everything α believes or partially believes is expressible in $\mathcal{L}$—if this is false, then the presence of inexpressible propositions in the domain of Cr is to be expected, not shunned. Perhaps we have just discovered that sometimes our partial beliefs towards expressible propositions comes hand-in-hand with partial beliefs towards inexpressible propositions; the latter are perfectly legitimate objects of thought, but not all such objects are expressible. I will discuss this possibility in the next section. Secondly, we could make alterations to the basic probabilistic model (or its interpretation) that was developed in §3; I will discuss this line of response in §6.

5. The Expressibility Assumption (Again)

There is surprisingly little philosophical discussion regarding whether every possible object of thought is linguistically expressible, though to the extent that the question has been discussed the general presumptive answer has been affirmative; e.g., (Searle 1969, pp. 19ff), (Katz 1978), (Schiffer 2003, p. 71), (Priest 2006, p. 54), and especially (Hofweber 2006; see also his 2016). Michael Dummett goes so far as to state *a priori* that:

> Thoughts differ in all else that is said to be among the contents of the mind in being wholly communicable: it is of the essence of thought that I can convey to you the very thought I have [...] It is of the essence of thought, not merely to be communicable, but to be communicable, without residue, by means of language. (1978, p. 142)

One would by no means be alone in presupposing the existence of $\mathcal{L}$.

Moreover, the existence of something much like $\mathcal{L}$ is strongly suggested by a wide variety of positions in philosophy. The assumption plays a role in important attempts to explain mental representation; see, e.g., (Field 1978), whose approach is premised on our ability to decompose attitudinal relations between thinkers and propositions into (a) relations between thinkers and sentences, and (b) between sentences and meanings. Likewise, the existence of a language like $\mathcal{L}$ is a background presupposition of inferential role semantics as cashed out by, e.g., (Boghossian 1993, pp. 73-4), whereby we assign contents to sentence-like symbol structures by taking causal patterns between them to mirror truth-preserving patterns of implication between propositions. In order for this kind of story to work, every thought that a particular believer might have had better be expressible by some sentence in some language or other. More generally, if one accepts the arguments for the existence of a Language of Thought as the psychological basis for our capacity to have propositional attitudes, then the existence of $\mathcal{L}$ seems hard to deny.

Likewise, the existence of a language rich enough to express each of our beliefs is implicit in several accounts of the nature of mental content itself. In Chalmers' epistemic two-dimensionalism, the contents of thoughts—including our partial beliefs—are modelled as

functions from 'scenarios' to extensions, with each 'scenario' being a complete description of an epistemically possible world in an idealised language consisting primarily of vocabulary for describing the microphysical and phenomenal characteristics of the world (see Chalmers 2011a; 2011b; 2012; see also Chalmers and Jackson 2001). And, as was noted in §1, Mark Jago argues for the existence of a richly expressive Lagadonian 'world-building' language for just the purpose of modelling hyperintensional belief contents as sets of possible and impossible worlds (where a 'world' in his framework is a set of Lagadonian sentences).

With that said, the recent literature has seen some purported counterexamples to my assumption about the expressibility of belief. James Shaw (2013) develops a variation on the Berry paradox to argue for the existence of a kind of inexpressible thought content—an instance of a case which he says "happens on extremely rare occasions due to a particular kind of linguistic technicality" (p. 70). Benj Hellie (2004) has also argued that there may be truths about phenomenal experience which we can appreciate but not express linguistically. And if one thinks that there is a one-to-one correspondence between ways the world might be and possible belief contents, then there are also classic expressive inadequacy arguments involving qualitatively indiscernible individuals and alien properties, to the effect that no language can describe every possibility (e.g., Lewis 1986, pp. 157ff; Bricker 1987).

I will not discuss any of these cases in any detail. Perhaps each case is a genuine problem for my assumption that $\mathcal{L}$ exists as characterised. But acquiescing on this point hardly seems to help with the problem currently at hand. The inexpressibility of most of $\mathcal{C}r$'s domain cannot be explained by an occasional linguistic technicality. And moreover, the inexpressible propositions that we have been describing are not plausibly *about* some ineffable aspect of our phenomenal experience, alien properties, or qualitatively indiscernible individuals. If $\mathcal{L}$ lacks the expressive power to represent our thoughts about such things—so be it. Let $\mathcal{L}$ represent a language capable of expressing only those more mundane beliefs which are expressible, like the belief that *there are dogs*. What kind of content could the set of worlds where 'There are dogs' is *not* true represent, if not that *there are no dogs*? Clearly, it has something to do with the existence of dogs—but what?

