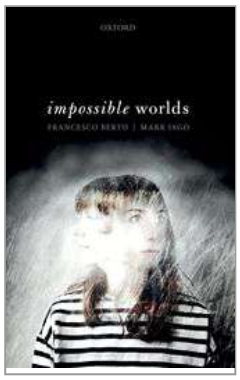

Impossible Worlds
Francesco Berto and Mark Jago

Print publication date: 2019
Print ISBN-13: 9780198812791
Published to Oxford Scholarship Online: August 2019
DOI: 10.1093/oso/9780198812791.001.0001

Hyperintensionality

Francesco Berto
Mark Jago

DOI:10.1093/oso/9780198812791.003.0008

Abstract and Keywords

This chapter asks whether hyperintensionality is a genuine phenomenon, or rather, a feature to be explained away. It then focuses on the epistemic case, considering arguments from Stalnaker and Lewis which attempt to explain away hyperintensionality. The argument for a genuinely hyperintensional notion of content is subsequently considered. Having made the case for genuine hyperintensionality, the chapter turns to the *granularity issue*: how fine-grained are impossible worlds? This is one of the most difficult issues any theory of hyperintensionality faces. The focus then returns to the *compositionality objection* and it is argued that some accounts of impossible worlds deliver a fully compositional theory of meaning.

Keywords: hyperintensionality, Stalnaker, Lewis, granularity, fine-grained worlds, compositionality

8.1 Is Hyperintensionality Real?

In §1.3, we introduced the concept of *hyperintensionality*. We said that an operator $\mathscr{H}$ is hyperintensional when $\mathscr{H}A$ and $\mathscr{H}B$ can differ in truth value, even when A and B are necessarily (logically, mathematically, or metaphysically) equivalent. We identified concepts which have been given an intensional possible worlds account, such as knowledge, belief, meaning and semantic content, propositions, information, counterfactual conditionals. We claimed that these concepts are hyperintensional and hence that their possible worlds account falls short of capturing the concept.

Some philosophers deny that some of these concepts are genuinely hyperintensional. Some deny that there are any genuinely hyperintensional concepts. These denials sometimes derive from other theoretical commitments. A philosopher who analyses content or meaning in terms of possible worlds will be unable to find any content or meaning in hyperintensional operators. She will view any purported hyperintensional operators as meaningless. But since knowledge, belief, and so on are meaningful concepts, she will deny that these are hyperintensional. She'll then argue that the appearance of hyperintensionality (in examples involving knowledge and belief reports, for example) are illusory. She may also offer a positive analysis of the concept, showing why it can't be hyperintensional.

In the cases of knowledge and belief, we find these moves in Lewis (1982, 1986b) and Stalnaker (1984). Both authors support the **(p.162)** possible worlds analysis of knowledge, belief, and content. They base their support on considerations of the nature of these concepts. Both acknowledge that we do not *seem* to know all consequences of what we know, that logical reasoning *seems* to be informative, and so on. But, they argue, the appearances are misleading. We'll discuss their suggestions in §8.2.

Belief and knowledge are everyday concepts, and so we may draw on everyday reflections on those concepts as (defeasible) evidence for hyperintensionality. Proposition, essence, grounding, and (to some extent) meaning and content are, by contrast, terms of art. A philosopher has more scope to rebut purported examples of hyperintensionality using these concepts. 'That's not how I use that concept', she may say. She may even reject the concept outright. So our claims here will be somewhat conditional in nature. If one wants to adopt one of these concepts (in the way that we think is the most philosophically useful), then one had better accept hyperintensionality. We'll discuss these cases in §8.3.

8.2 The Epistemic Case

For Stalnaker, the nature of belief, as a rational attitude, rules out a hyperintensional analysis. He begins with an 'impressionistic' picture of human representational states:

> Representational mental states should be understood primarily in terms of the role that they play in the characterization and explanation of action. What is essential to rational action is that the agent be confronted, or conceive of himself as confronted, with a range of alternative possible outcomes of some alternative possible actions. The agent has attitudes, pro and con, toward the different possible outcomes, and beliefs about the contribution which the alternative actions would make to determining the outcome. One explains why an agent tends to act in the way he does in terms of such beliefs and attitudes. And, according to this picture, our

> conceptions of belief are conceptions of states which explain why a rational agent does what he does.
>
> (Stalnaker 1984, 4)

(p.163) The key idea here is that human agents take *attitudes, pro and con, toward the different possible outcomes*. Suppose we assign a 1 to each possible outcome to which the agent takes a pro attitude, and a 0 when she takes a con attitude. Then the content of the attitude is captured as a *characteristic function* from possible outcomes to 1 or 0. Such functions are mathematically equivalent to the set of outcomes assigned 1. The content of any rational attitude is (equivalent to) a set of possibilities, on this analysis. Importantly, this argument doesn't proceed from a prior commitment to possible worlds semantics. Rather, it proceeds from an analysis of 'the role of beliefs and desires in the explanation of action' and concludes that 'the contents of those attitudes [must] distinguish between the alternative possibilities' (Stalnaker 1984, 23). Hence the possible worlds analysis.

Stalnaker acknowledges that we do not seem to be logically omniscient. The situation, he thinks, is that 'we have an argument to show that the identity conditions [on contents] are right, as well as examples that seem to show that they are wrong', and so 'the proper response is not so clear' (Stalnaker 1984, 24).

