HYPERINTENSIONALITY AND SYNONYMY

A logical, philosophical, and cognitive investigation

MSc Thesis *(Afstudeerscriptie)*

written by

Levin Hornischer
(born October 17th, 1991 in Filderstadt, Germany)

under the supervision of **Prof. Dr. Franz Berto** and **Prof. DDr. Hannes Leitgeb**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
June 23rd, 2017	Dr. Maria Aloni
	Prof. Dr. Franz Berto
	Dr. Luca Incurvati
	Prof. DDr. Hannes Leitgeb
	Prof. Dr. Benedikt Löwe (*chair*)
	Prof. Dr. Frank Veltman

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

ABSTRACT

We investigate the related and important concepts of synonymy and hyperintensionality (i.e. criteria for identity that are more fine-grained than necessary equivalence).

We show how, for every language, validity uniquely determines a co-hyperintensionality relation (that ensures substitution salva veritate). The ordering of operators by their ability to discriminate sentences is not linear (e.g. truthmaking and necessity are incomparable). However, co-hyperintensionality is not a cognitively adequate individuation of content.

Instead, we analyze cognitive synonymy (or likeness in cognitive role) conceptually via defeasible rules, algorithmically via logic programming, and neurally by constructing an appropriate neural network. This explains the contextual stability of synonymy (given by the rules) and its flexibility (some contexts defeat the rules).

We introduce the notion of a scenario that has a representational component (like Fregean senses) and an interpretational component (like a possible world). Scenarios can be grounded, e.g., in neural networks, and they have a constructive notion of distance. We use them to provide a hyperintensional semantics including a counterfactual and belief and conceivability operators.

This scenario framework allows reconstructing many notions of synonymy. We provide a logic for content identity in scenario semantics. We observe that it is inconsistent to hold both (i) if no scenario distinguishes two sentences, they are identical in content, and (ii) content identity entails identity in subject matter. Scenario semantics satisfies (i) and Fine's logic of analytic containment satisfies (ii), though it is not the most coarse-grained one above scenario semantics. A semantics with a content-granularity like analytic containment requires moving from scenarios to sets of scenarios.

We conclude our investigation with a pluralistic conception of synonymy: because of the sheer number of notions of synonymy, and because it is the only way to reconcile the many opposing features of synonymy.

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisors Franz and Hannes: Franz, thank you for your many helpful comments, for the inspiring meetings discussing the thesis, for your honest interest in my ideas, for encouraging me to further pursue this research, for inviting me to give a talk at your *LoC seminar*, for your academic support—and for so much more. Hannes, thank you for your astute remarks, for suggesting very helpful literature, for your fruitful questions, for the great conversations, and – especially – for accepting to supervise the thesis from abroad—this means a lot to me.

Moreover, I would like to thank Michiel van Lambalgen for introducing me to logic programming, its neural implementation, and its fruitfulness in modeling cognitive phenomena—as well as for many other stimulating insights ranging from Kant, over the philosophy of spacetime, to large cardinals in set theory. Also, I would like to thank Mark Jago and Greg Restall for their comments when I presented chapter 2 in Franz' seminar. And, in general, I would like to thank the audiences at that seminar and at the 2017 *Logik zwischen Mathematik und Philosophie* conference in Göttingen for their remarks and questions.

Furthermore, I'm grateful to all those from whom I've learned during my studies. I'm glad I could spend two years at the ILLC. I enjoyed the open-minded and inspiring community of researchers and fellow students—and the fact that you can keep talking logic all night long without having to worry weird looks.

Last but not least, I want to thank my family, my partner, and my friends for their great support during my studies—without you this thesis wouldn't exist!

Contents

1st *Chapter*

INTRODUCTION

We start with a broad and informal introduction to the topic of the thesis, including a description of the perplexing phenomenon of synonymy.

Then we provide the background for our investigation of synonymy and hyperintensionality. This includes a brief history of synonymy and a characterization of the concept of hyperintensionality, also indicating why it is tightly linked to synonymy. Moreover, we sketch existing approaches to provide a hyperintensional notion of meaning—that is, a notion of meaning according to which some necessarily equivalent sentences still can differ in meaning.

Next, we outline the main open problems of hyperintensionality and mention why they are important.

Finally, we summarize what we will do in this thesis including a list of its main contributions.

1.1 General introduction

We give a broad and informal introduction to the topic of the thesis. The focus is not on precision and completeness but on "giving a feel" for the topic.

One of the big questions in philosophy is: *What is meaning?* More specific subquestions include: What is the meaning of words and sentences? Is "the meaning" of a sentence an entity? And if so, is it something like a mind-independent abstract object or is it more like a mental idea? How do we have access to the meaning of sentences? Is meaning correlated with thought? Is the meaning of a sentence at all determined by facts (for example by facts about the speaker community)? And much more.

There are at least two reasons why this question is one of the big ones. The first reason is that we want to understand the foundations of language: How are we able to communicate thoughts? How do we make sense of the world and reason about it by representing it in language? So this reason is much like any other reason for a scientific enterprise to investigate a certain phenomenon: we want to understand the phenomenon. The second reason is more of a "meta-reason". Doing philosophy crucially involves understanding the language that we use to describe philosophical phenomena. Thus, understanding meaning can yield both further philosophical insights and a deeper understanding of philosophical methodology. An often invoked example for the former is the argument of Kripke (1980) that the meaning of "Hes-

perus is Phosphorus" is such that if true, it must be necessarily true—yielding an a posteriori metaphysical truth.[1] An example of the latter is the notorious philosophical method of conceptual analysis: the correctness of a suggested analysis of a concept is often judged by whether the meaning of the words expressing the concept coincides with the meaning of the words used in the analysis.

So the question "what is meaning?" is central to philosophy. However, following the famous slogan "no entity without identity" of Quine (1969, p. 23), understanding meaning first requires to understand when two sentences have the same meaning—that is, when they are synonymous.[2] In other words, before starting with the question of what meaning is, we have to start with the (equally hard) question of when two sentences have the same meaning.[3] This question – *when should two sentences be regarded as having the same meaning?* – will be the topic of this thesis.

But if we start thinking about when two sentences should mean the same thing, we quickly find ourselves at a loss: On the one hand, there are sentences that we commonly would call synonymous. For example, in the "Dead Parrot Sketch" from *Monty Python's Flying Circus* we find many sentences that we would usually take as synonyms for "The parrot is dead":

> It passed on! This parrot is no more! He has ceased to be! It's expired and gone to meet its maker! … Bereft of life, it rests in peace! … Its metabolic processes are now history! … It's kicked the bucket, it's shuffled off its mortal coil, run down the curtain and joined the choir invisible! *This is an ex-parrot*!

On the other hand, if we ask ourselves whether these sentence *really* have the same meaning, we start to wonder: Wait, there might be a difference between "dead" and "ceased to be" or between "dead" and "no metabolic processes". So looking at sentences with a *credulous* stance, we find many synonyms, but as soon we adopt a *critical* stance, we find the synonymy gone. Thus, on the one hand, across many contexts where we might use those sentences they will be regarded as synonymous—so their synonymy has a certain contextual stability. But on the other hand, it seems that for any two (non-identical) sentences we always can – by adopting a critical stance – cook up a context in which the sentences differ in meaning.

So synonymy has two seemingly opposing features: contextual *stability vs. flexibility*. Looking a bit further we find more of these pairs of opposing features of synonymy. (Some of them are related in some way, but each nonetheless stresses an important feature.)

Objective vs. subjective. Usually, we would think that whether or not two sentences are synonymous is a matter of the language and the world alone, and does not depend

[1] Cf. Putnam (cf. 1975a, 232f.). For more, including some critical remarks, see LaPorte (2016, sec. 3).

[2] Quine (1951, 22f.) even says that we should recognize "as the business of the theory of meaning simply the synonymy of linguistic forms …; meanings themselves, as obscure intermediary entities, may well be abandoned".

[3] There is a subtlety here: Prima facie there might be a difference between (i) the identity criterion for the meanings expressed by sentences and (ii) the identity criterion for meaning entities (regarded independently from the sentences expressing them). We focus on (i) here (which yields (ii) if we take meaning entities not as independent but as always expressed by sentences), but for our purposes here, we can ignore this distinction.

on a particular speaker. For example, whether or not "water" is synonymous to "H_2O" does not depend on what I think about these terms.[4] After all, dictionaries provide synonyms without reference to particular speakers. So synonymy appears to be objective or, at least, inter-subjective. However, if we take the critical stance again, we might wonder whether "water" really means the same as "H_2O" for someone to whom water is very important (say due to her job or religion): We might imagine her say that water is so much more than just H_2O-molecules because of all its minerals or because of its flourishing powers. So on this critical stance synonymy appears to be subjective.

Externalism vs. internalism. Sometimes it seems that whether or not two expressions are synonymous is settled purely on external grounds: To take the previous example, whether or not "water" is synonymous to "H_2O" only depends on whether or not the stuff we refer to as water has the chemical structure of H_2O (this is claimed by so-called semantic externalists). However, sometimes it seems that whether or not two expressions are synonymous requires not only external facts but also facts internal to the speakers of the language. For example, that the speakers know or believe that the chemical structure of water is H_2O.[5]

Respecting logical equivalence vs. not. One might think that at least certain logically equivalent sentences are synonymous. For example, one might think that any sentence φ is equivalent to its double-negated version $\neg\neg\varphi$. But again adopting a more critical stance we might wonder whether the sentence "She is friendly" really is synonymous to "She is not not friendly" (which seems more to say that she is neither fully unfriendly nor fully friendly).

Equivalence vs. similarity. On the one hand, we often think of two synonyms as being identical in meaning (taking synonymy to be an equivalence relation). However, on the other hand, we also often think of synonyms not as identical but only very similar in meaning (taking synonymy to be a similarity relation).

Extensional vs. intensional. Oftentimes, it is enough for two terms to be regarded as synonymous to have the same extension[6]. For example, in a travel guide we might see the sentences "Amsterdam has many canals" and "The biggest city of the Netherlands has many canals" to be treated synonymously. In such cases, synonymy acts like an extensional relation: whether or not sentences are in this relation only depends on the extensions of the sentences.

Sometimes, however, we demand that two terms necessarily have the same extension to be regarded as synonymous. In a philosophy class, for example, the just mentioned sentences are not regarded synonymous (since it might have been that Amsterdam failed to be the biggest city in the Netherlands). Yet, "half full" and "half empty" will be regarded as synonymous on this view.

However, sometimes even this is not enough. For example, when someone is pouring some water in a glass it makes sense to say "The glass is already half full" while it doesn't make much sense to say "The glass is already half empty". Roughly,

[4] The example is, of course, a reference to Putnam (1975b).

[5] The issues behinds this externalism vs. internalism distinction are far more intricate than it is suggested in this short paragraph. For an overview, see Lau and Deutsch (2016).

[6] The extension of a singular term (like "the person who proved Fermat's Last Theorem") or a name (like "Amsterdam") is the object it refers to. The extension of a general term (like "animals") or a predicate (like "being red") is the set of things it refers to. The extension of a sentence is its truth-value.

we can describe this situation thus: Both sentences talk about the state of the glass where half of its volume is taken up by water, but "half full" *represents* the state as being reached from below by filling, while "half empty" represents the state as being reached from above by emptying.