The point here is general. An adequate response to the argument of §4 can't be to just point out that there may be *some* possible things α *could* believe which are not expressible. The odd inexpressible object of thought here and there isn't too much cause for concern—but the underlying problem survives mere counterexamples to the existence of $\mathcal{L}$. Unless we make serious changes to the basic probabilistic model of our beliefs, so long as Booleanism and (Really) Unrestricted Comprehension are true, then if you have a degree of belief x towards P_S then you have a degree of belief $(1 - x)$ towards the inexpressible proposition $P_S{}^C$; and if you have degrees of belief x and y towards P_S and P_{S^*} then you'll have some degree of belief $z \leq x, y$ towards the inexpressible $P_S \cap P_{S^*}$ and $((x + y) - z)$ towards $P_S \cup P_{S^*}$. Inexpressibility on this model is not some esoteric phenomenon resting on a technicality, nor does it seem to be limited to a specific kind of topic (e.g., phenomenology, alien properties) about which we *might* have beliefs. We get to keep the model only if we're happy with the implication that thinkers in general have at least as many partial beliefs towards inexpressible propositions as they do towards expressible propositions. And that is a hard pill to swallow.

For similar reasons, I am not moved by simple cardinality arguments aimed at showing that we must accept the existence of inexpressible propositions, regardless of whether we adopt impossible worlds into our ontology or not. Some vigorously intuit that for any subset $\mathcal{S}$ of any language $\mathcal{L}$, α might (partially) believe that all and only the sentences of $\mathcal{S}$ are true. If $\mathcal{L}$ is set-sized, then the cardinality of the $\wp(\mathcal{L})$ is strictly greater than that of $\mathcal{L}$. It follows that $\mathcal{L}$ cannot contain a unique sentence S for each subset $\mathcal{S} \subseteq \mathcal{L}$ to the effect of 'All and only the elements of $\mathcal{S}$ are true'. Thus, either the content in question is not expressible at all, or it cannot be expressed in $\mathcal{L}$—either way, $\mathcal{L}$ is not up to the task of expressing everything that α might

believe. But even if the basic intuition which underlies this argument is correct—and it is by no means obvious that it is—the conclusion is merely that we must accept that we *might* have some inexpressible (partial) beliefs. What the argument doesn't do is give us any reason to think that the algebra of propositions $\mathcal{B}$ that constitutes what α actually has partial beliefs towards is filled to the brim with inexpressible propositions. Indeed, it's perfectly consistent with the argument's conclusion that $\mathcal{B}$ contains no inexpressible propositions at all.

Finally, although I have tried to present my argument in a way that is neutral with respect to different theories of what worlds are, there are ways of approaching the metaphysics of worlds which cannot avoid a version of my argument no matter how expressively inadequate any given language is. In particular, Daniel Nolan (1997) favours an approach where propositions are taken to be the fundamental entities from which worlds are constructed, rather than *vice versa*. On his picture, possible worlds are maximal consistent sets of propositions in the style of (Adams 1974), while impossible worlds are those sets of propositions which are inconsistent and/or non-maximal. Adopting this view, we could let $\mathcal{L}$ be the class of all propositions, trivialising the question as to whether $\mathcal{L}$ is 'expressively rich enough' to capture every belief α might have. We can then easily see that once something like Unrestricted Comprehension holds, there will be sets of worlds with no proposition in common amongst their members. These sets of worlds will not only be *linguistically* inexpressible, but quite literally unthinkable.[16]

6. Probabilities without Booleanism

Assuming that we want to avoid an excessive attribution of attitudes towards inexpressible propositions, then, our only other option looks to be an alteration of the probabilistic model outlined in §3. The obvious thing to change would be the assumption of Booleanism, which (in combination with the comprehension principles) leads directly to the problems with inexpressible propositions. I will say a few words about dropping Booleanism in a moment, but first I want to note some alterations to the model (or its interpretation) which I don't think will be very fruitful.

First of all, I think it would be a mistake to try to pin the blame on the fact that $\mathcal{C}r$ is a probability function, defined using a simple probability distribution $\mathcal{D}$ over Ω^+ (or an arbitrarily fine-grained partition thereof). To be sure, I have relied on the properties of these kinds of functions at several points in making my arguments; most notably, in motivating the use of impossible worlds in the model in the first place, and in presenting the reasons for assuming Minimal Richness. But if what I've said is taken to supply reasons to reject the use of probability distributions in modelling α's degrees of belief, then it gives us reason to reject plenty more besides.