One suggestion he offers is a *metalinguistic* approach, on which

> the apparent failure to see that a proposition is necessarily true, or that propositions are necessarily equivalent, is to be explained as the failure to see what propositions are expressed by the expressions in question.
>
> (Stalnaker 1984, 84)

How might this explanation go? Stalnaker elaborates:

> Relative to any propositional expression one can determine two propositions: there is the proposition that is expressed, according to the standard rules, and there is the proposition that relates the expression to what it expresses. If sentence *S* expresses (according to the standard rules) proposition *p*, then the second proposition in question is the proposition that *S* expresses *p*. In cases of ignorance of necessity and equivalence, I am suggesting, it is the second proposition that is the object of doubt and investigation.
>
> (Stalnaker 1984, 84–5)

It seems that if I do not know a necessary truth, *A*, then what I do not know is that the sentence '*A*' expresses the proposition *that A*. I do **(p.164)** not know that '*A*' expresses the set of all possible worlds, according to Stalnaker. I countenance possibilities in which the words in '*A*' mean something other than their actual meaning and so, at some level, I fail to understand what '*A*' means. For

> Whenever the structure of sentences is complicated, there will be a nontrivial question about the relation between sentences and the propositions they express, and so there will be room for reasonable doubt about what proposition is expressed by a given sentence.
>
> (Stalnaker 1984, 84)

This is clearly a genuine phenomenon. Anyone who's read enough German, or even old-fashioned English, knows the feeling. You know what each word in turn means but, as the sentence runs on and on, you lose the overall meaning. You don't know what proposition is expressed. It happens. But we question whether it happens enough to make Stalnaker's case. Is it the case that, whenever one seemingly fails to know (or believe) a necessary truth, one is confused about what the sentence in question means? That seems implausible.

Here are two examples (from Jago (2014a)) to make the case. The first is mathematical. You learn, in school, about integers, equations, addition, and exponentiation. You don't learn all there is to know about them, of course, but enough to be comfortable with what those terms mean. (Let's say, you know how to relate exponentiation to multiplication to addition, and you know well what addition is.) You also know how to understand notation involving variables (quantified, implicitly, over integers). So you have no trouble understanding an equation like

$$x^n + y^n = z^n$$

Do you thereby know that, when $n > 2$, there are no three integers x, y, z fitting the equation? Of course not! It took mathematicians centuries to find a proof for that claim. Nevertheless, it's a mathematical and logical necessity (with 'logic' understood so as to contain basic arithmetic). In this case, Stalnaker must claim that you don't understand something about that statement. But, given the way we set up the example, that's hardly plausible.

(p.165) The second example is more practical. Suppose you're playing chess (with no time controls) and agree to count a draw as a win for black. (Or suppose you're in a competition, and black needs only a draw to win the competition overall.) Here's a surprising mathematical fact: at each stage of the game, either you or your opponent has a *winning strategy* available. That's a function for generating the next move which, if followed to the letter, will result

in victory (including a draw for black), regardless of what the other player does. A winning strategy exists right from the first move, either for white or for black (but we don't know which). If players always followed a winning strategy, then the same colour would always win (always white, or always black). In reality, that doesn't happen. Either the players don't know the winning strategy when it's there (they don't know that, say, Qe7 is the first move of a winning strategy), or else they're really not that fussed about a guaranteed win. (And it's obviously not the latter.)

In this case, Stalnaker again has to claim that we fail to know some of the meanings involved. But all that's required (mathematically speaking) to entail a winning strategy are (1) a precise specification of the rules of chess, (2) a precise specification of the conditions for winning, and (3) a precise description of the current state of the board. Given those facts, a description of a winning strategy (for one or other of the players) follows mathematically. So, the claim has to be that no chess player (who ever lost without wanting to) understands the meaning of the rules, the winning conditions, or the way we describe the state of the board. That is highly implausible.

In short, the phenomenon Stalnaker describes, not knowing which proposition is expressed by a given sentence, is genuine. But it is not able to account for many of the cases in which an agent seems not to know some necessary truth.

A different approach is to think of an agent's overall epistemic state as being split into *fragments*, or multiple 'frames of mind'. An agent may believe something in one frame of mind, something else in some other frame of mind, and never combine the two bits of information. She fails to believe all consequences of what she believes, on this **(p.166)** approach, because (as it were) she never puts two and two together. Lewis (1982) describes how to make sense of the phenomenon:

> I used to think that Nassau Street ran roughly east-west; that the railroad nearby ran roughly north-south; and that the two were roughly parallel. … So each sentence in an inconsistent triple was true according to my beliefs, but not everything was true according to my beliefs. … My system of beliefs was broken into (overlapping) fragments. Different fragments came into action in different situations, and the whole system of beliefs never manifested itself all at once. … The inconsistent conjunction of all three did not belong to, was in no way implied by, and was not true according to, any one fragment. That is why it was not true according to my system of beliefs taken as a whole.
>
> (Lewis 1982, 436)

On Lewis's approach, each fragment of belief can be treated in the possible worlds tradition. So, each fragment corresponds to a proposition, representing a consistent and logically closed set of beliefs. Yet the agent's beliefs in total are not closed under entailment and need not be consistent, as in Lewis's example. Stalnaker (1984) supports a similar view (which he holds in combination with the metalinguistic approach described above). Fagin and Halpern (1988) develop the idea into a formal semantics, which they call a model of 'local reasoning'. (The idea is that the logically omniscient 'reasoning' is local to each fragment, rather than a global feature of the agent's total belief state.)

