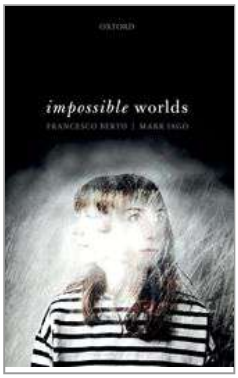

Impossible Worlds
Francesco Berto and Mark Jago

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

From Possible to Impossible Worlds

Francesco Berto
Mark Jago

DOI:10.1093/oso/9780198812791.003.0001

Abstract and Keywords

Possible worlds are ways things might have been. They find applications in analysing possibility and necessity; propositions; knowledge and belief; information; and indicative and counterfactual conditionals. But possible worlds semantics faces the issue of hyperintensionality, generated by concepts that require distinctions between logical or necessary equivalents. The problems of distinguishing equivalent propositions, of logical omniscience, of information overload, of irrelevant conditionals, and of counterpossible conditionals, are all instances of the general issue. Adding impossible worlds promises to help with these puzzles. But can we genuinely think about the impossible? It is argued that we can.

Keywords: hyperintensionality, possible worlds semantics, logical omniscience, impossible worlds, counterpossible conditionals

1.1 Worlds as Ways

Things might have been otherwise. David Bowie may still have been with us, the sun may have been shining on Nottingham, and the Axis powers may have won the Second World War. Such alternative ways we call *possible worlds*. Each possible world is a way things could have been. (This initial characterization says nothing of what possible worlds are, metaphysically speaking. That’s the topic of Chapters 2 and 3.) The actual world is the most general and comprehensive way in which things in fact are. In the actual world, the Nazis

lost the Second World War, the sky one of us sees from his office in Nottingham is cloudy, and David Bowie died at the beginning of 2016.

Ways things could have been can resemble the way things actually are. A world where the Axis powers won the Second World War is still a world where there was a war in which the Nazis fought, though with a different outcome from the actual world. Some possible worlds involve only small changes from ours: think of a world exactly like the actual one, except that you are one inch taller. Others are very different: think of one where the laws of biology and physics are turned upside down, so that you can be born twice, or travel faster than the speed of light. As we will see, the idea that it makes sense to speak of relations of similarity between possible worlds is important for some applications.

(p.12) Possible worlds have a *vast* array of applications. According to some, this is the main reason for accepting them: 'it may be that the best philosophical defence that one can give for possible worlds is to use them in the development of substantive theory' (Stalnaker 1991, 141). Since the late twentieth century rejection of the Quinean and Davidsonian idea that only extensional concepts should be allowed in serious philosophical inquiry, the notion of possible world has become ubiquitous in contemporary philosophy. It plays a key role in most branches of the discipline, ranging from logic to metaphysics and ontology, the philosophy of mind, the philosophy of information, moral and political philosophy, and aesthetics. But it has been used also outside of philosophy, in fields that range from the semantics of natural language to game theory, artificial intelligence, and cognitive science. We start with an overview of these applications. (Parts of the following section draw on Berto and Plebani 2015, chapter 11.)

1.2 Possible Worlds at Work

Possibility and Necessity

Perhaps the most typical application of possible worlds is in modal logic. This is, first of all, the logic of expressions like 'necessarily', 'possibly', 'contingently'. Such expressions are used in two different ways. A first use consists in qualifying the truth of a sentence, or of the proposition expressed by the sentence:

(1.1) It is necessary that 7 + 5 = 12.

(1.2) It is possible that Scotland leaves the UK.

(1.3) Possibly, Anna wins the music contest.

(1.4) Necessarily, Valeria is human.

Modalities of this kind are called *de dicto*. Expressions like 'necessarily' or 'it is possible that', or the concepts they express, are attached **(p.13)** to *dicta*, that

is, to pieces of language, or language-like entities, such as sentences or propositions. They express the way that sentence, or proposition, bears its truth value. Thus, according to (1.1), that seven plus five is twelve is necessarily true, and according to (1.3), that Anna wins the music contest is possibly true.

Modal expressions can also be used to qualify the features of objects:

(1.1a) Seven is necessarily an odd number.

(1.2a) Scotland is such that it could leave the UK.

(1.3a) Anna is a possible winner of the music contest.

(1.4a) Valeria is necessarily human.

Modalities of this kind are called *de re*, for the modals are used here to express the way in which a thing, a *res*, has some feature. Thus, according to (1.1a) and (1.4a), seven has the property of being odd, and Valeria that of being human, in a necessary way.

Contemporary logicians and philosophers follow Leibniz's insight that the necessary is what holds *no matter what*, in any way things could have been: that is, in all possible worlds. What is possible, on the other hand, is what holds at some possible world. What is contingent is what holds at some, but not all, possible worlds. Necessity and possibility are thus interpreted as quantifications over possible worlds. Using '$\Box$' for 'necessarily', '$\Diamond$' for 'possibly', 'iff' for 'if and only if', and letting W be the total set of possible worlds, we get:

'$\Box A$' is true at world w iff A is true at all worlds $w_1 \in W$

'$\Diamond A$' is true at world w iff A is true at some world $w_1 \in W$

(The notation '$w_1 \in W$' here means that w_1 is a member of the set W. It's a way of expressing that w_1 is a possible world.)

The two notions $\Box$ and $\Diamond$ are *duals* of one other, just as the universal and particular quantifiers, $\forall x$ and $\exists x$, are of one another. Each modal can be defined via the other and negation. That it is necessarily the **(p.14)** case that A means that it is not possible that $\neg A$ ('not-A'). And that it is possible that A means that it is not necessary that $\neg A$.

Necessity and possibility are highly ambiguous notions. (For a taxonomy, see chapter 1 of Divers 2002.) Although there is no universal consensus on this, many philosophers adopt three kinds of *absolute necessity*, holding in all possible worlds unrestrictedly:

LOGICAL NECESSITY fixed by the laws of logic broadly conceived (e.g., that if A, then either A or B);

MATHEMATICAL NECESSITY fixed by mathematical truths (e.g., that 7 + 5 = 12); and

METAPHYSICAL NECESSITY fixed by the identity and nature of things (e.g., that water is H_2O; that Valeria is a human being).