While in the first case synonymy is an extensional relation, it is not extensional in the two latter cases—and thus said to be intensional. The second case involved the modal profile of the sentences and the third case involved the representations of a situation provided by the sentences. In the current debate, difference in the modal profile is called an intensional difference, while difference in representation is an instance of what is called a hyperintensional difference—but we'll introduce those terms more precisely in section 1.2.2 below.

Now, in addition to the "left-right" opposing features of synonymy, there also is a "top-down" dimension of granularity of synonymy. We can think of this dimension as *levels of zoom*. For example, in the credulous stance we take "dead" and "no metabolic processes" to be synonymous, but when we "zoom in" far enough we find a difference between the two. Similarly, on an objective level we take "water" and "H_2O" to be synonymous, but when we zoom into the subjective level we find differences—and so on. Each level of zoom corresponds to a decision on what aspects of the expressions in question we deem to be important and what aspects we want to neglect because they don't seem relevant on this level.

So, as always in philosophy, we've started with a fairly simple looking question – when are two sentences synonymous? –, and then we've quickly found that the issue is far more complex. We end this general introduction by noting that the importance of understanding synonymy is not restricted to philosophy. For example, in natural language processing, it is important to understand whether or not two utterances (maybe of the same sentence-type) have the same meaning. And in cognitive science synonymy is important to understand the framing effect: why it is that given two equivalent formulations of a choice between two options we still choose differently depending on the formulation (cf. Tversky and Kahneman 1981).

In the remainder of the introduction chapter, we do the following. In the next section, we provide a background for our investigation of synonymy and hyperintensionality (including a characterization of the concept of hyperintensionality). Then we list the main open problems of hyperintensionality and sketch their importance. Finally, we provide a summary of the thesis including a list of its main contributions.

1.2 The background

To provide a background for our investigation of synonymy and hyperintensionality, we outline the cornerstones in the history of synonymy and introduce the concepts of intensionality and hyperintensionality. Then we sketch modern approaches to provide hyperintensional content and have a brief glance at synonymy in linguistics and cognitive science.

1.2.1 A very brief history of synonymy

In this section, we want to briefly mention some important contributions to understanding synonymy. For reasons of space, this history has to be incomplete, overly simplified, and anachronistic in its formulation. We don't aim to give a precise and exegetically adequate description of the authors, rather we aim at conveying the rough ideas without mentioning their problems.

Leibniz (substitution salva veritate synonymy). One of the most famous ways of characterizing synonymy is by saying that synonyms can be substituted for each other in all sentential contexts without changing the truth-value of the whole sentences in which the substitution took place. This is the *substitution salva veritate* criterion for synonymy. It is usually credited to Leibniz (1686).[7]

Kant (analytic synonymy). In the spirit of Kant's distinction between analytic and synthetic judgments, one can characterize two sentences φ and ψ as synonymous by saying that the statement "φ if and only if ψ" is analytic. And this is sketched, for example, by saying that the equivalence between φ and ψ can already be known just by knowing the meaning of φ and ψ. Or by saying that the meaning of ψ is already "contained" in the meaning of φ and vice versa.[8]

Frege (cognitive synonymy).[9] The substitution salva veritate principle plays a central role in Frege's work.[10] However, Frege additionally has the notion of equipollence to individuate senses—the abstract objects that he alleges as being the meanings of expressions. Roughly, two sentences are equipollent if and only if one could not rationally regard one as true but not the other.[11] This yields a cognitive notion of synonymy: Two sentences are synonymous if we cannot rationally conceive them to come apart. (We will come back to this in the next chapter.)

Carnap & Church (structural synonymy). Carnap (1947, §13–15) notes that necessary equivalence is not enough for substitution salva veritate in all contexts. For example, the sentence "The sun is shining or it is not shining" and a very long logical tautology are necessarily equivalent, but it might well be that someone believes the former but – due to its complexity – not the latter. He then describes the notion of *intensional structure* according to which two sentences have the same intensional structure if, roughly, the two sentences are built in the same way out of corresponding, necessarily equivalent basic sentences. He then shows that this structural synonymy can be used in many cases where a notion of synonymy is needed that is stronger than necessary equivalence. Church (1954) provides the notion of a synonymous isomorphism that is similar in spirit but different in detail.

Goodman & Mates (similarity synonymy). Goodman (1949) and Mates (1952) – and so many others that it seems to be a commonplace – observed that strictly speaking there is no synonymy but only high degree in likeness of meaning. No two (non-identical) sentences can have the same meaning in all respects. Using the terms of the previous

[7] To be more precise, it is credited to Leibniz' 1686 *Generales Inquisitiones de Analysi Notionum et Veritatum*. For a recent discussion see e.g. Malink and Vasudevan (2016).

[8] It should be noted that the analytic/synthetic distinction is one of the most discussed topics in philosophy. For an overview see Rey (2016).

[9] One might be perplexed to read "Frege" and "cognitive" so close to each other because of Frege's famous strong anti-psychologism. However, as the upcoming notion of synonymy hopefully makes clear, he nonetheless provided a cognitive criterion for synonymy.

[10] See, e.g., Frege (1892, 2008[1891]), or Mendelsohn (2005, sec. 2).

[11] See, e.g., Frege (1891, p. 14) or Frege (1979, p. 197). For a recent discussion see Schellenberg (2012).

section, if we just "zoom in" far enough, we can detect a difference between any two non-identical sentences. So synonymy can at most amount to having similar meaning. In short, synonymy is not meaning identity but only meaning similarity.

Quine (skepticism about synonymy). The substitution salva veritate criterion received much interest throughout, but it is particularly prominent in the work of Quine on modal notions and analyticity.[12] For example, Quine (1951, pp. 28-30) famously argued that neither cognitive synonymy nor analyticity can non-circularly be provided by substitution salva veritate. Or Quine (1960, 181f.) argued that the substitution salva veritate principle collapses all modal notions (i.e., making "φ" equivalent to "Necessarily, φ"). Especially the latter argument received a careful analysis by Quine's student Føllesdal showing that it can be generalized to showing that all intensional operators collapse and that its drastic conclusion can be avoided if we sharply distinguish between singular and general terms.[13,14]

1.2.2 Intensionality and hyperintensionality

In this section, we specify the notions of intensionality and hyperintensionality (as they are used in the current debate).

Intensionality. The intension of an expression is a function that assigns each possible world to the extension that the expression has in that world. This is the core of possible-worlds semantics: The meaning of an expression is modeled by its intension. For example, the intension of the predicate "is red" maps each world to the set of red things in that world. The intension of a sentence maps a world to 1 if the sentence is true in that world and to 0 otherwise. Thus, according to possible-worlds semantics, the meaning of a sentence – also called a proposition – is just the set of possible worlds where the sentence is true.[15]

Possible world semantics is a success-story of philosophy: In the second half of the 20$^{\text{th}}$ century, it was used to analyze many philosophically important concepts (like meaning, knowledge, counterfactuals, causality, etc.). Its revolutionary idea was to replace the purely extensional theories of the time by intensional ones. According to Nolan (2014), we're currently seeing another "revolution" replacing intensional theories by hyperintensional ones. So let's see what this hyperintensionality is.

Hyperintensionality. While it is commonly agreed what "extension" and "intension" refer to, the only agreed characterization of "hyperintensional" is negative as something that is neither intensional nor extensional. So hyperintensionality is about individuating entities finer than necessary equivalence. The term was introduced by Cresswell (1975, p. 25) to describe the phenomenon that in some sentential contexts – like the already mentioned belief contexts – even logical equivalence is not enough to ensure substitution salva veritate. Such contexts are called *hyper*intensional, because *more* than co-intensionality (in the sense of necessarily having the same extension) is needed for substitution salva veritate. (Nowadays, Cresswell's "logical equivalence"

[12] This work of Quine originated in the 1940s and '50s. Starting with Quine (1941), including the famous Quine (1951), and culminating in Quine (1960, esp. § 41).

[13] See Føllesdal (2004) which is a reprint of his 1961 PhD thesis. For a short version see Føllesdal (2013, pp. xxiii–xxvi).

[14] For reasons of space, we don't mention Quine's positive take on synonymy. For example in Quine (1960, e.g. pp. 29,41) via stimulus meaning and stimulus synonymy. Also see Føllesdal (2013, pp. xvi–xxi).

[15] A function $f : A \to \{0,1\}$ exactly correspond to the set $\{a \in A \mid f(a) = 1\} \subseteq A$.

has been replaced by "necessary equivalence".[16]) Since then, many things have been called "hyperintensional": not only sentential contexts, but also sentential operators, notions of content, principles of individuation, relations, concepts, and theories. For definiteness, we'll characterize them here (as precisely as such a general setting allows for).

Hyperintensional context A sentential context[17] is hyperintensional, if co-intensionality (i.e. necessary equivalence) does *not* ensure substitution salva veritate in that context.

Similarly, a sentential context is extensional (resp. intensional), if co-extensionality (resp. co-intensionality) ensures substitution salva veritate in that context. A sentential operator is a function that takes, for a natural number n, n-many sentences as input and produces a new sentence (e.g., negation, conjunction, or "belief that ..."). We call n the arity of the operator.

Hyperintensional operator A sentential operator of arity n is hyperintensional, if there are sentences $\varphi_1, \ldots, \varphi_n$ and $\psi_1, \ldots, \psi_n$ such that φ_i and ψ_i are co-intensional (for all $i \leqslant n$), but such that the operator applied to the φ's yields a true sentences and applied to the ψ's it doesn't.

A notion of content provides some – loose or precise – entity that can be regarded as the content, semantic value, or meaning – in the widest sense of the word – of a sentence (or, in general, of an expression). Examples are: extension, intension, Fregean sense, subject matter, truthmakers, cognitive value, psychological association, poetic value, or pragmatic implicature. The word "content" is often used instead of "meaning" since it is more neutral—in current philosophy of language, "meaning" usually has an externalist understanding.[18] Though, sometimes the word "content" is used more restrictively as a notion of content that is more (or equally) fine-grained than intensions but excludes subjective and pragmatic aspects like association and implicature.

Hyperintensional content A notion of content is hyperintensional, if there are two sentences that have the same intension but that have different content.

Hyperintensional relation A binary relation between sentences is hyperintensional, if there are two co-intensional sentences that are not in that relation.

A principle of individuation says when two sentences should be regarded as having the same content. Thus, it is a binary relation between sentences, and it is hyperintensional if that relation is hyperintensional. Finally, a theory is hyperintensional if it is formulated in a language containing hyperintensional operators, and a concept is

[16] See e.g. Nolan (2014, p. 151) or Jespersen and Duží (2015, p. 525).

[17] A sentential context is a position in a sentence. For example the dots in the following sentences are contexts in their surrounding sentences: "The sun is shining or ...", "She believes that ...", "If she believes that ..., then she'll pass the exam". In a formal language, any subformula of a formula forms a sentential context. Here we've only mentioned sub-sentences as sentential contexts because we will will only deal with those below. But also a predicate-place or a noun-place in a sentences can be regarded as a context, and the characterization of them being hyperintensional works analogously.

[18] That is, what a sentence means depends not only on what we think it means but also on facts about the world we live in. Cf. section 1.1.

hyperintensional if it needs a language containing hyperintensional contexts to be adequately described (Nolan 2014, 151f.).[19]

Note that synonymy and hyperintensionality are tightly linked: If synonymy is considered to be more than – or just different from – mere co-intensionality, then synonymy is a hyperintensional relation. (And in section 1.1 we've already seen indications that synonymy plausibly is different from co-intensionality.)