For instance, Dubois and Prade's (1988) possibility theory allows us to systematically construct a degree of belief function on the basis of what they call a *possibility distribution*; i.e., a function $\mathcal{D}$ from a space of worlds into [0, 1] such that $\mathcal{D}(\omega) = 1$ for at least one world ω. We then define $\mathcal{C}r$ on some Boolean sub-algebra on the space of worlds as follows:

$$\mathcal{C}r(P) = sup\{\mathcal{D}(\omega): \omega \in P\}, \text{ and } \mathcal{C}r(\emptyset) = 0$$

Defining $\mathcal{C}r$ in this way implies that it is sub-additive:

[16] This point is unknown to Nolan, who notes that there are sets of worlds in his model which correspond to no proposition (as he uses the term): "However, not every arbitrary set of worlds should count as a proposition once enough impossible worlds are admitted—since an impossible world *w* which is obtained by adding further things true to all of the things true at a possible world *v* … will be such that *v* will occur in every proposition in which *w* occurs (on pain of a proposition being true at *v* which is not true at *w*): so those sets containing *w* but not *v* should not count as propositions." (Nolan 1997, p. 563).

$$\mathcal{C}r(P \cup P^*) = max\{\mathcal{C}r(P), \mathcal{C}r(P^*)\} \leq \mathcal{C}r(P) + \mathcal{C}r(P^*)$$

However, $\mathcal{C}r$ so-defined will still satisfy Nonnegativity, Normalisation, and Monotonicity, so it faces its own logical omniscience problems whenever we limit the space of worlds only to those which are possible.

More generally, the large majority of formal systems for the representation of partial beliefs will have $\mathcal{C}r$ satisfy at least one of Nonnegativity, Normalisation, and Monotonicity—e.g., Choquet capacities (Choquet 1954; applied in, e.g., Tversky and Kahneman 1992), Dempster-Shafer belief and plausibility functions (Dempster 1968; Shafer 1976), possibility measures (Dubois and Prade 1988), ranking functions (Spohn 2012), and the interval-valued functions of (Levi 1974) and (Kyburg 1992). (It's worth noting that these systems are generally structured around the assumption of Booleanism, or something much like it.) Given that it would be better not to throw out most of what we've managed to achieve *vis-à-vis* the formal representation of partial belief, it might be best to take a deeper look at the Booleanism assumption.

We could keep Booleanism *if* we made some changes to how we interpret $\mathcal{C}r$. For instance, instead of saying that $\mathcal{C}r(P) = x$ if and only if α has degree of belief x towards some object of belief represented by P, we might instead say that $\mathcal{C}r$ represents α's degrees of belief only where the propositions in question are expressible. But what then of the values that $\mathcal{C}r$ assigns to inexpressible propositions? One thought would be to say that while $\mathcal{C}r$ represents α's degrees of belief when P is expressible, it represents some other propositional attitude ϕ when P is inexpressible. For instance, one might think that if P is expressible, then $\mathcal{C}r(P^C)$ represents α's *degree of rejection* towards P, which plausibly is $1 - \mathcal{C}r(P)$.[17] However, this kind of 'rejectionist' proposal will only work if the complement of every inexpressible proposition is expressible, which is not in general the case. In particular, the domain of $\mathcal{C}r$ has to be closed under unions, and the complement of the (inexpressible) union of two expressible propositions will often be itself inexpressible. Of course, we could still suppose that there exists some broadly 'doxastic' attitude ϕ that we have directed towards inexpressible propositions—but what reason do we have for positing the existence of ϕ, beyond the desire to preserve some modelling assumptions?

A better option would be to drop the assumption of Booleanism altogether. The definition of $\mathcal{C}r$ in terms of $\mathcal{D}$ does not rely on it, and we can still get from $\mathcal{D}$ to a value for any collection of expressible propositions without it. However, we shouldn't be too quick to dismiss Booleanism, which might still play an important role when it comes to understanding our degrees of belief and related phenomena.

Most saliently, it frequently comes up in various representation theorems, where the assumption that $\mathcal{B}$ has some minimally rich algebraic structure is a basic component of our capacity to assign numerical values to the propositions within $\mathcal{B}$ in a meaningful and systematic way. For example, the assumption plays a role throughout Richard Jeffrey's (1990) representation theorem for expected utility theory—where, if we were to assume that the space of thinkable propositions $\mathcal{B}$ was such that none of its members is a subset of any other members, almost all of his axioms would be either meaningless or trivial. Booleanism is a standard assumption for theories of decision making and uncertainty, with most decision-theoretic representation theorems being built around it.[18]

Or consider Stefánsson's (forthcoming) account of numerical degrees of belief in terms of qualitative belief orderings over propositions, a common approach which goes back to (de Finetti 1931) and importantly dependent on $\mathcal{B}$ having a rich algebraic structure. Without

[17] Thanks to [anonymous] for raising this suggestion.