We will not get into the issue of whether one of these is reducible to another (e.g., the mathematical to the logical, as claimed by *logicists* in the philosophy of mathematics, including Dedekind (1901), Frege (1879), Peano (1889), and Russell (1903)).

We also talk of things being necessary, or impossible, only in a relative sense, or from a certain viewpoint. We are stuck in a traffic jam in Paris at 2 pm; our flight is leaving from De Gaulle airport at 2:10 pm. We moan: 'There's no way that we can make it to the airport in time'. What we mean is that, given the timing, the means of transport available, and the laws of physics of our world, it is impossible for us to reach the airport in time. It is not unrestrictedly, absolutely impossible: if we had *Star Trek*'s transporter, we could make it. But a *Star Trek* world in which one can be instantaneously disassembled into atoms and reassembled exactly with the same atomic structure in a different place is a world quite different from ours. One may doubt that such a world is even physically possible, that is, compatible with our laws of physics.

Other modal notions, thus, are naturally understood as restricted forms of necessity or possibility. Something can count as *R*-necessary, for some relativized modal notion *R*, even if it fails to hold at some **(p.15)** possible world or other. Accordingly, the corresponding modals are understood as restricted quantifiers over possible worlds. Thus *nomological necessity*, compliance with the laws of nature of the actual world or of the world under consideration, is often (no universal consensus here either) taken to be a relative or restricted necessity. It is biologically impossible but not absolutely impossible for a human being to jump one mile up in the air; it is physically impossible (if Einstein was right) but not absolutely impossible for a body to travel faster than the speed of light.

Propositions

Possible worlds are extremely important for theories of *representation*, both in language and thought, and have been used to analyse key notions from the philosophy of language. Many of these approaches build on Wittgenstein's insight that understanding the meaning of a sentence is grasping its truth conditions: 'to understand a proposition means to know what is the case if it is true' (Wittgenstein 1921/1922, §4.024). Montague (1970) and Stalnaker (1976a) have claimed that *propositions*, the meanings or contents expressed by sentences and the primary bearers of truth values, should be understood as sets of possible worlds. The proposition expressed in English by 'raccoons like to somersault' is, on that view, the set of possible worlds where raccoons like to

somersault, precisely the same set of possible worlds making for the proposition expressed in Italian by ‘ai procioni piace fare le capriole’.

Knowledge and Belief

Another notion analysed via possible worlds is *knowledge*. Following Hintikka (1962), knowledge has been characterized in terms of what is true throughout all the ways things could be, for all the agent in question knows. On this approach, the possible worlds accessible to an agent represent her *epistemic possibilities*. Knowledge can then be treated as another restricted quantifier over possible worlds. If K stands for a given agent’s state of knowledge, and R is a binary **(p.16)** accessibility relation on the space of worlds W, the Hintikka-style characterization goes thus:

(H) KA is true at w iff A is true at all w_1 such that Rww_1

This thought is at the core of contemporary epistemic logic (see, e.g., Blackburn et al. 2002, Fagin et al. 1995, Van Benthem 2003). But several research programs in mainstream epistemology also rely on a similar viewpoint. Dretske’s *relevant alternatives* approach takes knowledge as ‘an evidential state in which all relevant alternatives (to what is known) are eliminated’ (Dretske 1981, 367). Lewis (1996) discusses a similar approach. Alternatives here work similarly to possible worlds, and the uneliminated relevant alternatives work similarly to accessible worlds.

Necessity (whether logical, mathematical, metaphysical, or nomological) and knowledge share the feature of being *factive*: what is necessary, and what is known, is true. Factivity can be expressed by claiming that the actual world must always be one of the (accessible) possible ones, with respect to the relevant kind of possibility. Unlike other factive modalities, though, knowledge is an *intentional* state: a state of the mind directed towards a certain content. There are also non-factive intentional states, including belief, desire, fear, hope, and imagination. These have also been understood using restricted quantifiers over possible worlds, where the accessible worlds are the ones where things are as the agent believes (imagines, etc.) them to be. (Fagin et al. 1995 is a comprehensive guide through epistemic and doxastic logics. Niiniluoto (1985) and Wansing (2017) each discuss the application to imagination; Berto (2018) gives a semantics for imagination using an enriched possible worlds approach.)

Knowing (or believing, imagining, etc.) that A is often taken to be a mental state whose content is the proposition expressed by A. As well as being the primary bearers of truth values, thus, propositions have been understood as the content of (*de dicto*) intentional states: they are what is known, believed, feared, or imagined when one knows, believes, fears, or imagines something (*de dicto*). Just as different **(p.17)** sentences like ‘raccoons like to somersault’ and ‘ai procioni piace fare le capriole’ can share the same content, so can different

people's mental states share the same content: John believes, and Mary fears, that Marine will win the elections. There is some doubt that one and the same kind of entity can cover both the role of primary truth bearers and the role of targets of *de dicto* intentional states (Jago 2018b, Lewis 1986b). Nevertheless, possible worlds stake a good claim at giving a unified account of a broad range of linguistic and mental contents.

Information

Information is connected to knowledge, or potential knowledge. If a sentence or proposition is informative, then one can come to know that information (say, by hearing the sentence uttered truthfully by a trusted speaker). Something may be informative even if no one yet knows it, however. We might think of information as embodying potential knowledge for some suitable cognitive agent in the right circumstances. If we analyse knowledge in terms of possible worlds, we should expect a similar approach to information to be available.

According to the *Bar-Hillel-Carnap theory of information* (Bar-Hillel and Carnap 1953, Bar-Hillel 1964), the informative job of a sentence A consists in partitioning the totality of possible worlds into those where A is true and those where it is false. We may identify the information with the partitioning function, which in effect says 'yes' to some possible worlds and 'no' to all the others. Or we might identify the information with the set of 'yes' worlds. (Mathematically speaking, the former is the *characteristic function* of the latter set. The two approaches are, in a straightforward mathematical sense, equivalent.)

You might notice the similarity to the possible worlds account of propositions: both are treated as sets of possible worlds. That's no coincidence. On this approach, the information contained in a sentence (in a context) is precisely the proposition expressed by an utterance of it (in that context).