Sometimes something stronger is required for content to be hyperintensional: namely, that it not only "sees" differences between co-intensional sentences but also that it determines the intension of a sentence.

Strictly hyperintensional content A notion of content is strictly hyperintensional, if there are two co-intensional sentences that have different content, *and* whenever two sentences have the same content, they are co-intensional.

Similarly, a relation is strictly hyperintensional, if it is hyperintensional and whenever two sentences are in that relation, they are co-intensional. However, usually, we'll work with the more general notion of hyperintensionality and not with strict hyperintensionality. And in the next chapter (section 2.3) we'll see interesting cases where they come apart.

It should be noted that especially before the advent of possible-world semantics, it used to be common to call something intensional whenever it cannot be understood in purely extensional terms. Here, however, we use the more recent terminology and reserve "intensional" for modal intensionality and "hyperintensional" for anything that exceeds even modal intensionality.

In section 1.4, we'll list several examples of hyperintensional contexts and concepts.

1.2.3 Modern approaches to hyperintensionality

In this section, we very briefly sketch several existing approaches providing hyperintensional content.

2D Semantics.[20] Among the many different forms of two-dimensional semantics, the most holistic arguably is that of Chalmers (2006a). The main idea is that the meaning of an expression is modeled as consisting of two parts (or "dimensions"). One is the so-called subjunctive- or 2-intension which is like the regular intension in possible worlds semantics. The other one – the so-called epistemic- or 1-intension – also assigns each possible world an extension but in a way that reflects the "cognitive role" of the expression rather than the "metaphysical role" represented by the usual (2-) intension. To determine the 2-intension, we fix the referents of our expressions in our world and then we move to a possible world and see how they might have been different there. Thus, even in a world where Venus is only visible in the morning and where the brightest object in the evening sky is Mars, both "Phosphorus" and "Hesperus" pick out the object Venus (because both names pick out the planet Venus in the actual world). To determine the 1-intension, we first go to a possible world and then fix the referents of our expressions there in the same way as we did in the

[19] This still is quite a vague characterization. But we refrain from making it more precise since we only mention it here for completeness and won't need it below.

[20] See e.g. Chalmers (2002, 2006a,b). For other versions and a more general introduction see Schroeter (2017).

actual world, and we then look at the properties of the objects we thus fixed. Hence, if we move to the just described possible world, "Phosphorus" (whose referent we picked in the actual world as the brightest star in the morning sky) refers to Venus and "Hesperus" picks out Mars. This two-dimensional meaning provides hyperintensional content because the (2-) intensions of two expressions – like with "Hesperus" and "Phosphorus" – can be the same while their 1-intensions still differ.

Impossible worlds semantics.[21] Possible worlds semantics is very useful but it has the serious shortcoming of collapsing all impossibilities and all necessary truths. The idea of *im*possible world semantics is to extend possible world semantics such that it keeps its merits but avoids the collapsing problem. This is done by adding to the possible worlds also impossible worlds. Heuristically speaking, just as possible worlds represent how things could have been, impossible worlds represent how things could *not* have been. Thus, there is an impossible world w where some bachelors are married but where all vixens still are female foxes. If the content of a sentence is taken to be the set of possible *and* impossible worlds that make the sentence true, then w draws a hyperintensional distinction between the necessary truths "All bachelors are unmarried" and "All vixens are female foxes".

Finesse built from worlds. What the two approaches so far have in common is that, roughly, they add – in different ways – more structure on top of possible worlds semantics. There are more approaches of this kind: They construct meaning entities purely by set-theoretic operations on possible worlds, and such that these meaning entities can draw hyperintensional distinctions between sentences. See, for example, Lewis (1970), Cresswell (1975), Berto (2010), or neighborhood semantics for modal logic (Pacuit 2017).

The next approaches move away increasingly further from possible worlds semantics—the final one basically doesn't have anything in common anymore.

Truthmaker semantics.[22] The idea of truthmaking is summarized by (Fine 2016a, § 1) as "the idea of something on the side of the world – a fact, perhaps, or a state of affairs – verifying, or making true, something on the side of language or thought – a statement, perhaps, or a proposition". For example, the fact that Anna walks makes true the sentence "Anna walks or doesn't walk" while it doesn't make true the sentence "Bob talks or doesn't talk"—so truthmakers can draw hyperintensional distinctions. Quite some work has to be put into making this precise, and this involves separating the metaphysical project of *what* truthmakers are from the semantic project of showing *how* truthmakers make sentences true. We will come back to Fine's truthmaker semantics in section 5.3.

Structured propositions.[23] This view agrees that the meaning of sentences are propositions but that these propositions are more than mere sets of possible worlds as it is claimed by possible worlds semantics—they have (more) structure. On this view, a proposition is built up from objects, properties, and relations (so ontologically speak-

[21] See e.g. Jago (2014), Priest (2005), or Berto (2017), and for more on impossible worlds see Berto (2013).

[22] See Fine (2016a) for a good overview including references. Truthmaker semantics has recently gained popularity due to the work of Yablo (2014a) and Fine (in a series of papers starting with 2014 and 2016, and many others are found on his *Academia* page).

[23] See e.g. Soames (1985, 1987) and King (1995). For an overview see King (2016). This approach is sometimes dubbed "Russellian propositions" because it's similar to the analysis of propositions given by Russell (1903, ch. IV).

ing a proposition is very different from a sentence).[24] For example, the proposition that the sentence "Anna walks" expresses is, roughly, the pair $\langle a, W \rangle$ where a is Anna and W is the property of walking. (How exactly the proposition looks like is not important here; it's the idea that matters.) Similarly, the sentence "apples are apples" expresses the proposition $\langle = \langle A, A \rangle \rangle$ where A is the property of being an apple. And "bananas are bananas" expresses $\langle = \langle B, B \rangle \rangle$ where B is the property of being a banana. Thus, even though the two sentences "apples are apples" and "bananas are bananas" are co-intensional, they express – on this view – different propositions: $\langle = \langle A, A \rangle \rangle$ is an object different from $\langle = \langle B, B \rangle \rangle$ since the object A is different from B (apples are not bananas).[25]

Meanings as algorithms.[26] The idea of this approach is to take the sense or meaning of an expression as the algorithm or procedure which allows one to compute or determine the extension of the expression. (On a related view, one could take a proposition to be an algorithm to construct the discourse model that the proposition induces.) Intuitively, although "All bachelors are unmarried" and "two plus two is four" are co-intensional, the procedures we use to determine that they are true are very different. Note that on this view whether or not two sentences have the same meaning is computationally not decidable (because algorithm identity is not decidable). This accounts for bounded rationality: We cannot just automatically compute whether, say, a simple and a complex (first order) tautology have the same meaning. Instead, we have to put in some creativity to see that they indeed both express a necessary truth.

For reasons of space, we cannot further analyze the approaches and compare them—though, see Gioulatou (2016) for an excellent discussion of impossible worlds semantics, structured propositions, and both Fine's and Yablo's truthmaker semantics.

In this thesis, we will develop a new approach to hyperintensional content (starting in chapter 3). It will integrate many merits of the approaches mentioned. For example, it will include how a speaker represents the world (cf. the epistemic intension in 2D semantics), it will allow for inconsistent scenarios (cf. impossible worlds), and there will be truthmakers. For reasons of space, we won't argue against the existing approaches or voice any dissatisfaction with them. We also won't compare our approach to all the existing ones (that would require another thesis), but – as already mentioned – we will compare it to the truthmaker semantics. Though, we believe that when we develop our approach, connections to other approaches (and advantages of our framework) will become apparent.

1.2.4 A very brief glance at synonymy in linguistics and cognitive science

This subsection concludes the background section and we will have – for the sake of completeness – a *very* brief glance at synonymy in linguistics and cognitive science. After all, so far we've only seen theories about synonymy or semantics from a philosophical or formal semantics perspective. Of course, linguistics and cognitive

[24] Though, there are variants where instead of taking objects, properties, and relations to build propositions one takes, roughly, their respective senses (cf. Zalta 1988).

[25] Note the similarity to Carnap's intensional structure mentioned in section 1.2.1.

[26] See van Lambalgen and Hamm (2005) and Moschovakis (1994, 2006). Moreover, the so-called procedural semantics of Duží et al. (2010) is very similar in spirit and it is based on Pavel Thichý's transparent intensional logic.

science is a way too big of a field to be covered in such a brief glance—though see Stanojević (2009) for an overview of synonymy from the cognitive perspective. So we only sketch the idea of what is called *distributional semantics* since it arguably provides the most prominent way to capture synonymy or similarity in meaning in current computational linguistics (and computational cognitive science)—but, of course, there is much more.

Distributional semantics is built around the idea that words that are similar in meaning occur in similar contexts. That is, words that are similar in meaning have a similar distribution across contexts (in large text corpora). This is called the *distributional hypothesis* and it is usually credited to Harris (1951, 1954).[27] The importance of this hypothesis is that according to it meaning similarity can be approximated by distributional similarity in large text corpora, and the meaning of a word can be approximated by its distribution pattern. These distribution patterns and their similarity – in turn – can be quantified (as vectors and vector similarity). This can also be implemented, and this is the key to the success of distributional semantics. For reasons of space we cannot go into more details, but for a recent introduction to distributional semantics see Clark (2015).

However, here is one worry that distributional semantics might not help us too much in our aim of gaining a deep understanding of (the concept of) synonymy. The notion of meaning similarity of distributional semantics only tells us *that* – as a matter of (contingent) fact – two words are similar in meaning (because they co-occur a lot), but it doesn't tell us *why* they are similar in meaning. The difference between the distributional meaning of a word and the (philosophical) meaning of a word is somewhat analogous to the difference between the extension of a concept and the concept itself. To illustrate, assume that the extension of a given concept is the set of numbers $\{2, 3, 5\}$. Then we don't know much about the concept itself yet. For it could be that the concept is *the first three prime numbers* or *the prime divisors of 30* or *my friend's three most favorite numbers*. We know that the concept has something to do with its extensions (like the word has to do something with the words it often co-occurs with) but we don't know why it has this extension (resp. why the word co-occurs a lot with the others). Of course, a lot more should be said here, but space doesn't allow. For a discussion of more worries see Lenci (2008, esp. sec. 1.1 and sec. 3) and Sahlgren (2008).

This is – to be sure – not to say that distributional semantics is of no use to understanding synonymy, but we do take it as an indication to not take it as a starting point for a philosophical and conceptual investigation of synonymy.

1.3 The problem(s)

We said that the big problem that we're after in this thesis is when we should regard two sentences as having the same meaning. In other words, when are they synonymous, or when is their content the same? In this section, we will briefly sketch several sub-problems that are currently discussed in the philosophical literature.

[27] Though, of course, many others were involved: For example, Leonard Bloomfield, John Rupert Firth, Margaret Masterman, and the late Wittgenstein.

1.3.1 The problem of co-hyperintensionality

In a recent special issue of *Synthese* on hyperintensionality, Jespersen and Duží (2015, p. 526) present arguably the most important open problem for hyperintensionality. We'll call it the problem of co-hyperintensionality. We'll introduce it here and deal with it in the next chapter.

As we've seen, there are operators where *more* than necessary equivalence of sentences is required to ensure substitution salva veritate. The problem of co-hyperintensionality is concerned with what "more" amounts to: What is the relation between sentences that has to hold for substitution salva veritate also in hyperintensional contexts?