[18] (Steele and Stefánsson 2015) contains a philosophical overview of expected utility theory and some of its most important representation theorems.

something like the axiom of *qualitative additivity*—that if P and P^* both have null intersection with P^{**}, then one holds P to be more likely than P^* if and only if one holds $P \cup P^{**}$ to be more likely than $P^* \cup P^{**}$—the qualitative belief ordering would lack adequate enough structure to be anything more than a simple (and representationally inadequate) ordinal scale.

The probabilistic analogues of logical omniscience require some solution if we're ever going to model the partial beliefs of ordinary agents. But it's hard to see how that can be done so long as the objects of thought are treated as sets of merely possible worlds. The solution we end up with *may* involve the introduction of impossible worlds, but this looks to be a viable solution only if we drop the very standard—and oft relied upon—assumption of Booleanism, or if we embrace the inexpressibility of most of our thoughts. Neither option seems particularly appealing, and we may well do better to look for a solution without the impossible.

References

Adams, R.M. 1974. Theories of Actuality. *Nous.* **8**, pp.211-231.

Berto, F. 2010. Impossible Worlds and Propositions: Against the Parity Thesis. *The Philosophical Quarterly*. **40**, pp.471-86.

Berto, F. 2013. Impossible Worlds. *Stanford Encyclopedia of Philosophy*. [Online]. Available from: https://plato.stanford.edu/archives/win2013/entries/impossible-worlds/

Bjerring, J.C. 2013. Impossible Worlds and Logical Omniscience: an Impossibility Result. *Synthese.* **190**, pp.2505-24.

Bjerring, J.C. 2014. Problems in Epistemic Space. *Journal of Philosophical Logic.* **43**, pp.153-170.

Bjerring, J.C. and Schwarz, W. 2017. Granularity Problems. *The Philosophical Quarterly*. **67**(266), pp.22-37.

Boghossian, P.A. 1993. Does an Inferential Role Semantics Rest upon a Mistake? *Philosophical Issues.* **3**, pp.73-88.

Bricker, P. 1987. Reducing Possible Worlds to Language. *Philosophical Studies.* **52**(3), pp.331-355.

Chalmers, D. 2011a. Frege's Puzzle and the Objects of Credence. *Mind.* **120**(479), pp.587-635.

Chalmers, D. 2011b. The Nature of Epistemic Space. In: Egan, A. and Weatherson, B. eds. *Epistemic Modality*. Oxford: Oxford University Press, pp.60-107.

Chalmers, D. 2012. *Constructing the World.* Oxford University Press.

Chalmers, D. and Jackson, F. 2001. Conceptual Analysis and Reductive Explanation. *Philosophical Review*. **110**, pp.315-361.

Choquet, G. 1954. Theory of capacities. *Annales de l'institut Fourier.* **5**, pp.131-295.

Cozic, M. 2006. Impossible States at Work: Logical Omniscience and Rational Choice. In: Topol, R. and Walliser, B. eds. *Contributions to Economic Analysis.* Elsevier, pp.47–68.

Creswell, M.J. 1973. *Logics and Languages.* London: Methuen.

Davies, M. 1981. *Meaning, Quantification, Necessity: Themes in Philosophical Logic.* London: Routledge & Kegan Paul.

de Finetti, B. 1931. Sul Significato Soggettivo Della Probabilita. *Fundamenta Mathematicae.* **17**(1), pp.298-329.

Dempster, A.P. 1968. A Generalization of Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological).* **30**, pp.205-247.

Dubois, D. and Prade, H. 1988. *Possibility Theory. An Approach to Computerized Processing of Uncertainty.* New York: Plenum.

Dummett, M. 1978. *Truth and Other Enigmas.* Harvard University Press.

Easwaran, K. 2014. Regularity and Hyperreal Credences. *Philosophical Review.* **123**(1), pp.1-41.

Field, H.H. 1978. Mental Representation. *Erkenntnis.* **13**, pp.9-61.

Halpern, J.Y. and Pucella, R. 2011. Dealing with logical omniscience: Expressiveness and pragmatics. *Artificial Intelligence.* **175**(1), pp.220-235.

Hellie, B. 2004. Inexpressible Truths and the Allure of the Knowledge Argument. In: Nagasawa, Y., et al. eds. *There's Something About Mary.* MIT press, pp.333-64.

Hintikka, J. 1962. *Knowledge and Belief: An introduction to the logic of the two notions.* Ithaca: Cornell University Press.