(p.18) This gives us a static notion of information, as something that's possessed by a sentence or proposition. The possible worlds approach also allows us to account for a dynamic notion of information, of *becoming informed* of such-and-such. When a cognitive agent gains the information (and, let's suppose, thereby learns) that raccoons like to somersault, we can model this in terms of ruling out the worlds where it is not the case that raccoons like to somersault. (Perhaps the raccoons of those worlds have different tastes; perhaps there are no raccoons there at all.) By 'rule out', we don't mean that the agent thereby treats those worlds as impossibilities. Rather, she rules them out as contenders for actuality: the ways things are, for all she knows.

Indicative Conditionals

Conditionality may also be dealt with using possible worlds. Many philosophers and logicians are unsatisfied with the material conditional, '$\supset$', taken as the operator given by the usual (two-valued) truth table: '$A \supset B$' is false when A is

true and B false, true otherwise. This delivers two inferences which have sometimes be called ‘paradoxes of the material conditional’ (Anderson and Belnap 1975, MacColl 1908, Routley et al. 1982):

(1.5) If $\neg A$, then $A \supset B$

(1.6) If B, then $A \supset B$

If we try to understand the English indicative conditional ‘if … then’ in terms of $\supset$, many seemingly false conditionals will come out true, just because their antecedent is false, or their consequent true:

(1.7) If Obama is Canadian, then the Moon is made of green cheese.

(1.8) If Strasbourg is in Germany, then Obama is American.

One reason to reject these is that there is no relevant connection between their antecedent and consequent: what’s Obama’s nationality got to do with the constitution of the moon, or the location of **(p.19)** European cities? Another reason to reject (1.7) and (1.8) is that any connection between their antecedent and consequent seems far too contingent. Even if Obama were Canadian, the moon would not be a giant cheeseball.

This suggests an alternative conception of the conditional on which, for ‘if A, then B’ to be true, it *cannot* be the case that A is true while B is false. This analysis gives us the *strict conditional*, ‘$\strictif$’. ‘$A \strictif B$’ is true just in case there is no possible world in which A is true but B is not. The strict conditional is the necessitation of the material conditional: $A \strictif B$ is understood as $\Box(A \supset B)$. It’s easy to see that (1.5) and (1.6) are invalid when we replace ‘$\supset$’ with ‘$\strictif$’. (Whether this move really avoids the worries is something we’ll come back to below, in §1.3 and Chapter 6).

Counterfactual Conditionals

Possible worlds have also been used to give a semantics for *counterfactual* conditionals. These are conditionals of the form ‘if it were (or, had been) the case that A, then it would be (or, have been) the case that B’, symbolized as ‘$A \boxright B$’. Counterfactuals are so-called because, in a typical use, they have a false antecedent, *contra factum*. In explaining why kangaroos have tails, for example, we might say ‘well, if kangaroos had no tails, they would topple over’ (Lewis 1973b).

(Many philosophers use ‘counterfactual’ for any conditional of the form, ‘if it were …, then it would be …’, even if the antecedent is true (Bennett 2003, Lewis 1973b, Williamson 2007). Others prefer to talk of ‘subjunctive conditionals’. We’ll will stick to the standard ‘counterfactuals’ terminology for all such conditionals.)

Counterfactuals are extremely important in our cognitive lives. We conceive counterfactual alternatives to reality in order to explore what would or would not happen, were those alternatives realized. Would John not have been injured, had he avoided crossing the road? They are also important in understanding history (Nolan 2016b): what if Hitler had had the A-bomb in 1944? They may help us to understand the concept of causation better (Lewis 1973a, Paul 2004; **(p.20)** see Paul 2009, Paul and Hall 2013 for in-depth discussion). So it's important to give a semantic analysis of counterfactuals.

How? That counterfactuals must be modal conditionals can be argued by comparing them to the corresponding indicative conditionals:

(1.9) If Kate Bush didn't write 'The Kick Inside', someone else did.

(1.10) If Kate Bush hadn't written 'The Kick Inside', someone else would have.

These have the same antecedent and consequent (in different moods), but different truth values. (1.9) seems true. We know that someone wrote 'The Kick Inside', so if it wasn't Kate Bush, it must have been someone else. By contrast, (1.10) seems false: 'The Kick Inside' might never have been written, if it hadn't been for Kate Bush. So even if one insists that (1.9) be taken as a material conditional, (1.10) seems to be of a different kind. The difference in mood between (1.9) and (1.10) has been understood as getting us to evaluate (1.10) by looking at alternative ways things could have been, that is, at alternative possible worlds.

Which worlds? The mainstream treatment of counterfactuals, due to Stalnaker (1968) and Lewis (1973b), says that we should evaluate 'if kangaroos had no tails, they would topple over' by looking to the *closest* possible worlds where kangaroos have no tails. We then see whether kangaroos topple over there. *Closeness* between worlds is understood as involving (contextually determined) similarity in the relevant respects. So evaluating a counterfactual will typically involve the *minimal change* (with respect to the world of evaluation) required to verify the antecedent. We disregard worlds where kangaroos have no tails but help themselves with crutches, or have evolved wings. Overall: '$A \mathrel{\square\!\!\rightarrow} B$' is true (at world w) iff the closest(-to-w) possible A-worlds are B-worlds.

(What if several possible worlds tie for closeness? Do we require *all* closest A-worlds to be B-worlds? Or some? Or most? What if the A-worlds get forever closer and closer to ours, with none being the **(p.21)** closest? These are tricky questions: Kratzer (1981), Lewis (1973b, 1981), and Nute (1975) discuss them in detail. We won't get into them here.)

Possible worlds have also been used in the analyses of *essence* and *de re modality* (Lewis 1986b), and of *metaphysical dependence* and *supervenience* (Bennett 2004, Davidson 1970). Many physicalist philosophers of mind, including Horgan (1982, 1993), Kim (1982, 1993), and Lewis (1983), express their commitment to physicalism about mental states in terms of supervenience, cashed out in terms of possible worlds. But, for reasons we won't go into here, we don't think a worlds-based approach is the best way to capture notions of essence or dependence. (We're largely persuaded by Fine's (1994) arguments.) So we won't discuss these applications any further.