Formally, the approach is similar to a non-adjunctive *subvaluational* semantics, on which premises A, B do not entail their conjunction, $A \wedge B$ (§4.4). On this approach, we consider a range of classical valuations, and consider a sentence to be true (simpliciter) when it's true-on-some-valuation. When A is true on one valuation and B true on another, but no valuation makes both A and B true simultaneously, we have a situation in which both A and B are true (simpliciter) but $A \wedge B$ is not. In the doxastic setting, this translates to the situation Lewis describes: he did not believe the inconsistent conjunction of the information he held about Nassau Street and the railroad.

(p.167) As with the metalinguistic approach, the phenomenon described by this approach is genuine, but it cannot explain away enough cases of non-omniscience. Consider our chess example from above. Let's assume that our players are highly competent players, who hold all the relevant information (the rules, what it takes to win, and the current state of the game) in their current frame of mind. (Not all players are like that, of course. But many who reach a high level of competence are.) According to the fragments of mind approach, those players will know (and be able to act upon) a winning strategy. But, as we know from experience, that just isn't the case, not even for the very best players.

Following Jago (2014a), we think that the fragments of mind approach misdiagnoses why real agents aren't logically omniscient. According to the view, an agent doesn't believe a consequence of what she believes because she hasn't put the relevant premises together. But when she does put those premises together, she thereby, all of a sudden, comes to believe all their consequences. It's as if all the agent's deductive effort goes into combining the premises into a single belief state: from $A_1, \ldots, A_n$ to $A_1 \wedge \cdots \wedge A_n$. But that's the easy bit, deductively speaking. The hard bit, for real agents, is deriving further non-trivial consequences. Whether she derives these from the individual premises taken together, or from their conjunction, doesn't make that much difference. The information conveyed by a non-trivial valid deduction does not correspond to the move from individual premises to their conjunction, but rather in the deductive move from those premises (or their conjunction) to the conclusion. The fragments of belief approach can't account for this phenomenon.

There are also technical problems with the approach. Although an agent's total belief state is not closed under under entailment, on the fragments view, it is closed under single-premise entailment. If A entails B and she believes that A, then she automatically believes that B (and similarly for knowledge). This is condition (C1) from §5.1. This holds because each fragment is logically closed. The approach also validates (C2)–(C5), (C7), and (C8) from §5.1. These are all forms of logical omniscience we would like to avoid, both for belief **(p.168)** and for knowledge (with the exception of (C3) for knowledge, since one can't know inconsistent things). So, even on technical grounds, the fragments of belief approach isn't a sufficient solution to the problems of logical omniscience.

We've phrased our discussion in terms of logical omniscience. How does this relate to hyperintensionality? Non-hyperintensionality is one form of logical omniscience: (C4) from §5.1, *Closure under logical equivalence*. It says that knowledge (or belief) never distinguishes between logically equivalent contents. This is the very definition of a non-hyperintensional concept.

Closure under logical equivalence (C4) is closely related to closure under single-premise entailment. Suppose that A logically entails B. Then A is equivalent to $A \wedge B$. Given (C4), an agent who knows that A thereby knows $A \wedge B$ too. And if she knows the conjuncts of each conjunction she knows, then she knows that B too. (Similarly for belief.) The latter principle is *Closure under conjunction*, (C7). So (C4) plus closure under conjunction, (C7), implies (C1). If agents are not logically omniscient at all, then (C1) is false. But as we argued above, we cannot pin this failure on (C7), and so (C4) must be false too. So if agents are not logically omniscient, as we've been arguing, then (C4) is false, then knowledge and belief are hyperintensional concepts.

Here's the situation, as we see it. Examples such as our chess case intuitively show that agents aren't logically omniscient. Lewis's and Stalnaker's attempt to convince us otherwise fail. Absent a successful attempt along those lines, the evidence directs us to accept that knowledge and belief are genuinely hyperintensional.

8.3 The Case from Content

We've been discussing doxastic and epistemic contents. But the general notion of content is much broader. Every meaningful expression has a semantic content, not just those that are known, believed, or uttered. If epistemic or doxastic contents are hyperintensional, then *content* **(p.169)** in general is a hyperintensional notion. But there are independent (non-epistemic) reasons for thinking that this is the right way to think about content. The content of a truth-apt sentence is a proposition. We'll talk in terms of propositions, since there's a straightforward notion of logical equivalence for propositions. (We can also talk about logically equivalent subsentential terms, such as descriptions, but this would needlessly complicate our discussion.)

Both Lewis (1973b, 1986b) and Stalnaker (1976a, 1976b, 1984) take propositions to be sets of possible worlds. There's a very simple argument for that view. It's common to identify propositions with truth-conditions. But what, ontologically speaking, is a *condition*? Let's say we're interested in whether something meets condition X in such-and-such situations. We treat that condition as a function from the situations to the answers, *yes* or *no*. Mathematically, such functions are *characteristic functions*, and each such function defines a set, containing all and only the input entities for which the function's output is *yes*. It is then both very natural and mathematically elegant to identify the condition itself with the set of 'yes' situations. In the case of a truth-condition, the input situations are possible worlds and the outputs are *true* or *false*. So, on this very natural view, a truth-condition is a set of possible worlds. (This is similar to the argument we met at the start of §8.2.)