Let's work toward a more precise formulation of the problem. We're asked to find a relation between sentences of a given language such that, roughly, whenever two sentences are in that relation, applying any operator of the language to those sentences will yield two equivalent statements (and being in that relation is the weakest sufficient condition for this). For instance, we want to find a relation two natural language sentences φ and ψ have to be in such that this is the weakest sufficient condition for both of the following inference-schemata to be valid

$$\frac{\text{Agent } a \text{ knows that } \varphi}{\text{Agent } a \text{ knows that } \psi} \qquad \frac{\text{Agent } a \text{ knows that } \psi}{\text{Agent } a \text{ knows that } \varphi.}$$

Now, this should not just work for the knowledge operator but for any operator of the given language.[28] (Recall that we call ∇ an n-ary operator if ∇ takes n sentences as arguments and outputs another sentence.) Thus, *the problem of co-hyperintensionality* reads as follows.

Given a language (be it a formal or natural language), we want to find a relation $\sim$ between the sentences of the language such that for any n-ary operator ∇ of the language we have

(i) If $\varphi_i \sim \psi_i$ for all $i = 1, \ldots, n$, then both inferences

$$\frac{\nabla(\varphi_1, \ldots, \varphi_n)}{\nabla(\psi_1, \ldots, \psi_n)} \qquad \frac{\nabla(\psi_1, \ldots, \psi_n)}{\nabla(\varphi_1, \ldots, \varphi_n)}$$

are valid (which we'll also denote $\nabla(\varphi_1, \ldots, \varphi_n) \Leftrightarrow \nabla(\psi_1, \ldots, \psi_n)$).

(ii) If we replace the condition '$\varphi_i \sim \psi_i$ for all $i = 1, \ldots, n$' in (i) by a weaker condition, then (i) won't be true in general anymore.

Given a language, we'll call a relation $\sim$ satisfying (i) and (ii) a *co-hyperintensionality relation*.

In the next chapter, we'll further discuss the problem and present a solution (in the sense of reducing the problem to validity).

[28] Given our informal formulation of the problem in the preceding paragraph or those of Jespersen and Duží (2015, p. 526) or Faroldi (2016, p. 1), one might object that this should not work for *all* operators but just for the hyperintensional ones. However, if it works for hyperintensional operators, it will also work for less fine-grained intensional or extensional operators. Thus, by quantifying over all operators we can neatly avoid to first be able to distinguish between hyperintensional and non-hyperintensional operators.

1.3.2 Related problems

We mention several further problems of hyperintensionality and indicate why the problem of co-hyperintensionality is the most fundamental one.

One problem is to ask whether the notions of synonymy – mentioned in section 1.2.1 – that aim to capture content identity (and not mere content similarity) are jointly consistent. (We'll find that they are not.)

The *granularity problem* of possible worlds semantics asks how to avoid that all necessarily equivalent sentences express the same meaning entity, while preserving the benefits of possible worlds semantics.

The *problem of normative relations* asks us, for instance, how exactly the content of a conjunction should be related to the content of the conjuncts to capture the behavior in, say, belief contexts. Or, more generally, how to "reconcile hyperintensional contents with the normative principles governing epistemic concepts" Jago (2014, p. 10f.).

The *problem of bounded rationality* ask us to account for the fact that by knowing one thing we, as humans, don't know all of its logical consequences (that is, necessary equivalence is not enough to ensure substitution salva veritate in knowledge-contexts). Note that this problem is independent of possible worlds semantics: "We lack a satisfactory understanding, from any point of view, of what it is to believe that P while disbelieving that Q, where the 'P' and the 'Q' stand for necessarily equivalent expressions" (Stalnaker 1984, p. 24). In the context of epistemic logics or possible worlds semantics, this problem is called the *problem of logical omniscience*, since there knowledge is analyzed as truth in all (epistemically) possible worlds which notoriously renders the agents to know all logical consequences and all logical truths.

We end this subsection by briefly commenting in what sense the problem of co-hyperintensionality is the most fundamental one among the problems of hyperintensionality—and hence should be dealt with first. Jespersen and Duží (2015, p. 529) note that the problem of co-hyperintensionality is still in the realm of logic, while the others are already part of philosophy of language or related fields. Moreover, it needs less notions to state it compared to the other problems of hyperintensionality, and finding the correct criteria for identity is in general the first step in showing the existence of a potential kind of entity.

1.3.3 The Lewis-Stalnaker objection to hyperintensionality

There is one objection to hyperintensional content raised by David Lewis and Robert Stalnaker. We present it here and then dismiss it.

The Lewis-Stalnaker objection to hyperintensional content runs as follows.[29] There is the troublesome data of necessarily equivalent sentences that we intuitively take to differ in meaning. However, this data should *not* be accounted for by making meaning more fine-grained than intensions. Rather it should be accounted for by the fact that we don't always know what intension is expressed by a sentence.

This is a fair point: For example, if one wants to defend a conception of synonymy that is – in the terminology of section 1.1 – objective, externalist, (modally) intensional, and respects logical equivalence. Or, if one wants to defend the position that for an externalist description of the world intensions have exactly the right granularity. (For

[29] See e.g. Stalnaker (1984, ch. 5) , Lewis (1982), Lewis (1986, sec. 1.4), or Stalnaker (2002).

example, Williamson (2013, p. 217 & p. 266) seems to be in favor of this position, while Nolan (2014, p. 156-8) claims that not only our representation of the world but also the world itself is hyperintensional.)

However, there are two issues. First, if the objection is taken to defend the just mentioned conception of synonymy, then – as seen in section 1.1 – this is a highly specific reconstruction of the intuitive notion of synonymy. There are many features of synonymy that are not captured on this conception: to mention but one example, it renders synonymous "Hesperus is Hesperus", "Hesperus is Phosphorus", and any mathematical truth like Fermat's Last Theorem. Second, and more importantly, the objection (merely) shifts the problem from explaining the troublesome data to explaining when we *take* two sentences to refer to the same intension. In other words, even if one holds that hyperintensionality is a phenomenon only occurring on the level of our representation (or conceptualization) of the world and not in the world itself, one still has to explain representational hyperintensionality.[30,31]

Here we're interested in a more general framework that can also capture both the other features of synonymy and representational hyperintensionality. We start constructing such a framework in chapter 3.

1.4 Importance of hyperintensionality and synonymy

In this section, we briefly sketch the importance of hyperintensionality and synonymy.

In section 1.1, we've already seen that synonymy (and hence hyperintensionality) is essential both for a broad range of philosophical questions and for philosophical methodology. We also noted that synonymy is important in areas of cognitive science and linguistics. And the width of the problems of hyperintensionality discussed in the last section, too, indicates its importance.

Furthermore, let's look at examples of hyperintensional contexts and concepts to see that many – if not most – philosophically interesting concepts are hyperintensional. We'll first list them and then discuss them below.

(i) Cognitive operators: believe, perceive, conceive, surprise, being funny, intend.

(ii) Epistemic operators: knowledge, intuition, a priori.

(iii) Semantics: sense, subject matter, truth-making, synonymy, passive/active, counterfactuals & counterpossibles.

(iv) Explanation (Schnieder 2011), causation, verification, confirmation, proof.

(v) Metaphysics: hyperintensional accounts of essence (Fine 1994), metaphysical grounding (Fine 2012; Schaffer 2009), intrinsicality (Bader 2013; Eddon 2011).

For reasons of space we won't discuss all of these examples—they only serve as an illustration here (so we're not committed to all of them actually being hyperintensional).

[30] For a discussion of the view that hyperintensionality is a representational and not a worldly phenomenon, see Nolan (cf. 2014, sec. 3), and for a motivation see Jespersen (2010, esp. p. 97).

[31] Also see, e.g., Jago (2014, sec. 2.4–2.6) for a discussion and criticism of the Lewis-Stalnaker objection.

Ad (i). We've already seen that belief contexts are hyperintensional. A more mundane example is "it is funny that": Arguably, it's mildly funny that a skeleton has *no body* to go with, while it's not mildly funny that a skeleton has *no one* to go with—although the two sentences after the that-clause are co-intensional. Also, I can see that there is a glass on the table without seeing that there is a glass on the table and apples are apples.

Ad (ii). It's the same with knowledge: I can know that apples are apples without knowing a very complicated mathematical truth or a very complex logical truth (both are necessary and hence co-intensional with a simple logical truth).

Ad (iii). The two tautologies "apples are apples" and "bananas are bananas" are co-intensional, but they have different subject matter (one is about apples and the other about bananas), they have different senses, they have different truthmakers, and intuitively they are not synonymous (on a reasonably fine notion of synonymy). The concept of active/passive needs a hyperintensional analysis because the sentences "The child scratched the car" and "The car was scratched by the child" are co-intensional but one is active while the other is passive. Also, the counterfactual "If Anna squared the circle, Anna would be famous" is true while "If apples weren't apples, Anna would be famous" intuitively is not—although the premisses are co-intensional (also cf. section 4.2.1, esp. footnote 109).

Ad (iv). The notion of a proof needs an hyperintensional analysis, too: In mathematics there often exist several different proofs of a single theorem—for example, Meŏstrović (2012) provides 169 proofs of Euclid's theorem of the infinitude of primes. But obviously, their difference or sameness cannot be accounted for by possible worlds, for if the proofs are correct, they are necessarily so. Moreover, going back to an example from Carl Hempel, a raven that is black is a piece of confirmation for "All ravens are black" but arguably not for the logically equivalent "All non-black things are not ravens" (cf. e.g. Yablo 2014a, sec. 1.5).

Ad (v). The examples from metaphysics serve to show that we're witnessing more and more hyperintensional analyses of important metaphysical notions. As already indicated, Nolan (2014, p. 149) even goes as far as to claim that the "twenty-first century is seeing a hypterintensional [*sic*] revolution".

1.5 Outlook

We provide a non-technical summary of what lies ahead.

Chapter 2 – The structure of hyperintensionality. We start with the problem of co-hyperintensionality since it is considered to be one of the most pressing open problems in the foundations of hyperintensionality and since the hope is that it finds exactly the right granularity for synonymy. We develop a simple and general framework to show that for every language there is one unique co-hyperintensionality relation that, moreover, coincides with substitution salva veritate in every context. This relation can be defined in terms of validity, so without committing to a particular semantics, we reduce co-hyperintensionality to validity.

We then apply this framework to investigate the structure of hyperintensionality. We find: The ordering of operators by granularity (i.e. by their ability to discriminate sentences) is partial but in general not linear. Even natural operators like necessity

and inexact truthmaking are incomparable. Moreover, incomparable operators point to a plurality of logics for hyperintensional operators. Yet, if validity is transitive, individuating content by co-hyperintensionality is – in general – not a cognitively adequate individuation of content. This is an impossibility result: co-hyperintensionality cannot provide a notion of cognitive synonymy.

Chapter 3 – Cognitive synonymy. Given this impossibility result, we start over and describe the concept of cognitive synonymy as likeness in cognitive role: Two sentences have the same cognitive role, if, roughly, what we usually (should) interpret, presuppose, understand, and – inductively, abductively, or deductively – conclude given one sentence we also do given the other. We can recognize that two sentences φ and ψ have the same cognitive role in this sense, if, roughly, we have the rules "Usually, if φ, then ψ" and "Usually, if ψ, then φ" in our knowledge base.