Possible worlds are a success story of philosophical theorizing. Still, most of the accounts using them, which we have just sketched, face issues. The umbrella under which many of these can be gathered is the concept of *hyperintensionality*, to which we now turn.

1.3 The Problem(s) of Hyperintensionality

Hyperintensionality can be characterized as a feature of concepts. A concept is hyperintensional when it draws a distinction between necessarily equivalent contents, where the relevant necessity is unrestricted: logical, mathematical, or metaphysical, if we stick to the threefold distinction mentioned above. If the relevant concept is expressed by an operator $\mathscr{H}$, then $\mathscr{H}$ is hyperintensional when $\mathscr{H}A$ and $\mathscr{H}B$ can differ in truth value, in the face of A and B's being necessarily (logically, mathematically, or metaphysically) equivalent.

(Cresswell (1975) originally defined 'hyperintensional' to pick out a position in a sentence in which logical equivalents cannot be replaced *salva veritate*. But, as Nolan (2014, 151) notes, it is now common to use the term more broadly, with 'necessary equivalence' in place of 'logical equivalence'.)

(p.22) This characterization of hyperintensionality is a contrastive one. It tells us that a concept or operator is hyperintensional when it is more fine-grained than intensional or (normal) modal concepts or operators, marking a distinction invisible to the latter. It does not yet provide us with a full-fledged characterization of hyperintensionality, and it says nothing about 'just "how hyper" hyperintensions are' (Jespersen and Duzi 2015, 527). Different hyperintensional notions may display different degrees of fine-grainedness. We discuss this key issue in Chapter 8.

The problems we are about to examine affect the possible worlds accounts introduced in §1.2. The problems emerged over the last few decades in piecemeal fashion. But a single issue underlies them all: they are hyperintensional notions, making distinctions more fine-grained than the standard possible worlds approach can easily model.

Propositions: Triviality

If we take propositions, the meanings or content of sentences, as sets of possible worlds, then necessarily equivalent propositions are one and the same proposition: possible worlds never disagree on necessarily equivalent sentences. Assuming again that mathematical and logical necessity are unrestricted, ‘if Obama is human, then Obama is human’ and ‘7 + 5 = 12’ are true in the same possible worlds: all of them. So they express the same proposition, viz., the total set of worlds.

This seems wrong: the sentences should have different meanings. They speak of different things: only one is about Obama. We have a dual problem with sentences that cannot be true, like ‘Obama is both human and not human’ and ‘7 + 5 = 13’. These would also express the same possible-worlds proposition: the empty set of possible worlds. This seems just as bad a result as the first: the two sentences have different meanings and are about very different things: the first is about Obama, the second is not.

This problem is particularly evident when we turn to the kinds of propositions typically expressed when we do metaphysics. Many metaphysical claims are such that, if they are true, or false, they are **(p.23)** necessarily so. This includes claims of modal metaphysics, such as statements about the very nature of possible worlds. If we say that possible worlds have such-and-such natures, then we seem committed to that claim being necessarily true. After all, if it were possibly false, then it would be false at some possible world, which seems to make little sense. (Just how powerful this argument is depends on how we take worlds to represent a particular state of affairs: we’ll discuss this issue in chapters 2 and 3.)

Many other metaphysical claims seem to be necessary (if true at all): Hegel’s doctrine of the Absolute *Geist*, Plato’s view of ideas as purely intelligible forms, and Armstrong’s claim that there are immanent multiply instantiated universals, do not seem to be contingent claims. Defenders and objectors of these views alike agree that these are distinct viewpoints, expressed by distinct propositions. If that’s right, then each view corresponds either to the set of all worlds (if true), or to the empty set (if false). But since these are three distinct views, expressed through three distinct propositions, those propositions are not plain sets of possible worlds.

Here’s a further puzzle to bring out the problem. Suppose Anna and Valeria are debating the nature of properties. Anna says (*P*) that they’re transcendent Platonic universals, whereas Valeria says (*I*) that they’re immanent universals. Each view corresponds either to the set of all worlds (if true), or to the empty set (if false). Suppose further than both Anna and Valeria believe that propositions are sets of possible world. Then Anna must accept that her claim is identical to

the claim that $P \vee I$ (since she believes P is necessarily true, I is necessarily false, and hence that P and $P \vee I$ each correspond to the set of all worlds).

Similarly, Valeria must accept that her claim is identical to the claim that $P \vee I$ (since she believes that P is necessarily false, that I is necessarily true, and hence that I and $P \vee I$ each correspond to the set of all worlds). But if each accepts that their claim is identical to the claim that $P \vee I$, they must accept that their claims are identical, which neither will accept. If they are to have a serious debate about the nature of properties, therefore, they should reject **(p.24)** their beliefs that propositions are sets of possible worlds. Genuine, rational metaphysical debate is possible only on the assumption that propositions are not sets of possible worlds.

Knowledge and Belief: Logical Omniscience

Historically, one of the first manifestations of the hyperintensionality issue came from the modal treatment of epistemic and doxastic concepts. Here, the issue is *logical omniscience*: a cluster of closure conditions on knowledge and belief, which come as a spin-off of Hintikka's (1962) possible-worlds approach (§1.2). Perhaps the most important closure effects of Hintikka's clause (H) above are:

(C1) If KA and A entails B, then KB

(C2) If A is valid, then KA

(C3) It is not the case that: KA and $K\neg A$

and similarly for belief. (We will find a more comprehensive list of closure conditions in §5.1.)

(C1), often dubbed *Closure under entailment* or *Full omniscience*, says that one knows all the entailments or logical consequences of what one knows. The principle also applies to the possible-worlds semantics for belief: one believes all the logical consequences of what one believes. (C2), *Knowledge of all valid formulas*, says that one knows all the logical truth (and similarly for beliefs). When we define validity as entailment by the null set of premises, (C2) is a special case of (C1). (C3) guarantees *Consistency* of knowledge: one can never have inconsistent knowledge, and the corresponding principle says one can never have inconsistent beliefs.