On this approach, the proposition $\langle A \rangle$ is the set of all possible worlds according to which it is the case that A. (Alternatively, the worlds which are such that, were they actual, then it would be the case that A.) This story comes with a ready-made account of what logically complex propositions are, for the set-theoretic account gives a Boolean algebra for propositions: $\langle A \wedge B \rangle$ is simply the set of worlds $\langle A \rangle \cap \langle B \rangle$ and similarly, $\langle A \vee B \rangle = \langle A \rangle \cup \langle B \rangle$, $\langle \neg A \rangle = \langle A \rangle^{c}$ (the set-theoretic complement of $\langle A \rangle$, on the domain of all possible worlds) and $\langle A \rightarrow B \rangle = \langle A \rangle^{c} \cup \langle B \rangle$. A consequence of this account is that necessarily equivalent propositions are identical (Stalnaker 1976a, 9). That is to say, the approach is intensional, not hyperintensional.

Despite the simplicity and naturalness of the approach, its shortcomings are many. We'll discuss two (related) ways to bring these out **(p.170)** (following Jago (2018b)). First, the *truthmaker objection*. Consider these two propositions:

(8.1) ⟨Lenny exists⟩

(8.2) ⟨Lenny exists ∧ 3 exists⟩.

True propositions have truthmakers. (Or, at the very least, true existential propositions have truthmakers.) Lenny alone truthmakes (8.1). But Lenny on his own does not truthmake (8.2): the number 3 gets in on the act, too. It's in virtue of the existence of both Lenny and the number 3, taken together, that (8.2) is true. Lenny is a *full* truthmaker for (8.1), but not for (8.2). Hence (8.1) stands in a relation to Lenny which (8.2) does not and so, by Leibniz's Law, (8.1) and (8.2) cannot be one and the same proposition. But they are necessarily equivalent and hence, according to the possible worlds view, one and the same proposition. Contradiction.

A related objection to the possible worlds account is the *aboutness objection*. *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, 1). Research on

aboutness has been flourishing recently, mainly thanks to Fine (2016) and Yablo (2014). Even before introducing a ‘grand-sounding name for something basically familiar’ (Yablo 2014, 1), logicians and semanticists were looking for a content-preserving entailment relations: relations that holds between two meaningful items A and B only when B introduces no content alien to what A is about. *Tautological entailment* (Van Fraassen 1969), *analytic containment* (Angell 1977, Correia 2004), and *analytic implication* (Parry 1933) are all variations on this theme.

As an example of how aboutness creates problems for the standard possible worlds account, consider the propositions:

(8.3) ⟨Lenny is stretching ∨ ¬Lenny is stretching⟩

which is about Lenny, not Bertie, and

(8.4) ⟨Bertie is barking ∨ ¬Bertie is barking⟩

(p.171) which is about Bertie, not Lenny. They’re about different things, and hence are different propositions. *Aboutness* should thus relate (8.3) to Lenny, not Bertie, and (8.4) to Bertie, not Lenny. So (8.3) and (8.4) stand in different relations, and hence are distinct entities. But they are logically equivalent. The possible world view then says they are one and the same proposition. Contradiction. In each case, we say, it’s the possible worlds view of propositions that’s at fault. It cannot model aboutness and aboutness-preserving entailment. There is broad agreement that a semantics that works properly for these notions should be hyperintensional (Yablo 2014, 62). (That said, Yablo himself works within the possible worlds framework, but with additional hyperintensional resources.)

We’ve argued that at least some concepts are hyperintensional. And unsurprisingly, we find the best way to account for such concepts is in terms of impossible (as well as possible) worlds. Impossible worlds give us a very natural and flexible way to account for hyperintensional concepts.

Let’s continue with our focus on propositions. As we said above, it’s common to identify propositions with truth conditions, and in general to identify conditions with characteristic functions. So a proposition is identified with the set of situations in which it is true. As we emphasized above, that’s a very natural view. But there’s nothing in this argument that says the situations in question have to be possible worlds. We take them to be sets of worlds, both possible and impossible. So we can accept the motivation whilst accepting that propositions are hyperintensional.

Just what logical properties such propositions have depends on the properties of impossible worlds. It depends on which logical rules, if any, impossible worlds must preserve. This is the *granularity issue*, our topic for §8.4.

(p.172) 8.4 The Granularity Issue

Given that a world represents such-and-such, what else must it represent? This is one of the most difficult issues surrounding impossible worlds. A logically possible world which represents that A and that B must thereby represent that $A \wedge B$, that $A \vee B$, that $\neg\neg A$, and so on. Which of these closure principles, if any, apply to impossible worlds? We call this the *granularity issue*. In this section, we'll discuss some considerations on granularity in general. We'll then look at how different considerations might apply to different hyperintensional concepts.

Let's begin with *Nolan's Principle* (Nolan 1997, 542):

> (NP) If it is impossible that A, then there's an impossible world which represents that A.

Nolan thinks of this as a kind of unrestricted 'comprehension principle' for impossibilities. If true, it tells us something about both the nature and the scope of impossible worlds.

Given (NP), you might think we can show that impossible worlds needn't obey any particular logical closure principle (other than *identity*, $A \vDash A$). If that's right, then impossible worlds are *open worlds* (Priest 2005, Rantala 1982a). We defined open worlds in §5.3 as worlds which, in general, obey no logical closure principle other than $A \vDash A$. (From now on, when we say 'closed under no principle', we'll always mean 'no principle except $A \vDash A$'.)

The argument from (NP) to open worlds goes as follows. Take any putative closure principle, taking us from some premises $A_1, A_2, \ldots$ to C. If the principle isn't valid, then there's a logically possible world such that $A_1, A_2, \ldots$ but not C. If it is valid, by contrast, then it cannot be the the case that $A_1, A_2, \ldots$ but not C. But then, by (NP), there is an impossible world such that $A_1, A_2, \ldots$ but not C. Generalizing, for any closure principle, there is a world not closed under that principle.