Hintikka, J. 1975. Impossible Possible Worlds Vindicated. *Journal of Philosophical Logic.* **4**, pp.475-84.

Hofweber, T. 2006. Inexpressible properties and propositions. In: Zimmerman, D. ed. *Oxford Studies in Metaphysics.* Oxford: Oxford University Press.

Hofweber, T. 2016. Are there ineffable aspects of reality? In: Bennett, K. and Zimmerman, D. eds. *Oxford Studies in Metaphysics, Vol. 10.* Oxford: Oxford Unviersity Press.

Jago, M. 2009. Logical information and epistemic space. *Synthese.* **167**, pp.327-341.

Jago, M. 2012. Constructing Worlds. *Synthese.* **189**, pp.59-74.

Jago, M. 2013. Are Impossible Worlds Trivial? In: Puncochar, V. and Svarny, P. eds. *The Logica Yearbook 2012.* College Publications, pp.35-50.

Jago, M. 2014a. *The Impossible: An Essay on Hyperintensionality.* Oxford University Press.

Jago, M. 2014b. The Problem of Rational Knowledge. *Erkenntnis.* **79**, pp.1151-1168.

Jago, M. 2015a. Impossible Worlds. *Nous.* **49**(4), pp.713-728.

Jago, M. 2015b. Hyperintensional Propositions. *Synthese.* **192**(3), pp.585-601.

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

Kaplan, D. 1995. A Problem in Possible Worlds Semantics. In: Sinnott-Armstrong, W., et al. eds. *Modality, Morality and Belief: Essays in Honor of Ruth Barcan Marcus.* Cambridge: Cambridge University Press, pp.41-52.

Katz, J. 1978. Effability and Translation. In: Guenthner, F. and Guenthner-Reutter, M. eds. *Meaning and Translation.* New York: NYU Press, pp.157-189.

Kyburg, H.E. 1992. Getting Fancy with Probability. *Synthese.* **90**, pp.189-203.

Levi, I. 1974. On Indeterminate Probabilities. *The Journal of Philosophy*. **71**(13), pp.391-418.

Lewis, D. 1979. Attitudes De Dicto and De Se. *The Philosophical Review*. **88**(4), pp.513-543.

Lewis, D. 1986. *On the Plurality of Worlds.* Cambridge University Press.

Lewis, D.K. 2004. Letters to Beall and Priest. In: Priest, G., et al. eds. *The Law of Non-contradiction: New Philosophical Essays.* Clarendon Press, pp.176-177.

Lipman, B.L. 1997. Logics for Nonomniscient Agents: An Axiomatic Approach. In: Bacharach, M., et al. eds. *Epistemic Logic and the Theory of Games and Decisions.* Springer, pp.193-216.

Lipman, B.L. 1999. Decision Theory without Logical Omniscience: Toward an Axiomatic Framework for Bounded Rationality *The Review of Economic Studies.* **66**(2), pp.339-361.

Nolan, D. 1997. Impossible Worlds: A Modest Approach. *Notre Dame Journal of Formal Logic.* **38**, pp.535-72.

Nolan, D. 2013. Impossible Worlds. *Philosophy Compass.* **8**(4), pp.360-372.

Perry, J. 1979. The Problem of the Essential Indexical. *Nous.* **13**(1), pp.3-21.

Priest, G. 2006. *In contradiction: a study of the transconsistent.* Oxford: Oxford University Press.

Rantala, V. 1982. Impossible Worlds Semantics and Logical Omniscience. *Acta Philosophica Fennica.* **35**, pp.106–15.

Schiffer, S. 2003. *The things we mean.* Oxford: Oxford University Press.

Searle, J.R. 1969. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press.

Shafer, G. 1976. *A Mathematical Theory of Evidence.* Princeton: Princeton University Press.

Shaw, J.R. 2013. Truth, Paradox, and Ineffible Propositions. *Philosophy and Phenomenological Research.* **86**(1), pp.64-104.

Spohn, W. 2012. *The Laws of Belief: Ranking Theory and Its Philosophical Application.* Oxford: Oxford University Press.

Stalnaker, R. 1996. Impossibilities. *Philosophical Topics.* **24**, pp.193-204.

Steele, K. and Stefánsson, H.O. 2015. Decision Theory. In: Zalta, E.N. ed. *The Stanford Encyclopedia of Philosophy.* [Online]. Available from: <http://plato.stanford.edu/archives/win2015/entries/decision-theory/>

Stefánsson, H.O. forthcoming. What Is 'Real' in Probabilism? *Australasian Journal of Philosophy.*

Tversky, A. and Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty.* **5**(4), pp.297-323.