These conditions follow directly from interpreting the relevant epistemic notions as restricted quantifiers over possible worlds. For instance, (C1) holds once we understand A's entailing B as the claim that B is true at all possible worlds (of all models of the epistemic logic at issue) where A is true. Then, if A is known (believed), it holds at all the epistemically accessible possible worlds. But if A entails B, **(p.25)** then B holds at all those worlds too, and so B is known (believed) as well. (C2) holds when we understand the logical validity of A as its

holding in all possible worlds (of all models, etc.). Then, in particular, a valid A holds at all the epistemically accessible worlds, and so is automatically known (believed).

For applications in computer science, such principles are often taken as harmless (Fagin et al. 1995, chapter 9). However, it is generally admitted that they deliver implausibly idealized notions of knowledge and belief, having little to do with *human* cognition. Against (C1), for instance: we know basic arithmetic truths like Peano's postulates, and these entail (let us suppose) Goldbach's Conjecture; but we don't know whether Goldbach's Conjecture is true. Against (C2): Excluded Middle is (let us suppose) valid, but intuitionist logicians do not believe it, and so do not know it either.

As for (C3): there cannot be inconsistent knowledge, given that knowledge is factive and assuming there are no true contradictions. But real, finite, and fallible cognitive agents may well have inconsistent beliefs. They may even believe the relevant inconsistencies explicitly, and take themselves as justified in doing so (e.g., *dialetheists* believe that the Liar sentence is both true and false (Priest 1987)).

An answer one sometimes hears is that K in (H) expresses not knowledge or belief, but rather some derivative attitude, characterized in terms of knowledge or belief: what an agent is logically committed to, given what else she knows or believes. This leaves us in want of a logical account of *knowledge* and *belief* for real agents, as opposed to some conditional commitment. One may also question this account of epistemic or doxastic commitment. Is an intuitionistic logician really committed to Excluded Middle (given classical logic)? Are those of us with inconsistent beliefs – all of us! – really committed to everything being true, given that a contradiction classically implies every sentence?

Information: Triviality and Overload

The possible worlds Bar-Hillel-Carnap analysis of information has similar issues to the account of knowledge and belief. 'If Obama is **(p.26)** human, then Obama is human' and '$x^n + y^n = z^n$ has no integer solutions for $n > 2$' are both necessarily true. So there is no possible world ruled out by learning either. On the Bar-Hillel-Carnap analysis, neither are genuinely informative, and so neither are learnable. But while the former is easily deemed true by competent speakers of English, the truth of the latter is non-trivial in the extreme. For the latter is Fermat's Last Theorem, a proof of which took centuries to find. The first, by Andrew Wiles, was 130-something pages long.

The problem generalizes. A possible worlds analysis of information entails that *no* logical, mathematical, or metaphysical truth can be informative. It denies, in particular, the informativeness of any logical deduction or mathematical proof, and thus the epistemic value of devoting one's time to the study of mathematics

or logic. But some deductions and proofs are obviously informative. This can depend on the fact that the conclusion has high syntactic or semantic complexity, but it need not be so. Fermat's Last Theorem is expressed by a sentence anyone with high school maths can understand. But recognition of its truth, via proof, is extremely complicated.

Even simple proofs, like short truth table calculations, can be informative. Students who have just mastered the truth table for the material conditional may be surprised to find out that Frege's Law, $(A \supset (B \supset C)) \supset ((A \supset B) \supset (A \supset C))$, is a tautology, or that for all A and B, either $A \supset B$ or $B \supset A$. It is part of the explanation of why they are surprised, that they acquire new information. It seems, then, that there is a legitimate notion of information whereby one *can* learn, or become informed of, a tautology.

Indicative Conditionals: Irrelevance

The possible worlds treatment of conditionality is not free from problems either. We have seen that the strict conditional '$A \dashv B$' is free from the paradoxes of the material conditional. But it has its own so-called 'paradoxes of the strict conditional':

(1.11) If $\neg\Diamond A$, then $A \dashv B$

(p.27) (1.12) If $\Box B$, then $A \dashv B$

If B is true in all possible worlds, or A in none, then there is no possible world where A is true and B is false, so '$A \dashv B$' is true too. Interpreting the 'if …, then …' of English as the strict conditional, this makes many seemingly false conditionals true, just because their antecedent is impossible, or their consequent is necessary:

(1.13) If 5 + 7 = 13, then Obama is Canadian.

(1.14) If Obama is American, then 5 + 7 = 12.

These look bad because of the irrelevance phenomenon. There seems to be no connection between the antecedent and consequent: they are about wholly distinct things. Given the necessity of logical truth and the impossibility of logical falsity, we also get true strict conditionals whose consequent is a truth of logic, or whose antecedent is a falsity of logic, e.g., of the following form:

(1.15) $A \dashv (B \dashv B)$

(1.16) $A \dashv (B \vee \neg B)$

(1.17) $(A \wedge \neg A) \dashv B$

These also look bad due to irrelevance: what A is about may have nothing to do with what B is about. Take an instance of (1.16), 'if the Moon is made of green cheese, then either Nottingham is in Scotland, or not'. Does that sound correct? (We will come back to this kind of irrelevance phenomenon in Chapter 6.)

Counterfactual Conditionals: Counterpossibles

Counterfactuals with impossible antecedents are called *counterpossibles*. The Lewis-Stalnaker treatment of counterfactuals delivers *vacuism*: the view that all counterpossibles are vacuously true. If '$A \mathrel{\square\!\!\rightarrow} B$' is true when all the closest A-worlds are B-worlds, and there are no A-worlds, then it comes out automatically true. (Just as **(p.28)** 'all hobbits in this room are tiny' is true, trivially, given that there are no hobbits in the room.) So a counterfactual '$A \mathrel{\square\!\!\rightarrow} B$' will be trivially true whenever its antecedent is impossible. To add insult to injury, the conditional with the same antecedent and negated consequent, '$A \mathrel{\square\!\!\rightarrow} \neg B$', will also be trivially true.

Some philosophers believe that, appearances to the contrary notwithstanding, this is all right (we discuss this kind of view in Chapter 12). However, many – including Nolan (1997), Brogaard and Salerno (2013), Priest (2008), Krakauer (2012), Bjerring (2014), and Bernstein (2016) – think these results to be problematic. Nolan (1997) gives a nice example:

> (1.18) If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would have cared.
>
> (1.19) If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would not have cared.