This argument is flawed, for it misapplies Nolan's Principle. As stated, Nolan's Principle allows us to take any single sentence 'A' **(p.173)** which cannot be true, and infer the existence of an impossible world which represents that A. But the argument above considered multiple sentences, 'A_1', 'A_2', … and 'C'. These cannot together be substituted into (NP) as stated. (We'll discuss more general principles, which Nolan may have had in mind, later in this section.)

We can always substitute for 'A' in (NP) the single sentence, '$A_1 \wedge A_2 \wedge \cdots \wedge \neg C$'. This is impossible iff it cannot be the case that $A_1, A_2, \ldots$ but not C. So, using this

sentence, (NP) legitimately gives us an impossible world w which represents that $A_1 \wedge A_2 \wedge \cdots \wedge \neg C$. However, there is no guarantee that w represents each A_i but does not represent that C. We haven't yet fixed which principles hold of impossible worlds and so, in particular, we can't assume those principle. It may be that w represents a conjunction without representing each of the conjuncts. It may represent both $\neg C$ and C.

We could instead try substituting a sentence of the metalanguage for 'A' in (NP), such as this one: ''A_1', 'A_2', ... are true, but 'C' is not true'. Then (NP) gives us an impossible world w where that entire sentence is true. But again, this need not be a world where the A_is hold but C does not. To make that inference, w would need to support the inferences from A's being true to A's being the case and from A's not being true to A's not being the case. But since w is an impossible world, we can't assume that those principles hold of it. So however we try, we can't get directly from (NP) to the failure of all closure principles.

Priest (2016a) adopts two principles that are similar to, but stronger than, (NP): 'everything holds at some worlds, and everything fails at some worlds' (Priest 2016a, 5) and, for any distinct A, B, 'there are worlds where A holds and B fails' (Priest 2016a, 7). More specifically, in our terminology:

(8.5) For any A, there is a world which represents that A and a world which does not represent that A.

(8.6) For any distinct A and B, there is a world which represents that A but does not represent that B.

(p.174) Priest calls these the 'primary directive' and 'secondary directive' on impossible worlds, respectively. The latter implies the former, which in turn implies (NP), but neither converse holds. To see this, first suppose that it is impossible that A. Then (8.5) says there is a world such that A, which by definition is an impossible world. So (8.5) implies (NP). Suppose instead that A is necessary. Then (8.5) implies that there is an impossible world which does not represent that A. But (NP) does not imply this. Since A is necessary, $\neg A$ is impossible, and so (NP) says there is an impossible world which represents that $\neg A$. But this need not be a world which does not represent that A: it may represent *both* that A and that $\neg A$, as some of the FDE-worlds from §5.4 do. In general, (NP) tells us only about the existence of worlds which represent such-and-such. It does not tells us about the existence of worlds which *fail* to represent such-and-such. That's why (8.5) is strictly stronger than (NP).

It's clear that (8.6) implies (8.5), but the converse does not hold. For any pair of sentences 'A' and 'B', (8.5) tells us that there is a world which represents that A and a world which fails to represent that B. But these could be distinct worlds. Since there's no way to infer that they're the same world, as (8.6) requires, (8.5)

doesn't imply (8.6). So we get increasingly strong principles as we go from (NP) to (8.5) to (8.6).

To illustrate the extra power (8.6) gives us (over (8.5) and (NP)), consider Simplification, the inference from $A \wedge B$ to A, or Disjunction Introduction, from A to $A \vee B$. (8.6) directly entails that there are worlds where these rules fail. These are not FDE-worlds (which respect the classical rules for conjunction and disjunction).

But even (8.6) does not get us to the existence of open worlds, where any logical rule (except $A \vDash A$) can fail. For consider *adjunction*, from A and B to $A \wedge B$. Given (8.6), we may infer that there's a world which represents that A but not $A \wedge B$, and that there's a world which represents that B but not $A \wedge B$. But we can't get to a world which represents *both* that A and that B, but not that $A \wedge B$. (It clearly won't help to consider the premises conjunctively: there's no world which represents and does not represent that $A \wedge B$.)

(p.175) To infer the existence of open worlds from such-and-such being impossible, we need a stronger principle still, such as the following:

> (NP^+) If it is impossible that $A_1, A_2, \ldots$ but not C, then there's an impossible world which represents that $A_1, A_2, \ldots$ but not C.

This is very much in the spirit of the original (NP). Given it, for any logical principle (other that $A \vDash A$), there's a world which breaks that principle. If (NP^+) is true, then we need to include open worlds in our theory. However, there is virtually no gap between (NP^+) and that conclusion: (NP^+) says, in effect, that for any logical rule, there's a world which breaks that rule. It too needs an argument. And whereas Nolan's original principle (NP) has plenty of intuitive force (for those who believe in the existence of impossible worlds to begin with), we can hardly claim that for (NP^+).

Despite the failure of this argument, we believe that there exist open worlds. There are three arguments we can offer. The first is an argument from logical rules. If there are impossible worlds at all, then there are worlds which break *some* logical rule. More carefully: if there are qualitatively distinct impossible worlds, then there are impossible worlds which do not represent that A for every A. Such worlds are not governed by all the usual logical inference rules. Let's fix on the standard introduction and elimination rules for each connective, such as Adjunction and Simplification for conjunction. Each connective's meaning is intimately related to its introduction and elimination rules. (Just think about how Adjunction and Simplification fix what '$\wedge$' means.) But if one of these meaning-fixing rules can be broken by an impossible world, then surely any of them can. And if our domain of worlds does not in general respect any logical rule, we have a domain of open worlds. This argument doesn't give us a positive reason for

thinking that all rules are broken by impossible worlds. Rather, it looks to convince us that there's no reason to reject that approach.