This is a "conceptual" analysis of our cognitive ability to recognize two sentences as playing the same cognitive role. We then provide this analysis on the other two Marrian levels: on the algorithmic level by using logic programming (precisely capturing the notion of a defeasible rule), and on the neural level by providing a neural network that explains why cognitive synonymy has both rule-like features but still is (contextually) flexible.

Chapter 4 – Scenarios. The framework thus developed gives rise to a notion of a scenario. Roughly, a scenario consists of a set of rules and facts that represent a part of a (possible) world and of an interpretation of these facts and rules that thus constitutes the intended model of that part of the world. One way these scenarios can be grounded is in the states of (idealized) neural networks with which a (possible) agent conceptualizes and reasons about the part of the (possible) world she perceives.

We observe that the class of scenarios is rich in structure: it has, for example, a notion of distance and various modal accessibility relations. We use this to describe when a scenario makes true (or false) a logically complex sentence—including cognitive operators and a counterfactual. This scenario semantics combines the advantage of possible worlds semantics that it can speak of "truth at a world" with taking into account different modes of representation provided by the rules. Thus scenario semantics provides hyperintensional content and various interesting notions of validity that we will characterize.

Chapter 5 – Notions of synonymy. We will use the scenario framework to describe and characterize various notions of synonymy. These notions range from "world based" ones that only take the current scenario into account, over those that take surrounding scenarios into account, to "logical" ones that only take logical relations between sentences into account. Also, we provide a logic where $\varphi \equiv \psi$ is derviable if and only if φ and ψ have the same hyperintensional content according to the scenario semantics.

We observe something paradoxical about the notion of content identity (or absolute synonymy). The following two principles are jointly inconsistent: (i) If no scenario can be imagined in which two sentences differ in meaning, they are identical in content, and (ii) content identity entails identity in subject matter. Content identity according to the scenario semantics satisfies (i) but not (ii). Looking for notions of synonymy that satisfy (ii), we find that the logic of analytic containment of Fine (2016a) satisfies (ii), though it is not the first above scenario semantics that does that. Moreover, if we

want to account for Fine's finesse of content with scenarios, it is both sufficient and in a precise sense necessary to move from (single) scenarios to sets of scenarios making sentences true (or false).

We conclude that our investigation leads us to a pluralistic conception of synonymy: This is not only because of the sheer number of independently well-motivated notions of synonymy that we'll see, but also because of the many opposing features of synonymy that can only be reconciled by acknowledging a plurality of synonymy.

1.5.1 The main contributions of this work

We summarize the main and – to the best of my knowledge – novel ideas and results of the thesis.

- The terminology of opposing features of synonymy: contextual stability vs. flexibility, objective vs. subjective, externalism vs. internalism, respecting logical equivalence vs. not, extensional vs. intensional.
- The terminology of hyperintensional and strictly hyperintensional.
- For every language the co-hyperintensionality relation is unique, determined by validity, and coincides with the relation ensuring substitution salva veritate in every context.
- The ordering of operators by granularity (i.e. by their ability to discriminate sentences) is partial but in general not linear. For example, knowledge and explanation are incomparable and so are inexact truthmaking and necessity.
- Incomparable operators indicate a plurality of logics for hyperintensional operators.
- An impossibility result: co-hyperintensionality is not a cognitively adequate individuation of content.
- We analyze cognitive synonymy (or likeness in cognitive role) conceptually via defeasible rules, algorithmically via logic programming, and neurally by constructing an appropriate neural network.
- This explains the contextual stability of synonymy (given by the rules) and its flexibility (in some contexts the rules get defeated).
- Our neural implementation extends existing ones by allowing not only binary states but also continuous ones. This allows to take evidence into account and explains why some contexts defeat a rule while others don't.
- The notion of a scenario: Representational and interpretational component, grounded e.g. in (states of) neural networks, various accessibility relations, construction of a pseudometric on the class of scenarios, a notion of a negligible set of scenarios.
- Scenario semantics: provides cognitive operators (belief and conceivability), a counterfactual, and hyperintensional content.

- Various interesting notions of validity (including a novel one called exclusion preserving validity), "well-behaved" scenarios yield the strong Kleene three-valued logic, the class of all scenarios yields the first-degree entailment four-valued logic.
- The scenario framework allows to reconstruct many notions of synonymy: They can be ordered from local to global, and from providing a contentful link between the sentences to a purely logical link.
- A sound and complete logic for content identity in scenario semantics.
- The following two principles are jointly inconsistent: (i) If no scenario can distinguish two sentences, they are identical in content, and (ii) content identity entails identity in subject matter.
- Characterizing the logic of analytic containment (AC) of Fine (2016a): Two sentences are AC-equivalent if and only if, roughly, they are equivalent in the logic of first-degree entailment and their disjunctive forms share the same literals.
- AC-equivalence does satisfy (ii) but it is not the first notion above scenario semantics that does so. The first notion is axiomatized by

$$\mathrm{AC} + \text{“}\varphi \vee (\varphi \wedge \psi) \equiv \varphi \vee (\varphi \wedge \neg\psi)\text{”}.$$

- Moving from scenarios to sets of scenarios as truth- or falsemakers of sentences is sufficient and necessary to make scenario semantics more fine-grained such that it individuates as AC-equivalence.
- Pluralism about synonymy: because of the sheer number of independently well-motivated notions of synonymy, and because it is the only way to reconcile the many opposing features of synonymy.

2nd *Chapter*

THE STRUCTURE OF HYPERINTENSIONALITY

The problem of co-hyperintensionality asks us to find a relation between the sentences of a given language such that applying any (n-ary) operator to (n-tuples of) sentences (pairwise) in that relation yields equivalent statements (and being in that relation is the weakest sufficient condition for this). This should particularly work for alleged hyperintensional operators, like belief or explanation, for which it is not clear whether such a relation exists. In this chapter, we develop a simple and general framework to show that for every language there is one unique such relation that, moreover, can be defined in terms of validity. Thus, without committing to a particular semantics, we reduce co-hyperintensionality to validity.

We then apply this framework to investigate the structure of hyperintensionality—an important task since most philosophical notions are considered to be hyperintensional. We find: The ordering of operators by granularity (i.e. by their ability to discriminate sentences) is partial but in general not linear. Even natural operators like necessity and inexact truthmaking are incomparable. Further, co-hyperintensionality indeed coincides with substitution salva veritate in any context. Moreover, incomparable operators point to a plurality of logics for hyperintensional operators. Yet, if validity is transitive, individuating content by co-hyperintensionality is inconsistent with the Fregean equipollence criterion for the sameness of content and not cognitively adequate—this yields an impossibility reading of our results and should make us rethink the purpose of co-hyperintensionality (which will be the starting point for the next chapter).

2.1 The problem of co-hyperintensionality

We recall from the introduction (section 1.3.1) that – according to the literature – one of the most fundamental open problems concerning the concept of hyperintensionality is the problem of co-hyperintensionality.

Problem 2.1.1. Given a language (be it a formal or natural language) we want to find a relation $\sim$ between the sentences of the language such that for any n-ary operator ∇ of the language we have

(i) If $\varphi_i \sim \psi_i$ for all $i = 1, \dots, n$, then both inferences

$$\frac{\nabla(\varphi_1, \dots, \varphi_n)}{\nabla(\psi_1, \dots, \psi_n)} \qquad \frac{\nabla(\psi_1, \dots, \psi_n)}{\nabla(\varphi_1, \dots, \varphi_n)}$$

are valid (which we'll also denote $\nabla(\varphi_1, \dots, \varphi_n) \Leftrightarrow \nabla(\psi_1, \dots, \psi_n)$).

(ii) If we replace the condition '$\varphi_i \sim \psi_i$ for all $i = 1, \ldots, n$' in (i) by a weaker condition, then (i) won't be true in general anymore.

Given a language, we'll call a relation $\sim$ satisfying (i) and (ii) a *co-hyperintensionality relation*. Three remarks.

First, it is a commonplace to take validity to be necessary truth-preservation. So we could replace the above intended interpretation of $\Leftrightarrow$ as validity by necessary truth-preservation. Though, our results are independent of this choice. So even if validity were to differ from necessary truth-preservation, our results would obtain for both of these intended interpretations of $\Leftrightarrow$.

Second, as already seen, the problem is also often formulated as finding a relation between sentences that ensures substitution salva veritate in any context. It seems like it often is tacitly assumed that both formulations are equivalent. However, prima facie this is not obvious, since in the substitution salva veritate formulation we only substitute one sentence at a time, but in the co-hyperintensionality formulation we substitute n-many sentences in one go. Among others, proposition 2.2.7 below shows that this makes indeed a big difference. Yet, in corollary 2.2.12 we show that being substitutable salva veritate in every context still is equivalent to being co-hyperintensional.

Third, and finally, note that condition (i) demands that a co-hyperintensionality relation is fine-grained enough: If one of the inferences is not valid, this can be traced back to the fact that one of the pairs of corresponding sentences is not co-hyperintensional. On the contrary, condition (ii) demands that a co-hyperintensionality relation is not too fine-grained: As validity is reflexive, we could always choose the identity relation to fulfill (i). However, in general, this would be unsatisfactorily strong, because for example even though p and $p \wedge p$ are not syntactically identical, we arguably still can reason validly from 'Agent a knows that p' to 'Agent a knows that $p \wedge p$' and vice versa.[32] In this case being syntactically identical is not the weakest condition for the inferences to hold, in contradiction to (ii).

2.2 There is exactly one co-hyperintensionality relation

Our goal of this section is to show that for every language there is exactly one co-hyperintensionality relation $\sim$ as problem 2.1.1 asks us to find, and that this relation can be defined in terms of validity. The aim of our outline is instructiveness and not brevity: Rather than just giving an unmotivated proof, we want to provide more general notions that can be applied beyond the proof. In section 2.2.1, we start by defining the notion of a language that the problem requires. In section 2.2.2, we define what a co-hyperintensionality relation for a single operator is and prove that there always exists a unique one. We use this in section 2.2.3 to define what it is to be a co-hyperintensionality relation for the whole language (in the sense of our problem), and show that there always exists a unique one that is definable in terms of validity.

[32] For more reasons for coarse-graining see Bjerring and Schwarz (2017, sec. 4).

2.2.1 The notion of a language underlying the problem

The problem of co-hyperintensionality talks about languages, and in order to even formulate the problem we need to assume that languages come with a set of sentences with operators on it together with a relation of validity. So we *have to* make these assumptions about languages, but we also don't want to assume more in order to capture *all* the languages that the problem talks about. This gives us the following notion of a language that we are going to work with.

Definition 2.2.1 (Language)**.** We call the tripel $\mathcal{L} = (S, \Omega, \Leftrightarrow)$ a *language*, where

- S is a countable[33] set of sentences created by some syntax.
- Ω is a set of operators of finite arity on S. An n-ary operator ∇ is a function $\nabla : S^n \to S$. An operator of arity 0 is a propositional constant.
- $\Leftrightarrow$ is a binary relation on S that is reflexive, symmetric and transitive (and hence an equivalence relation).

Three remarks. First, instead of 'language' we also could have chosen the name 'logic' or 'the core of a language'. This notion of a language is the most general notion that the problem allows for, and which we hence should adopt. But there also is another, independent reason for working with this notion. In general, a language can be more complex than what is represented by the triple $(S, \Omega, \Leftrightarrow)$. For example, its sentences can be further structured and contain quantifiers and subject- and predicate-clauses, or it could take contexts of utterance into account. By adopting the above notion of a language, and thus individuating languages only up to their core, we avoid blurring our problem by other problems that come along with more complex languages. We leave it to future work to investigate in detail the special case of first-order languages or the setting where the relation of co-hyperintensionality takes the context of utterance into account.