Hobbes' squaring the circle would have made absolutely no difference to the suffering of those sick children. So (1.19) should come out true true for *this* reason, and not merely because there are no worlds verifying the antecedent. Similarly, (1.18) should come out false. Some counterpossibles are false; and where they are true, typically, they are not trivially so.

The problem connects in an obvious way to the triviality problem for possible worlds propositions. We often reason counterfactually in matters of mathematics, logic, and metaphysics. Vacuism about counterpossibles can hardly account for this. We make counterfactual suppositions in all of these areas, perhaps with the purpose of criticizing a theory by drawing unpalatable consequences. Such a practice is trivialized if counterpossibles are all true. Imagine Hilbert arguing against Brouwer that, if intuitionism were true, then much of standard mathematics would be lost, for we could not then resort to impredicative definitions. Vacuism makes any such claim devoid of dialectical content. For given vacuism (and classical logic), we **(p.29)** can also truthfully assert that, if intuitionism were true, nothing of standard mathematics is lost.

Indeed, this claim would be equivalent to Hilbert's; yet it is what Hilbert wanted to deny.

Taken together, these problems provide a strong case against the possible worlds approach. It might even be that they provide reason to abandon all attempts to analyse these notions in terms of worlds (Fine 1975a, 2012a, 2019).

However, we needn't abandon a worlds-based analysis of these concepts. The problems we've just sketched show that we can't give good enough analyses using only *possible* worlds. We have to be more open minded. We can give good worlds-based analyses of these concepts, so long as the worlds in question include possible *and impossible* worlds. Using impossible worlds, we can solve a number of problems faced by possible worlds accounts of knowledge, belief, meaning, information, and conditionality.

1.4 Impossible Worlds

This book is concerned with worlds that are not possible, with 'possible' understood in an unrestricted sense. You might worry that impossible worlds are metaphysically weird, logically disreputable, or not really useful for this or that purpose; or that they lose some crucial benefit of possible worlds accounts. We'll argue that it ain't so. In Chapters 2 and 3, we'll show that impossible worlds can be metaphysically acceptable even if, as we'll discuss there, some attempts to make them metaphysically reputable fail. In Chapters 4–7, we'll show how impossible worlds have useful logical applications, and that they may, but need not, involve a departure from classical logic. We'll discuss philosophical applications of impossible worlds in Chapters 8–12. As we go, we'll address some of the objections that have been raised against impossible worlds, including, for example, that they don't allow for a compositional account of meaning (Chapter 8).

(p.30) The possible worlds framework seems still to be a dominant conceptual framework of our time for philosophical theorizing. We'll argue that the impossible worlds framework constitutes a net theoretical gain. The late twentieth century saw an intensional revolution centred on the notion of possible world. The early twenty-first century is seeing what Nolan (2014) called a *hyperintensional revolution*. Impossible worlds are at home in this revolution.

They aren't the only theoretical tool that's been suggested for analysing hyperintensionality. An alternative is the *structured propositions* approach of King (1995, 1996, 2007), Soames (1985, 1987), and others. Another one comes from Pavel Tichy's *transparent intensional logic* (Duzi et al. 2010). Other recent approaches include Fine's *truthmaker semantics* (Fine 2012a, 2014, 2019) and Yablo (2014)'s work on *aboutness*, which enriches possible worlds semantics with divisions of the space of worlds itself. We shan't discuss these here in any detail. We've discussed structured propositions elsewhere Jago (2014a, 2015, 2017). Whatever its merits in accounts of content, it cannot claim to be a general

account of hyperintensional logical or philosophical notions. Ripley (2012) compares the impossible worlds and structured propositions approaches to hyperintensionality, coming down forcefully on the side of the former.

(Truthmaker semantics is an exciting recent development. As a general philosophical approach to hyperintensionality, it is at present underdeveloped, but has great potential. We don't think it can be a general approach to hyperintensional notions. It seems that any truthmaker involving James Newell Osterberg is thereby a truthmaker involving Iggy Pop, since Osterberg is Iggy Pop. But one can believe that Iggy Pop co-wrote David Bowie's 'China Girl' without believing that Osterberg did. So we don't see how epistemic or doxastic contents could be modelled using truthmaker semantics.)

Our aim is to investigate, develop, and defend the way of impossible worlds. Of the other ways, we won't say much. (For an assessment of the relative merits of the structured propositions approach, aboutness à la Yablo, truthmaking à la Fine, and impossible worlds, in the treatment of hyperintensionality, see Gioulatou (2016).) So let us turn **(p.31)** to the obvious question: what's an impossible world? (The following material draws on Berto and Jago 2018.)

A look at the literature on impossible worlds (which is rapidly growing: see Nolan 2013 for a survey) presents us with a number of different definitions. These can be reduced to four main ways of treating impossible worlds, ordered from the more to the less general:

IMPOSSIBLE WAYS: Just as possible worlds are characterized as ways things could have been, so are impossible worlds often characterized as ways things could *not* have been. The initial insight is that not everything is possible. Some things just (absolutely) cannot happen. Anything that just can't happen must be an absolute impossibility; and these ways the world just couldn't be are impossible worlds. Beall and van Fraassen (2003), Restall (1997),Salmon (1984), and Yagisawa (1988) think of impossible worlds in this way.

LOGIC VIOLATORS: Another definition has it that impossible worlds are worlds where the laws of logic fail. This approach depends on what we take the laws of logic to be. Given some logic L, an impossible world with respect to the L-laws is one in which some of those laws fail to hold (Priest 2008, chapter 9). An impossible world in this second sense will also be impossible in the first sense, so long as the logic L in question is no stronger than the logic governing logical possibility. But for dialethists or intuitionists, a world violating containing contradictions, or failing excluded middle, won't count as a way the world couldn't be, and so won't count as an impossible world in the first sense. Whichever logic is operative, there are worlds which count as impossible in the first but not in this second sense. If the Continuum Hypothesis of set theory is true (and, logicists are wrong!), some world where the Continuum Hypothesis

fails may well be impossible in the first sense without thereby violating any law of the chosen logic.

CLASSICAL LOGIC VIOLATORS: Another definition has it that impossible worlds are worlds where the laws of classical logic fail (Priest 1997a). **(p.32)** This definition gives the same results as the previous one if we take the laws of logic to be the classical ones, but not otherwise. A world complying with intuitionistic logic, but where instances of Excluded Middle fail, will be impossible in this third sense.