The second argument is an argument from epistemic states such as belief. (Jago (2014a) and Priest (2016a) give a similar argument.) **(p.176)** Take any putative closure principle P, from $A_1, A_2, \ldots$ to (distinct) C. If all worlds are closed under P, and we analyse belief in terms of accessible worlds, then any agent who believes that $A_1, A_2, \ldots$ will thereby be modelled as believing that C, too. But it is at least possible for some agent to believe $A_1, A_2, \ldots$ but not believe that C. So not all worlds are closed under P.

The argument is quite general. If some possible agent can believe the premises of some logical principle but not the conclusion, then there must be an epistemically accessible world which represents those premises but not the conclusion. That was basically our argument, in §5.4, against using a weaker-than-classical logic to account for epistemic states. If we restrict the worlds accessible to an agent to, say, the FDE worlds, then the agent remains logically omniscient, with respect to the first-degree entailments of her beliefs. To avoid that situation, for any closure principle, there must be an accessible world which doesn't obey that principle. That's the argument for open worlds from epistemic states.

The third argument for open worlds is an argument from counterpossible reasoning. (Priest (2016a) gives a similar argument.) Suppose we're debating whether some putative logical principle should be accepted. We might be studying philosophy of logic, and discussing arguments in favour of intuitionism or paraconsistency, for example. Then we might make suppositions about the validity (or otherwise) of Excluded Middle ($A \vee \neg A$), Double-Negation Elimination ($\neg\neg A \vDash A$), or the Explosion Principle ($A, \neg A \vDash B$). We'll consider what happens if that principle fails. If the principle is in fact valid, then we're making a counterpossible supposition (§1.3).

If we want to analyse counterpossibles using impossible worlds (as we propose in Chapter 12), then we need impossible worlds in our account which violate that particular logical principle. It seems that we can engage in this kind of supposition for *any* logical principle (other than $A \vDash A$). For each such principle, we'll need impossible worlds in our account which violate it. So impossible worlds in general cannot be closed under any logical principle (other than $A \vDash A$). That's just to say that we need open worlds in our account.

(p.177) It doesn't follow from these arguments that one *single* world breaks all closure principles. It might be that each closure principle is broken by some world, although no world breaks them all. So, although the arguments above establish that we need open worlds in our account, it doesn't establish that the full range of open worlds (including worlds which respect no closure principle

whatsoever) is required. Nevertheless, the simplest semantics available (given that open worlds are required at all) is to allow models to be built from any collection of open worlds, with no restrictions. Then some models will involve only worlds which meet a certain condition; but there will also be models containing worlds which don't meet that (or any other) condition.

For this perspective to make sense, we need a very fine-grained notion of how worlds represent. We need world-representation to be at least as fine-grained as the sentences of the object language. For otherwise, there would be distinct object-language sentences A, B such that a world represents that A iff it represents that B. But this gives us a closure principle ($A \dashv\models B$) which all worlds must obey, contrary to what we established above.

Linguistic ersatzism is the most straightforward way to realize this very fine-grained notion of world-representation. If impossible worlds are themselves sets of sentences, and any such set counts as a world, then world-representation will be as fine-grained as the worldmaking language itself. That is, for any pair A, B of worldmaking sentences, there will be a world which represents that A but not B. Moreover, as we argued in §3.6, linguistic ersatzism is probably the best way to make sense of impossible worlds. (This argument is neutral between linguist ersatzism proper and the linguistic variant of the hybrid account from §2.5, on which the possible worlds are Lewisian and the impossible worlds are sets of sentences.)

If we then make use of the class of all such worlds, we obtain extremely fine-grained analyses of hyperintensional concepts. But this is a problem in itself. We saw in §5.3, in our discussion of logical omniscience in epistemic logic, that this 'anything goes' approach to knowledge or belief tells us nothing about those concepts. If an agent can believe anything, whilst disbelieving anything else, then in **(p.178)** what sense do we have a model of belief at all? This 'anything goes' approach to belief seems equivalent to the purely syntactic approaches of Eberle (1974) and Moore and Hendrix (1979). These contain an arbitrary set of sentences, which are taken as representations of the agent's beliefs. If we can do things that way, why bother with the worlds apparatus in the first place?

Here, we have reached a deep puzzle about hyperintensionality. Hyperintensional concepts require our models to be extremely fine-grained. Yet when we achieve that fine granularity, we seem to have surrendered the benefits of the approach.

We find the arguments above, over the granularity of worlds, convincing. So we accept the open worlds picture: in general, worlds need be closed under no particular logical rule. Our metaphysics of what worlds are, and of how they represent, must be in accord with this principle. But it doesn't follow that 'anything goes' with any analysis of a hyperintensional concept, for two reasons.

First, a particular concept may require a certain kind of world, obeying certain conditions. When analysing possibility and necessity, for example, we restrict our attention to the possible worlds only. Although these are not intensional concepts, a similar restriction applies to many hyperintensional concepts. If we're interested in what's possible or necessary *with respect to intuitionistic logic*, or *with respect to paraconsistent logic*, we restrict our attention to the worlds which obey those logics.