Second, let's understand why $\Leftrightarrow$ should be an equivalence relation. The intended interpretation of $\Leftrightarrow$ is that it holds between sentences φ and ψ if and only if ψ follows validly from φ and vice versa. Because of the 'vice versa' clause, $\Leftrightarrow$ should be symmetric. Further, since validity surely is reflexive, we demand $\Leftrightarrow$ to be reflexive. Finally, we assume here that transitivity is an essential property of validity and hence demand $\Leftrightarrow$ to be transitive. This is a standard assumption and it is uncontested for many domains, but for some domains, it has been objected (see e.g. Cobreros et al. 2012; Ripley 2013). For reasons of space we won't discuss this assumption here, and only at the end we formulate the further question of what happens if we drop it.

Third, formally speaking, a language $(S, \Omega, \Leftrightarrow)$ has the structure of a universal algebra (together with the additional equivalence relation $\Leftrightarrow$). In principle, this allows for a fruitful application of the theory of universal algebras to these languages (cf. Humberstone 2015).

Let's consider classical propositional logic (CPL) as an example for a language in our sense (which we will use several times below). There, the set of sentences S is recursively defined by: all propositional variables $p, q, r, p_0, p_1, p_2, \ldots$ are in S; $\bot$ and $\top$ is in S; and if φ and ψ are in S, then $\neg\varphi$ and $\varphi \wedge \psi$ are in S. Now, the set of operators Ω contains the following: $\bot$ and $\top$ are 0-ary operators, $\neg\cdot : S \to S, \varphi \mapsto \neg\varphi$

[33] Though, nothing hinges on the countability assumption.

is a 1-ary operator, and $\wedge(\cdot,\cdot) : S^2 \to S, (\varphi,\psi) \mapsto \varphi \wedge \psi$ is a 2-ary operator. The relation $\Leftrightarrow$ is the usual equivalence relation: $\varphi \Leftrightarrow \psi$ iff any valuation V assigns φ and ψ the same truth-value.

2.2.2 Co-hyperintensionality for an operator

In this subsection, we define what it means, given a language $(S, \Omega, \Leftrightarrow)$, to be a co-hyperintensionality relation for a fixed operator ∇ of the language (definition 2.2.3). The main result is that there is a unique such ∇-co-hyperintensionality relation and that it can be characterized in terms of the validity relation $\Leftrightarrow$ (theorem 2.2.6).

Given a language $(S, \Omega, \Leftrightarrow)$, we want to find a co-hyperintensionality relation among all the binary relations on S. But in fact, for philosophical reasons, we demand that any co-hyperintensionality relation, if it exists, should be reflexive: Syntactic identity should entail same meaning in whatever sense. (Innocuous as this may seem, it is still a non-vacuous assumption as we'll see in proposition 2.2.8.) Hence we want to find a co-hyperintensionality relation in the set

$$\mathbf{R}_S := \{R \subseteq S \times S \mid R \text{ is reflexive}\}.$$

We start by ordering the relations in $\mathbf{R}_S$ by how fine-grained they are. This is an important notion since – as we saw in the formulation of our problem – co-hyperintensionality relations have to be fine-grained enough but also not too fine-grained.

Definition 2.2.2 ($\approx_1$ is less fine-grained than $\approx_2$)**.** Let $(S, \Omega, \Leftrightarrow)$ be a language. Given $\approx_1, \approx_2 \in \mathbf{R}_S$, we say that $\approx_1$ *is less fine-grained than* $\approx_2$ (in signs $\approx_1 < \approx_2$) if

(i) $\forall \varphi, \psi \in S:$ If $\varphi \not\approx_1 \psi$, then $\varphi \not\approx_2 \psi$, and

(ii) $\exists \varphi_0, \psi_0 \in S : \varphi_0 \approx_1 \psi_0$ and $\varphi_0 \not\approx_2 \psi_0$.

That is, by (i), $\approx_2$ is as fine-grained as $\approx_1$ (or, equivalently, whenever there is a $\approx_1$-difference, there also is a $\approx_2$-difference), and by (ii), $\approx_2$ is also properly finer than $\approx_1$ (there is a $\approx_2$-difference that is not seen by $\approx_1$).

While definition 2.2.2 arguably is the most natural way to say that $\approx_1$ is less fine-grained than $\approx_2$, it is equivalent to simple set-theoretic (reversed) inclusion: For all $\approx_1, \approx_2 \in \mathbf{R}_S$ we have: $\approx_1 < \approx_2$ iff $\approx_1 \supsetneq \approx_2$. Thus, the granularity-relation $<$ allows us to order $\mathbf{R}_S$: $(\mathbf{R}_S, <)$ is a strict partial order and $(\mathbf{R}_S, \leqslant)$ is a partial order. The identity relation is the greatest element of these orders, and the trivial relation $S \times S$ is the least element.[34]

Now we can properly define what it means for a relation to be a co-hyperintensionality relation for a given operator.

Definition 2.2.3 (∇-co-hyperintensionality relation)**.** Let $(S, \Omega, \Leftrightarrow)$ be a language. Given $\nabla \in \Omega$ with arity $m \in \mathbb{N}$ we call $\approx \in \mathbf{R}_S$ a ∇-co-hyperintensionality relation if $\approx$ is $\leqslant$-minimal with the following property

[34] Thus, $\mathbf{R}_S$ even is a distributive lattice under the operations $\cap$ and $\cup$. Yet, we can't use the set-theoretic complement operation to get a boolean algebra, since the complement of the identity relation is not reflexive anymore.

(A) For all $\varphi_1, \ldots, \varphi_m, \psi_1, \ldots, \psi_m \in S$: If $\varphi_i \approx \psi_i$ for all $i = 1, \ldots, m$, then $\nabla(\varphi_1, \ldots, \varphi_m) \Leftrightarrow \nabla(\psi_1, \ldots, \psi_m)$.

Given the arity is clear from the context, we also write $\bar{\varphi} := (\varphi_1, \ldots, \varphi_m)$ and abbreviate "$\varphi_i \approx \psi_i$ for all $i = 1, \ldots, m$" by "$\bar{\varphi} \approx \bar{\psi}$". Thus, condition (A) above reads $\bar{\varphi} \approx \bar{\psi} \Rightarrow (\nabla(\bar{\varphi}) \Leftrightarrow \nabla(\bar{\psi}))$. The minimality-condition says that if there is another relation $\approx' \in \mathbf{R}_S$ that has property (A), and if $\approx' \leqslant \approx$, then $\approx' = \approx$.

We convince us that this definition indeed captures what we are looking for. If $\approx$ is a ∇-co-hyperintensionality relation according to definition 2.2.3 (regardless of whether it exists), then it is a relation that fulfills the conditions (i) and (ii) of problem 2.1.1 for the fixed operator ∇. This is because condition (A) just ensures that $\approx$ meets (i) and the condition that $\approx$ is $\leqslant$-minimal in having (A) is the most natural way to make precise the demand of (ii), that $\approx$ is not unnecessarily fine-grained. (Remember that in this subsection we're only concerned in finding a co-hyperintensionality relation for a single operator. Only in the next subsection we will be concerned with co-hyperintensionality relations *proper*, that is, relations that fulfill (i) and (ii) for every operator).[35]

Now, the most pressing question is whether ∇-co-hyperintensionality relations exist. This is indeed the case: We will define one below (definition 2.2.4) of which we will prove that it is a ∇-co-hyperintensionality relation (lemma 2.2.5) and that it even is the only one (theorem 2.2.6).

Definition 2.2.4. ($\approx_\nabla$) Let $(S, \Omega, \Leftrightarrow)$ be a language. Let $\nabla \in \Omega_S$ be an m-ary operator. Introducing some useful notation we set for $\bar{\chi} \in S^m$, $\varphi \in S$ and $i \in \{1, \ldots, m\}$

$$\bar{\chi}[\varphi/i] := (\chi_1, \ldots, \chi_{i-1}, \varphi, \chi_{i+1}, \ldots, \chi_m) \in S^m.$$

Thus we can define the relation

$$\approx_\nabla := \Big\{ (\varphi, \psi) \in S \times S \mid \forall \bar{\chi} \in S^m, \forall i \in \{1, \ldots, m\} : \nabla(\bar{\chi}[\varphi/i]) \Leftrightarrow \nabla(\bar{\chi}[\psi/i]) \Big\}.$$

We observe that $\approx_\nabla$ indeed is a ∇-co-hyperintensionality relation.

Lemma 2.2.5. *Let $(S, \Omega, \Leftrightarrow)$ be a language. For every operator $\nabla \in \Omega$ we have*

(i) $\approx_\nabla \in \mathbf{R}_S$.

(ii) $\approx_\nabla$ *has property (A):* $\forall \bar{\varphi}, \bar{\psi} \in S^m : \bar{\varphi} \approx_\nabla \bar{\psi} \Rightarrow (\nabla(\bar{\varphi}) \Leftrightarrow \nabla(\bar{\psi}))$.

(iii) $\approx_\nabla$ *is a ∇-co-hyperintensionality relation.*

Proof of 2.2.5. (i) is clear as $\Leftrightarrow$ is reflexive. For (ii), let $\bar{\varphi}, \bar{\psi} \in S^m$ such that $\bar{\varphi} \approx_\nabla \bar{\psi}$. Since $\varphi_1 \approx_\nabla \psi_1$ we have (choosing the i in the definition of $\approx_\nabla$ to be 1 and $\bar{\chi}$ to be $\bar{\varphi}$) that

$$\nabla(\varphi_1, \varphi_2, \varphi_3 \ldots, \varphi_m) \Leftrightarrow \nabla(\psi_1, \varphi_2, \varphi_3 \ldots, \varphi_m).$$

[35] We can already see that any ∇-co-hyperintensionality relation $\approx$ is not only reflexive by definition, but also symmetric and transitive: Because if not, consider its symmetric or transitive closure and get a contradiction to minimality.

Since $\varphi_2 \approx_\nabla \psi_2$ we further get (choosing the i to be 2 and $\bar{\chi}$ to be $(\psi_1, \varphi_2, \varphi_3, \ldots, \varphi_m)$) that

$$\nabla(\psi_1, \varphi_2, \varphi_3 \ldots, \varphi_m) \Leftrightarrow \nabla(\psi_1, \psi_2, \varphi_3 \ldots, \varphi_m).$$

We repeat this process and finally get

$$\nabla(\psi_1, \psi_2, \ldots, \psi_{m-1}, \varphi_m) \Leftrightarrow \nabla(\psi_1, \psi_2, \ldots, \psi_{m-1}, \psi_m).$$

Putting all the equivalences we gathered together and using the transitivity of $\Leftrightarrow$ we get

$$\nabla(\varphi_1, \varphi_2, \ldots, \varphi_m) \Leftrightarrow \nabla(\psi_1, \psi_2, \ldots, \psi_m),$$

that is $\nabla(\bar{\varphi}) \Leftrightarrow \nabla(\bar{\psi})$, as wanted.