CONTRADICTION-REALIZERS: A still narrower definition has it that an impossible world is one where sentences of the form A and $\neg A$ hold, against the Law of Non-Contradiction (Lycan 1994). Impossible worlds of the fourth kind will be impossible in the third sense, since they thereby violate classical logic. But not vice versa: an intuitionistic world will have the Law of Non-Contradiction hold unrestrictedly, and so will be impossible in the third, but not the fourth, sense.

Talk of impossible worlds as ways things could (absolutely) not have been might suggest that these worlds are, themselves, impossible objects. An impossible object is an object which could not possibly exist and so does not in fact exist. Yet defenders of impossible worlds claim that they do in fact exist. (Or rather, most of their defenders do. Those who don't have a different view of what existence is. We will discuss the issue of the existence of impossible worlds in §2.3.)

This isn't an issue for impossible worlds only. By the same reasoning, we could say: possible worlds (other than the actual world) are merely possible objects and so not actually existing objects. And yet their defenders say that they do in fact actually exist. (Or rather, most of their defenders do. Those who don't have a different view of what actuality is. We will discuss the *genuine realist* approach, on which possible worlds exist but may not actually exist, in §2.2.)

For now, it will suffice to stick to an analogy. Assume that some Escher drawings represent impossible situations. This does not make them impossible. They are not merely possible entities either: they really, actually exist. So actual entities can represent impossible situations. A core part of our investigation into impossible worlds will concern how they manage to represent the situations they represent. (Don't take the Escher drawing analogy too far: we don't want to claim that worlds represent *pictorially*, in the way pictures do.)

(p.33) The question of how worlds (possible and impossible) represent what they represent is tied up with the question of what they are, metaphysically speaking. This question will occupy Chapters 2 and 3. We'll investigate the issue by first looking at what possible worlds are, metaphysically speaking. Of each of

the plausible options, we then ask whether it may be extended to account for impossible worlds.

Some readers may discern no serious issue here. Some modal logicians take the instrumentalist line, on which the set of worlds W may be any old bunch of objects with some relations between them. Anything that does the job will do. This seems philosophically unsatisfactory, just as it seems unsatisfactory to talk of moral properties, or abstract universals, or truth, and yet to refuse to consider their nature. If it is good to understand various concepts in terms of worlds, possible or impossible, then we want to know why this is so. It is difficult to answer the question without saying something about what kind of things worlds are. (Of course, it's often fine to work with worlds without considering their nature, if one is merely postponing, rather than forever refusing to answer, that question.)

Before we get to the logical and philosophical applications of impossible worlds (Parts II and III), therefore, we will investigate the metaphysics of worlds. But, before we get to the metaphysics of worlds and the issue of how worlds represent impossibilities, we should ask whether *we* can represent impossibilities. For if we can't, there is less work to do for impossible worlds in logic and philosophy.

1.5 Conceivability and Possibility

Hyperintensionality is typically thought to involve representational contexts. Impossible worlds have a role to play, first of all, in modelling representational mental states, or thoughts, whose hyperintensional nature is tied to the fact that their content involves absolute impossibilities in some way or other.

(p.34) But can we actually think about the impossible? Can we have mental representations – intentional states of the mind – directed to impossible contents? A venerable philosophical tradition denies this. Hume is the most quoted authority:

> 'Tis an establish'd maxim in metaphysics, that whatever the mind clearly conceives includes the idea of possible existence, or in other words, that nothing we imagine is absolutely impossible.
>
> (Hume 1739/1978, I, ii, 2)

We think that Hume's maxim is wrong (as do Byrne (2007), Fiocco (2007), Kung (2010), and Priest (2016a), among others). Arguing for this requires us to say something about conceivability and imagination. These are highly ambiguous notions. One way to clarify them consists in asking how mental representations in general represent, and looking at answers provided by cognitive psychologists. (We now follow Berto and Schoonen 2018.) The literature presents two main candidate codings for mental representations: the *linguistic*

and the *pictorial*, the difference between the two consisting in the degree of arbitrariness of the representation relation (Paivio 1986).

Pictorial mental representations are gathered under the rubric of ‘mental imagery’, and characterized by reference to sensory perception. They are ‘quasi-perceptual experiences’ (Thomas 2014, Introduction), for they resemble perceptual representation, but can occur in the absence of the actual stimuli. Studies on neuroimaging, such as Ganis et al. 2004, seem to show that visual mental imagery (the most studied kind of imagery) activates about 90% of the same cerebral areas activated by visual perception, though the interpretation of such results is somewhat controversial.

Visual mental imagery is often claimed to have spatial or quasi-spatial features. When we entertain imagery of this kind, we represent objects and situations typically in three-dimensional egocentric space. These representations are available for ‘parallel processing’ because they have some kind of mereological structure (Paivio 1986, 198). You can represent to yourself in this way the arrangement of your living room and describe its contents from different viewpoints, mentally **(p.35)** scanning the objects included there from top to bottom or from left to right; you can mentally zoom into a corner, and so on. Of course, psychologists who work on mental images do not claim that they are real pictures, hence the frequent use of the ‘quasi-’ prefix. The claim that the parts of a pictorial mental representation correspond to the parts of the represented scenario, with the relative distances respected, comes with the proviso that ‘part’ and ‘distance’ should be understood functionally rather than spatially (see e.g. Kosslyn and Pomerantz 1977).

Linguistic mental representation, by contrast, is arbitrary in the same way that the connection between words and what they mean is arbitrary. Such representations are called ‘amodal’ to stress that they are disconnected from sensory modalities in a way pictorial representations are not. According to Paivio (1986, 198), linguistic mental representations are processed serially, the way we process the meanings of sentences through their subsentential components. This is taken as evidence that linguistic representations lack the mereological and quasi-spatial features of (visual) pictorial ones.

Paivio’s *dual coding* theory has it that there are precisely two codes for mental representations: the linguistic and the pictorial. Cognition works with two functionally independent (though interacting) systems handling representations of the two kinds. The usefulness of having two systems, according to some, lies in the different contents the two are apt to represent: pictorial imagery is more suitable for concrete situations which are proximal in space and time, whereas linguistic representation works better for abstract scenarios involving non-perceptual features (Amit et al. 2009).