There are (hyperintensional) notions of *semantic content* which, we think, require us to restrict the domain of worlds. (We discuss one such notion in §9.6.) These notions require worlds more fine-grained than classical possible worlds, but not as fine-grained as open worlds. We obtain those worlds by narrowing down the class of all open worlds, based on certain principles we want to enforce on our analysis. None of these cases allow that 'anything goes'. As we shall see in §9.6, there are substantial, non-trivial equivalences on semantic contents, which we can capture in our impossible worlds framework. That's the first reason why adopting open worlds doesn't imply 'anything goes' for a hyperintensional concept.

(p.179) The second reason is that an account of a hyperintensional concept may involve non-trivial structure, even if its worlds are themselves unstructured open worlds. One example is given by worlds-theories of counterfactuals (§1.2), which involve a similarity metric on worlds. We can adopt that approach even if we include open worlds in the account. (Again, we discuss counterpossible conditionals in detail in Chapter 12.) The metric can give us structure when we want it, but not when we don't. Let's counterfactually suppose that intuitionistic logic is the correct account of validity, as in our example above. Then the similarity metric will select worlds where intuitionistic principles operate but which (let's assume) are otherwise most similar to our own. This approach tells us that, if intuitionistic logic were correct, Double Negation Elimination would not be valid. But it also tells us that, if intuitionistic logic were correct, Nottingham would still be wet today. And it doesn't tell us that, if intuitionistic logic were correct, there would be flying elephants in Amsterdam. (This all depends on the details of the similarity metric.)

In this kind of approach, anything may be supposed. But it isn't the case that anything goes within a supposition. The similarity metric (or some other metric, if you prefer) gives us appropriate structure within any counterfactual supposition. There may or may not be principles governing counterfactual supposition across the board (Chapter 12). But even if there aren't, it's not the case that 'anything goes' within any counterfactual supposition. That's enough to convince us that the open worlds approach isn't trivial.

What of the worry, mentioned previously, that a model of belief with open worlds is trivial? This is one of the toughest problems to address, given that there seem to be no necessary closure conditions on what we can believe. Our answer (much as in the case of counterpossibles) is that a *model* may include important structure, even if the worlds it involves do not. We discuss the issue in §§10.3–10.4 and present a formal model in §10.5, based on some ideas discussed in §9.5.

There is one further worry the open worlds approach should address: that it results in disjunctive truth-conditions and implausible **(p.180)** consequences for a theory of meaning. (We touched on this issue in §4.3.) We'll discuss the worry and offer a response in §8.5.

8.5 The Compositionality Objection

Compositionality is the principle that the meaning of a complex expression is a function of the meanings of its constituent expressions. It's commonly taken to be a mandatory feature of any adequate theory of meaning. The argument is that, as competent speakers of a language, we are in principle capable of grasping the meanings of a potentially infinite number of sentences. And since we've learnt the meanings of a limited number of words, this is possible only if the meanings of complex sentences are obtainable recursively from the meanings of their constituent parts (Davidson 1965).

Take this very simple example. If you know what 'badgers are great' and 'and' mean in English, then you know what

> (8.7) Badgers are great, and badgers are great, and badgers are great, and badgers are great, and badgers are great, and badgers are great

means. Chances are, you never encountered that sentence before reading it here, yet you understood it straight away. How so? Because you understood all the component parts, of which the meaning of the whole is a function. That's compositionality.

The argument applies equally to notions of content: the content of a complex expression must be a function of the content of its constituents. The content of a conjunction $A \wedge B$, for example, must be a function of the contents of A and of B. The content of $A \vee B$ will be some other function, also of the contents of A and of B.

The possible worlds account of propositions (§8.3) clearly meets this requirement: if the contents of A and of B are sets of possible worlds, then the content of $A \wedge B$ is their intersection and the content of $A \vee B$ is their union. Since each pair of sets has a unique intersection **(p.181)** and union, this account of content (for the Boolean connectives, at least) is compositional.

In this example, we see a link between compositionality and truth-conditions. Given that ∧ requires the truth of both its conjunctions, the possible-worlds content of $A \wedge B$ must be all those worlds where both A and B are true. On any account of worlds, that's just the intersection of the A-worlds with the B-worlds. So here, there is a direct link between the connective's truth-condition and the function on contents. Similarly for the other truth-functional connectives.

This link breaks down with open worlds. At an open world, it is in general not the case that $A \wedge B$ is true just in case both A and B are true. There are open worlds where badgers are great but our badger sentence (8.7) is not true. So, when contents may include open worlds, we do not obtain the content of $A \wedge B$ by taking the intersection of the A-worlds with the B-worlds. Indeed, insofar as the truth of $A \wedge B$ is independent of the truth of A and of B at an open world, there appears to be no function from the open-world contents of A and B to the content of $A \wedge B$. But that's just to say that the account isn't compositional.

This is a serious worry for the open worlds approach. If it can't be met, then it could well be a fatal objection. (Note that it should be seen as a worry concerning meaning or content, not validity. For we analyse validity in terms of what's the case according to the *possible* worlds; and for possible worlds, the usual recursive clauses for the connectives hold.)

Our response to the problem is in two parts. First, we distinguish two aims for a formal theory. A formal theory might be (part of) a theory of meaning, in which case, it must be compositional. But a formal theory may be intended as a useful model of some notion (information, or belief, or logical consequence), without claiming to be a theory of meaning. In this sense, a model might get things extensionally correct, whilst not respecting the underlying mechanisms of *why* the modelled concept works they way it does. In short, the accounts we've presented might be fine for some tasks, but not for others, including giving a theory of meaning. (This kind **(p.182)** of attitude seems implicit in the work of those who have employed non-normal worlds for various logical purposes, which we examined in Chapters 4–7.)