For (iii), it remains to show that $\approx_\nabla$ is $\leqslant$-minimal in having the property (A). Assume for contradiction, that there is a $\approx \in \mathbf{R}_S$ with property (A) and $\approx < \approx_\nabla$. By the latter, there are $\varphi, \psi \in S$ such that $\varphi \approx \psi$ but $\varphi \not\approx_\nabla \psi$. The latter again means that there is a $\bar{\chi} \in S^m$ and an $i \in \{1, \ldots, m\}$ such that $\nabla(\bar{\chi}[\varphi/i]) \not\Leftrightarrow \nabla(\bar{\chi}[\varphi/i])$. Since $\approx$ has property (A) we thus get $\bar{\chi}[\varphi/i] \not\approx \bar{\chi}[\psi/i]$. Since $\chi_j \approx \chi_j$ for $j \in \{1, \ldots, m\} \setminus \{i\}$ (because co-hyperintensionality relations are reflexive), we have $\varphi \not\approx \psi$ ↯. □

In the proof of (ii), we essentially used the transitivity of $\Leftrightarrow$, and this will be the only essential use. Now we're ready to show that $\approx_\nabla$ is not just *a* but also *the* unique ∇-co-hyperintensionality relation.

Theorem 2.2.6 (Unique existence of ∇-co-hyperintensionality)**.** *Let $(S, \Omega, \Leftrightarrow)$ be a language. Let $\nabla \in \Omega$ be an m-ary operator. Then there is exactly one ∇-co-hyperintensionality relation, namely $\approx_\nabla$.*

Proof of 2.2.6. We already know by lemma 2.2.5 that $\approx_\nabla$ is a ∇-hyperintensionality relation. Hence it remains to show that it is unique. So let $\approx \in \mathbf{R}_S$ be a ∇-hyperintensionality relation. We have to show $\approx = \approx_\nabla$. It's easy to see that $\approx \subseteq \approx_\nabla$, that is, $\approx_\nabla \leqslant \approx$. Since $\approx_\nabla$ has property (A) and $\approx$ is $\leqslant$-minimal in having (A), $\approx_\nabla = \approx$, as wanted. □

This was the most crucial result of this section. We end it by two other results indicating that 2- and more-ary operators are far more ill-behaved than 1-ary operators. The first one shows that in the case of 1-ary operators the minimality condition in the definition of a ∇-co-hyperintensionality relation amounts to demanding the converse of property (A), while this is not the case for multi-ary operators.

Proposition 2.2.7. *Let $(S, \Omega, \Leftrightarrow)$ be a language. Let $\nabla \in \Omega$ be an m-ary operator. Let $\approx \in \mathbf{R}_S$ be a ∇-hyperintensionality relation. Then we have*

(i) If $m = 1$, then for all $\varphi, \psi \in S$:

$$\varphi \approx \psi \textit{ iff } \nabla(\varphi) \Leftrightarrow \nabla(\psi). \tag{2.1}$$

*(ii) For $m > 1$ (i) doesn't hold anymore. That is, in general we do **not** have that for all $\bar{\varphi}, \bar{\psi} \in S^m$*

$$\bar{\varphi} \approx \bar{\psi} \textit{ iff } \nabla(\bar{\varphi}) \Leftrightarrow \nabla(\bar{\psi}). \tag{2.2}$$

Proof of 2.2.7. Ad (i). This can be shown even without theorem 2.2.6, however, for brevity we won't do this here. By theorem 2.2.6, $\approx = \approx_\nabla$. The left-to-right direction is just property

(A), for the other direction we have that $\nabla(\varphi) \Leftrightarrow \nabla(\psi)$ just means $\varphi \approx_\nabla \psi$ if ∇ is 1-ary.

Ad (ii). We take $(S, \Omega, \Leftrightarrow)$ to be the language of classical propositional logic. We consider the operator $\wedge$ and the (only) $\wedge$-hyperintensionality relation $\approx_\wedge$. Then we have for $\bar{\varphi} := (p, \neg p)$ and $\bar{\psi} := (q, \neg q)$ that

$$\wedge(\bar{\varphi}) \;\Leftrightarrow\; p \wedge \neg p \;\Leftrightarrow\; q \wedge \neg q \;\Leftrightarrow\; \wedge(\bar{\psi}).$$

However, we do not have that $p \approx_\wedge q$, because for $\bar{\chi} := (p, p)$ and $i = 2$ we have $\wedge(\bar{\chi}[p/i]) = \wedge(p, p) = p \wedge p \not\Leftrightarrow p \wedge q = \wedge(p, q) = \wedge(\bar{\chi}[q/i])$. Consequently, we have $\wedge(\bar{\varphi}) \Leftrightarrow \wedge(\bar{\psi})$ but $\bar{\varphi} \not\approx_\wedge \bar{\psi}$, hence (2.2) indeed fails. □

The second result is that the philosophically emphatically demanded assumption that co-hyperintensionality relations should be reflexive is not vacuous.

Proposition 2.2.8. *Assume that we don't demand co-hyperintensionality relations to be reflexive.*[36] *Then there is a language $(S, \Omega, \Leftrightarrow)$ and a $\nabla \in \Omega$ such that there is a non-reflexive ∇-co-hyperintensionality relation $\approx$. Moreover, the reflexive closure of $\approx$ is not a ∇-co-hyperintensionality relation anymore (it is not minimal and it fails to have the substitution property).*

Proof of 2.2.8. Let $S = \{p_0, p_1, p_2, \ldots\}$ and $\Omega = \{\nabla\}$ where ∇ is defined as follows. Let $\nabla_0 : S \times S \to S \setminus \{p_0\}$ be an injection and define $\nabla : S \times S \to S$ as the function that maps (p_0, p_0) and (p_1, p_1) to p_0 and otherwise is like ∇_0. Let $\Leftrightarrow$ be the syntactic identity relation (so $\Leftrightarrow$ is as in CPL when restricted to our impoverished setting).

Consider the irreflexive relation $\approx = \{(p_0, p_1)\}$. By construction, $\approx$ has property (A). Moreover, we claim that $\approx$ is minimal in having property (A). Indeed, if not, there is $(p_i, p_j) \neq (p_0, p_1)$ such that $\nabla(p_0, p_i) = \nabla(p_1, p_j)$, and hence, by injectivity of ∇ outside $\{(p_0, p_0), (p_1, p_1)\}$, either

- $(p_0, p_i) = (p_1, p_j)$, in contradiction to $p_0 \neq p_1$, or
- $(p_0, p_i) = (p_1, p_1)$ and $(p_1, p_j) = (p_0, p_0)$, in contradiction to $p_0 \neq p_1$, or
- $(p_0, p_i) = (p_0, p_0)$ and $(p_1, p_j) = (p_1, p_1)$, in contradiction to $(p_i, p_j) \neq (p_0, p_1)$.

By minimality of $\approx$, the reflexive closure $\approx_r$ of $\approx$ cannot be a ∇-co-hyperintensionality relation anymore. In fact, $\approx_r$ not only fails to be minimal, it also fails to have the substitution property: $p_0 \approx_r p_0$ and $p_0 \approx_r p_1$ but $\nabla(p_0, p_0) \not\Leftrightarrow \nabla(p_0, p_1)$. □

2.2.3 Co-hyperintensionality for a language

So far we've been considering hyperintensionality relations only for a given operator. However, what our problem asks us to find is a hyperintensionality relation for all operators of the language. Hence we define:

Definition 2.2.9 (Co-hyperintensionality relation)**.** Let $(S, \Omega, \Leftrightarrow)$ be a language. A relation $\approx \in \mathbf{R}_S$ is a *co-hyperintensionality relation*, if $\approx$ is $\leqslant$-minimal with the property

(B) For every operator $\nabla \in \Omega_S$, for all $\varphi_1, \ldots, \varphi_m, \psi_1, \ldots, \psi_m \in S$: If $\varphi_i \approx \psi_i$ for all $i = 1, \ldots, m$, then $\nabla(\varphi_1, \ldots, \varphi_m) \Leftrightarrow \nabla(\psi_1, \ldots, \psi_m)$.

Now we can prove the result we were working toward all along.

[36] That is, given a language $(S, \Omega, \Leftrightarrow)$, call *any* binary relation on S a ∇-co-hyperintensionality relation, if it is minimal in satisfying property (A).

Theorem 2.2.10 (Unique existence of co-hyperintensionality)**.** *Let* $(S, \Omega, \Leftrightarrow)$ *be a language. Then there is exactly one co-hyperintensionality relation, namely the relation*

$$\sim := \bigcap_{\nabla \in \Omega} \approx_\nabla \in \mathbf{R}_S.$$

Proof of 2.2.10. In the same way as in the proof of theorem 2.2.6 we first show that $\sim$ is a co-hyperintensionality relation and then its uniqueness. □

Section 2.3 (after proposition 2.3.2) will shed some light on why the co-hyperintensionality relation $\sim$ is the intersection over the operator-co-hyperintensionality relations.

To get a feel for this theorem, let's consider an example for the co-hyperintensionality relation of a language. For the sake of simplicity, we consider classical propositional logic and see that its co-hyperintensionality relation is – non-surprisingly – that of extensional equivalence.

Example 2.2.11 (Co-hyperintensionality in CPL)**.** Let $(S, \Omega, \Leftrightarrow)$ be the language of CPL. We show that the hyperintensionality relation $\sim$ of propositional logic is $\Leftrightarrow$, that is, that $\Leftrightarrow = \sim = \approx_\wedge \cap \approx_\neg$. To show this, we show that $\approx_\wedge = \Leftrightarrow$ and $\approx_\neg = \Leftrightarrow$.

($\approx_\wedge = \Leftrightarrow$). Fixing some arbitrary $\varphi, \psi \in S$, we have to show

$$\varphi \Leftrightarrow \psi \text{ iff } \forall \bar{\chi} \in S^2, \forall i \in \{1,2\} : \wedge(\bar{\chi}[\varphi/i]) \Leftrightarrow \wedge(\bar{\chi}[\psi/i])$$

that is, we have to show

$$\varphi \Leftrightarrow \psi \text{ iff } \forall \chi_1, \chi_2 \in S : \begin{cases} \varphi \wedge \chi_2 \Leftrightarrow \psi \wedge \chi_2 & \text{(a), and} \\ \chi_1 \wedge \varphi \Leftrightarrow \chi_1 \wedge \psi & \text{(b)} \end{cases} \tag{2.3}$$

Indeed, the left-to-right direction of (2.3) holds because $(\varphi \leftrightarrow \psi) \rightarrow (\varphi \wedge \chi \leftrightarrow \psi \wedge \chi)$ is a propositional tautology. For the right-to-left direction of of (2.3) we choose $\chi_2 := \varphi$ and $\chi_1 := \psi$ and we get

$$\varphi \;\Leftrightarrow\; \varphi \wedge \underbrace{\varphi}_{=\chi_2} \overset{\text{(a)}}{\Leftrightarrow} \psi \wedge \underbrace{\varphi}_{=\chi_2} \;\Leftrightarrow\; \underbrace{\psi}_{=\chi_1} \wedge \varphi \overset{\text{(b)}}{\Leftrightarrow} \underbrace{\psi}_{=\chi_1} \wedge \psi \;\Leftrightarrow\; \psi.$$

($\approx_\neg = \Leftrightarrow$). Fixing some arbitrary $\varphi, \psi \in S$, we have to show $\varphi \Leftrightarrow \psi$ iff $\neg\varphi \Leftrightarrow \neg\psi$, and this clearly holds. ◁

We end with a first application of theorem 2.2.10 showing that substitution salva veritate and co-hyperintensionality do indeed coincide—which, as we saw, is prima facie not obvious but usually tacitly assumed. For this we first need two definitions.