(Some psychologists, including Pylyshyn (1973, 2002), think that there is really just one kind of mental representation. They attempt to reduce the pictorial to the linguistic. This involves the ‘imagery debate’ or ‘analog/propositional debate’, to which we return in Chapter 7.)

We argued in Berto and Schoonen (2018) that, if mental representations involved in conceivability represent linguistically, then Hume’s maxim cannot even get off the ground. If we make the plausible assumption that linguistic mental representations have at least the **(p.36)** same representational power as the expressions of natural languages like English, then of course we can conceive, by linguistically mentally representing it, the impossible. Logically impossible sentences of ordinary English can be perfectly meaningful.

Quine (1948) argues that contradictions can be meaningful. He makes the point as a response to his fictional philosopher Wyman, sometimes taken as representing Meinong’s view (to which we will come back in §2.3) that some things do not exist. Wyman believes that things like Pegasus ought to be admitted in our ontological catalogue, as *possibilia*, for otherwise it would make no sense to say that Pegasus is not. By parity of reasoning, says Quine, we ought to admit the round square cupola on Berkeley College; otherwise, it would make no sense to even say that *it* is not. But accepting this brings inconsistency. Wyman reacts by declaring that inconsistent conditions are meaningless. We find Quine’s reply spotless:

> Certainly the doctrine [that contradictions are meaningless] has no intrinsic appeal; and it has led its devotees to such quixotic extremes as that of challenging the method of proof by *reductio ad absurdum* – a challenge in which I sense a *reductio ad absurdum* of the doctrine itself.
>
> Moreover, the doctrine of meaninglessness of contradictions has the severe methodological drawback that it makes it impossible, in principle, ever to devise an effective test of what is meaningful and what is not. It would be forever impossible for us to devise systematic ways of deciding whether a string of signs made sense – even to us individually, let alone other people – or not. For it follows from a discovery in mathematical logic, due to Church (1936), that there can be no generally applicable test of contradictoriness.
>
> (Quine 1948, 34–5)

Graham Priest, a friend of true contradictions, agrees (for once!) with Quine:

> If contradictions had no content, there would be nothing to disagree with when someone uttered one, which there (usually) is. Contradictions do, after all, have meaning. If they did not, we could not even understand someone who asserted a **(p.37)** contradiction, and so evaluate what they say as false (or maybe true). We might not understand what could have

> brought a person to assert such a thing, but that is a different matter and the same is equally true of someone who, in broad daylight, asserts the clearly meaningful ‘It is night’.
>
> (Priest 1998, 417)

Now suppose that, instead, there are irreducibly pictorial mental representations. Then the question for supporters of Hume’s maxim is: does pictorial imagination work *purely* pictorially, or not? Does the relevant mental imagery represent a situation without any language-like arbitrary assignment of meaning, but just via the phenomenological similarity of the imagery to the worldly situation?

It is controversial whether mental representation can *ever* work purely pictorially. Fodor (1975) argues that pictorial mental representation has a role in cognition only insofar as it works as ‘imagery under description’, that is, insofar as the imagery comes endowed with linguistic labels: linguistic mental representations pinning down what the image is about. If so, then the arbitrariness of the relevant linguistic labels allows us to imagine the impossible.

Kung (2010) argues that this stipulative labelling component gives pictorial imagination its power to represent the impossible: for example, by stipulating the identity of the imagined objects. Imagine Tim kissing John. The phenomenology of the mental imagery can be such that the represented figures are relevantly similar to Tim and John: hair colour, eyes, bodies. But what makes the imagining count as a representation of a scenario in which *Tim* kisses *John* is that one takes one figure as representing *Tim* and the other as representing *John*. And just as one can imagine Tim kissing John (a possible scenario), so can one imagine Tim as a cleverly disguised robot. One labels the imagined person-lookalike, which turns out to be filled with circuits and transistors, as *Tim*. But Tim (suppose) is essentially human, so this scenario is metaphysically impossible.

What if we do have pictorial mental imagery that represents purely pictorially? Then it may be that scenarios imagined in this way must be possible. But mental imagery of this kind would be quite limited in scope (Berto and Schoonen 2018). Some labelling seems to be needed **(p.38)** whenever perceptual experience has a content that goes beyond mere shapes and colours (Siegel 2006, Siewert 1998). You see a face and a nose, rather than merely face-like and nose-like shapes. Your experience comes labelled: the nose-like shape as a *nose* and the face-like shape as a *face*.

Purely pictorial mental imagery on its own would be limited, in particular, as a tool of modal epistemology. Philosophers discuss whether the imaginability of intrinsic universals, time travel, or a supreme being having all perfections to the highest degree entails their absolute possibility. (Van Inwagen (1998) doubts

that we can imagine these things; see Hawke 2011 for a discussion.) Imagination here cannot be purely pictorial, for it involves abstract objects and properties far removed from sensory perception. In these debates, ‘imagination’ seems to be understood more broadly than the purely pictorial characterization allows.

If, finally, mental representations are taken to be neither linguistic nor pictorial, this leaves the supporters of Hume’s maxim with a heavy burden of proof. They seem forced to invoke a peculiar ‘third code’ of representation, with no counterpart in general theories of representation. It’s then up to Humeans to provide a plausible theory of how that notion works. Absent a workable theory, the approach is relying on representational magic. In short, we have good reason to hold that we can mentally represent absolute impossibilities.

Chapter Summary

Possible worlds are ways things might have been (§1.1). They find applications in analysing possibility and necessity; propositions; knowledge and belief; information; and indicative and counterfactual conditionals (§1.2). But possible worlds semantics faces the issue of hyperintensionality, generated by concepts that require distinctions between logical or necessary equivalents. The problems of distinguishing equivalent propositions, of logical omniscience, of information overload, of irrelevant conditionals, and of counterpossible **(p.39)** conditionals, are all instances of the general issue (§1.3). Adding impossible worlds promises to help with these puzzles (§1.4). But can we genuinely think about the impossible? We argued that we can (§1.5). **(p.40)**