The second, and less defensive, part of our response is to show that at least some open worlds-based approaches *are* compositional. This would show that a full theory of meaning can be given by including open worlds. To do this, we'll need to return to the metaphysics of worlds, which we discussed in Chapters 2 and 3.

Consider the various metaphysics of worlds we gave throughout those chapters, and what they say about truth-at-a-world. If w is a concrete world, then A is true at w iff things are such that A, once quantifiers are restricted to w. If w is an ersatz world, then truth-at-w might be a matter of which properties comprise (or are encoded by) w. Or it might be a matter of which propositions or worldmaking

sentences are members of w. Or it might be some other feature of w's construction.

We're going to focus on the linguistic ersatz approach, on which worlds are sets of sentences (§3.6). Such worlds represent *explicitly*, by containing a sentence that says such-and-such. These need not be sentences of the object language. Indeed, it's best they're not, else we make little progress in connecting the object language to reality. Let's write 'A^*' for the translation into the worldmaking language of the object language sentence 'A'. Truth-at-a-world is then defined in terms of world-membership:

(T_w) A is true at w iff $A^* \in w$

This definition applies to all worlds w and all sentences A, without regard for whether w is possible or impossible, open or not, and whether A is atomic or complex.

Next, let '$[\![A]\!]$' denote the content assigned to the object language sentence A: the set of all worlds at which A is true. This is the set of all worlds which contain A^*, $\{w \mid A^* \in w,\, w \in W\}$ (where W is the set of all worlds). But, given that any set of worldmaking sentences is a world in W, if we add A^* to every world in W, the set we obtain, **(p.183)** $\{w \cup \{A^*\} \mid w \in W\}$, is none other than $[\![A]\!]$. Moreover, these worlds all have A^*, and no sentence but A^*, in common. More precisely, their intersection, $\cap [\![A]\!]$, is none other than $\{A^*\}$.

Given these facts, we can always move back-and-forth between the content c of an object language sentence A and the corresponding worldmaking sentence A^*. So, as long as we can build up complex worldmaking sentences from more basic ones, in roughly the way we do for object language sentences, the semantics will be fully compositional. What follows is one way to make this precise.

Let's abstract from the precise grammatical details and suppose that complex worldmaking sentences are built using 1-place sentence operators, written prefix, and 2-place sentence operators, written infix. (These might include the Boolean connectives, various conditionals, and unary and binary modalities). Let O_1 be any such 1-place operator and O_2 be any such 2-place operator, and let $x^\frown y$ be the concatenation of worldmaking strings x and y. Then $O_1{}^\frown A^*$ and $A_1^*{}^\frown O_2{}^\frown A_2^*$ are worldmaking sentences.

Next, we define two semantic functions, f_1 (on the contents of a 1-place operator and a sentence, both from the object language) and f_2 (on the contents of a 2-place operator and two sentences):

$$f_1(o, c) = \left\{ w \cup \{ o^\frown x \mid x \in \bigcap c \} \mid w \in W \right\}$$

$$f_2(o, c_1, c_2) = \left\{ w \cup \{ x^\frown o^\frown y \mid x \in \bigcap c_1, y \in \bigcap c_2 \} \mid w \in W \right\}$$

To understand what's going on here, recall that, when c is the content of an object language sentence, $\cap c$ will be some singleton $\{A^*\}$. So the definition of f_1 sets x to be equal to some worldmaking sentence A^*, to which it concatenates o, and then returns the set of all worlds containing that complex worldmaking sentence. Let's write '$\neg^*$' for the translation into the worldmaking language of the object language negation '$\neg$'. Then $f_1(\neg^*, [\![A]\!])$ returns $[\![\neg A]\!]$, the set of worlds according to which $\neg A$ is true.

The same goes for f_2. Given an object language conjunction $A \wedge B$, we use f_2 on $\wedge^*$ (the worldmaking translation of '$\wedge$') and the contents of A and B. Then $f_2(\wedge^*, [\![A]\!], [\![B]\!])$ returns $[\![A \wedge B]\!]$, the content of **(p.184)** $A \wedge B$. That's all we need to demonstrate that the semantics is fully compositional.

How is it that this approach is compositional, when various semantics with open worlds presented throughout Chapters 4–7 are not? We gave those semantics in terms of a primitive valuation function or relation, assigning truth-values to sentences at worlds. For open worlds, the valuation (or relation) assigns truth-values to all sentences directly, and without restriction. That valuation is non-recursive. But the content assigned to a sentence depends on that valuation, and so this notion of content is non-recursive. Compositionality in the linguistic ersatz approach, by contrast, relies on a recursive translation from object language sentences to worldmaking sentences. The worldmaking sentence A^* does a lot of the heavy lifting in giving A's content. That's why the linguistic ersatz approach delivers a recursive characterization of the content of any object language sentence.

Chapter Summary

We began our discussion by asking whether hyperintensionality is a genuine phenomenon, or rather, a feature to be explained away (§8.1). We then focused on the epistemic case, considering arguments from Stalnaker and Lewis which attempt to explain away hyperintensionality (§8.2). We argued that they are not successful. We then considered the argument for a genuinely hyperintensional notion of content (§8.3). Having made the case for genuine hyperintensionality, we turned to the *granularity issue* (§8.4): how fine-grained are impossible worlds? This is one of the most difficult issues any theory of hyperintensionality faces. We then returned to the *compositionality objection*, and argued that some accounts of impossible worlds deliver a fully compositional theory of meaning (§8.5).