Given a language $(S, \Omega, \Leftrightarrow)$ we call a relation $\approx \in \mathbf{R}_S$ a *substitution salva veritate relation* if $\approx$ is $\leqslant$-minimal with the property

$$\forall \chi \in S,\ \forall \varphi \text{ subformula}^{37} \text{of } \chi,\ \forall \psi \in S : \text{ If } \varphi \approx \psi, \text{ then } \chi \Leftrightarrow \chi[\psi/\varphi], \tag{2.4}$$

where $\chi[\psi/\varphi]$ is the formula resulting from replacing the subformula φ of χ by the formula ψ. Clearly, (2.4) is just the precise formulation of "φ and ψ are substitutable

salva veritate in any context".

Moreover, we say Ω is *closed under substitution operators* iff for all $\chi \in S$ and any $\varphi \in S$ that is a subformula of χ, the operator $\nabla_{\chi,\varphi}$ defined by $\nabla_{\chi,\varphi}(\psi) = \chi[\psi/\varphi]$ is in Ω. Intuitively, closing Ω under substitution operators doesn't "really" change the language. We neither will discuss here how to make this statement more precise, nor how the condition that Ω is closed under substitution operators (which is used in the announced proposition below) can be relaxed for specific languages.

Corollary 2.2.12 (Substition salva veritate and co-hyperintensionality coincide)**.** *Let $(S, \Omega, \Leftrightarrow)$ be a language. Assume that Ω is closed under substitution operators. Then for all $\approx \in \mathbf{R}_S$, $\approx$ is a substitution salva veritate relation iff $\approx$ is a co-hyperintensionality relation. In particular, by theorem 2.2.10, $\sim$ is the unique substitution salva veritate relation.*

Proof of 2.2.12. For $\approx \in \mathbf{R}_S$ it is straightforward to check that we have the following equivalences:

$\approx$ is a substitution salva veritate relation, i.e. $\approx$ is $\leqslant$-minimal with property (2.4)

iff[38] $\forall \varphi, \psi \in S\Big(\varphi \approx \psi$ iff $\forall \chi \in S\big($If φ subformula of χ, then $\chi \Leftrightarrow \chi[\psi/\varphi]\big)\Big)$

iff[39] $\forall \varphi, \psi \in S\Big(\varphi \approx \psi$ iff $\forall \nabla \in \Omega \forall \bar{\chi} \in S^m \forall i \in \{1, \dots, m\}\big(\nabla(\bar{\chi}[\varphi/i]) \Leftrightarrow \nabla(\bar{\chi}[\psi/i])\big)\Big)$

iff[40] $\approx$ is a co-hyperintensionality relation. □

2.2.4 Putting the results into perspective

In this section, we show in what sense our results so far provide a solution to the problem of co-hyperintensionality, and we indicate what the value of the results is.

The problem of co-hyperintensionality (problem 2.1.1) asked us to find a co-hyperintensionality relation given a language $(S, \Omega, \Leftrightarrow)$. In that sense, theorem 2.2.10 is the solution. But let's be more precise to understand in which sense this is a solution. In order to even pose the problem of co-hyperintensionality we need to have a notion of validity for statements involving hyperintensional operators. If we don't have validity, we cannot pose the problem. We showed that as soon as validity is available and the problem can be formulated, the problem has one unique solution. That is, once we know what validity is for statements involving hyperintensional operators, we get co-hyperintensionality for free via our results. In other words, we reduced co-hyperintensionality to validity.

Let's indicate the additional value of the results by replying to two worries that our results are only of little value.

The first worry says that maybe our results provide a formal solution to the problem of co-hyperintensionality, but they don't provide a philosophical solution.

Reply: In one sense the objection is right. We don't claim that the results provide a "contentful" insight into what hyperintensionality *really* is. For example, the results don't tell us whether "Bachelors are unmarried" is co-hyperintensional with "Unmarried men are unmarried". This is because that still depends on what operators we

[37] Since, for greater generality, we haven't specified a syntax of S we think of subformulas in terms of operators: For example, ψ is a subformula of $\nabla_1(\varphi, \nabla_2(\psi))$.

[38] Demanding the right-to-left direction in $\forall \varphi, \psi \in S(\dots)$ corresponds to demanding $\leqslant$-minimality.

[39] Given $\varphi, \psi \in S$ show that the two conditions on the right of the iff's are equivalent. From the lower to the upper we use that Ω is closed under substitution operators.

[40] By theorem 2.2.10.

allow in the language and – especially – on what the underlying validity relation is. In fact, a result on that level of abstraction cannot (and should not) deliver such a philosophically contentful discovery. However, this also points at a sense in which the objection is wrong. As mentioned, the results reduce co-hyperintensionality to validity, and this is a philosophical insight: It reduces the more complex concept of co-hyperintensionality to the simpler one of validity or truth-preservation. In other words, it shows that nothing interesting can be found on the level of co-hyperintensionality that cannot already be seen on the level of validity.

The second worry starts from the informal understanding of co-hyperintensionality as substitution salva veritate also in hyperintensional contexts, and then wonders: why all the fuss? We're looking for the minimal relation ensuring substitution salva veritate—but, well, then this relation just is the substitution salva veritate relation.

Reply: First of all, the objection is not entirely right: As already noted, for co-hyperintensionality we demand simultaneous n-place substitution, while for substitution salva veritate we only demand 1-place substitution. And this prima facie difference does indeed make quite a difference (cf. e.g. proposition 2.2.7). So from a purely formal point of view the key insight of theorem 2.2.10 is that we can find a relation that not only guarantees individual substitution (this is indeed trivial as the objection notes) but also uniform substitution. And corollary 2.2.12 tops this by showing that for structures $(S, \Omega, \Leftrightarrow)$ uniform substitution (that ensures the outcomes being in the $\Leftrightarrow$ relation) is essentially the same as individual substitution. However, looking at other domains (e.g. multivariable calculus in mathematics), we see that this is something we cannot expect in general.[41]

Regardless of whether this is just a formal subtlety or a result really worth its name, the value of the theorem consists not so much in what it states but more in its applications and implications. We'll see the applications in the next section when we apply the framework and the theorem to investigate the structure of hyperintensional operators. Some of the main (philosophical) implications are that the results clarify and regulate the concept of hyperintensionality. One such implication is – as already mentioned – that the results reduce co-hyperintensionality to validity. Other implications will be discussed in section 2.4. They include that co-hyperintensionality (as the literature defines the term) is language dependent and – in general – not a cognitively adequate individuation of content. (And the implications thus discovered didn't seem to me to be very clearly understood in the literature.)

2.3 The structure of hyperintensional operators: The granularity order

In this section, we want to apply the framework and results of the preceding section to investigate the structure of the set of operators—an important enterprise when

[41] In multivariable calculus, we can have a function $f : \mathbb{R}^n \to \mathbb{R}$ such that (i) "individually it is continuous", i.e. for any variable, f is continuous in that variable while the other variables are fixed, but (ii) "uniformly it is not continuous", i.e. f is not continuous (so there is a converging sequence of argument vectors where the f-values of the members of the sequence don't converge to the f-value of the limit of the sequence). (Note that the heuristic "uniformly f is continuous" should not be confused with the mathematical notion "f is uniformly continuous".)

understanding hyperintensionality. This set comes with a natural ordering that compares operators by their granularity, that is, by their ability to differentiate between sentences. Thus, one operator is more fine-grained than another if the one can see all the differences the other can see, but also can see more. We shall see that this ordering is a partial order but in general not linear: There are operators, even natural ones, that cannot be compared by granularity.

The idea of ordering operators by granularity is precisely defined as follows.

Definition 2.3.1 (Ordering operators by granularity)**.** Let $(S, \Omega, \Leftrightarrow)$ be a language. For $\nabla_1, \nabla_2 \in \Omega$ we set

$$\nabla_1 \leqslant \nabla_2 \text{ iff } \approx_{\nabla_1} \leqslant \approx_{\nabla_2} .$$

Hence $(\Omega, \leqslant)$ is a partially ordered set.

It would be particularly nice if we could compare any two operators by granularity. One obvious reason is that if $(\Omega, \leqslant)$ is a linear order, it is much easier to investigate (also philosophically speaking)—the structure of hyperintensionality would be far less complicated. Another reason is the following. We've seen in section 1.2.2 that hyperintensional operators are characterized either in the general or in the strict version. We can express these two characterizations as follows (where $\Box$ is an intensional operator):

(I) An operator ∇ is a hyperintensional operator, if ∇ can see differences intensional operators cannot see, that is, $\nabla \not\leqslant \Box$.

(II) An operator ∇ is a strictly hyperintensional operator, if ∇ is more fine-grained than intensional operators, that is, $\Box < \nabla$.

Now, if $(\Omega_S, \leqslant)$ were a linear order, then hyperintensionality (I) and strict hyperintensionality (II) would be equivalent. However, concerning this we have a negative result.

Proposition 2.3.2 (Non-linearity of operator ordering)**.** *There are languages* $(S, \Omega, \Leftrightarrow)$ *such that* $(\Omega, \leqslant)$ *is not a linear order.*

Proof of 2.3.2. To construct $(S, \Omega, \Leftrightarrow)$, start with the language of classical propositional logic $(S, \Omega', \Leftrightarrow)$. We fix three propositional variables p, q and r and define the 1-ary operators

$$\nabla_1(\varphi) := \begin{cases} \top & , \varphi = p \text{ or } \varphi = r \\ \bot & , \text{otherwise} \end{cases} \qquad \nabla_2(\varphi) := \begin{cases} \top & , \varphi = p \text{ or } \varphi = q \\ \bot & , \text{otherwise.} \end{cases}$$

Adding the operators ∇_1 and ∇_2 to Ω' won't change the set S (nor $\Leftrightarrow$) as their outputs are always already in S. Thus let $\Omega := \Omega' \cup \{\nabla_1, \nabla_2\}$. We claim that for the language $(S, \Omega, \Leftrightarrow)$, $(\Omega, \leqslant)$ is not a linear order. For this we have to show that there are two operators whose co-hyperintensionality relations are not comparable.

Indeed, we claim that $\approx_{\nabla_1} \not\leqslant \approx_{\nabla_2}$ and $\approx_{\nabla_2} \not\leqslant \approx_{\nabla_1}$, that is, that $\approx_{\nabla_2} \not\subseteq \approx_{\nabla_1}$ and $\approx_{\nabla_1} \not\subseteq \approx_{\nabla_2}$. We have

$$\begin{aligned} &\nabla_1(p) = \top \Leftrightarrow \top = \nabla_1(r) \\ &\nabla_1(p) = \top \not\Leftrightarrow \bot = \nabla_1(q) \end{aligned} \tag{2.5}$$

$$\begin{aligned} &\nabla_2(p) = \top \not\Leftrightarrow \bot = \nabla_2(r) \\ &\nabla_2(p) = \top \Leftrightarrow \top = \nabla_2(q). \end{aligned} \tag{2.6}$$

Now since $\approx_{\nabla_1}$ is the ∇_1-co-hyperintensionality relation and since ∇_1 is 1-ary, we have by

proposition 2.2.7 and (2.5) that $p \approx_{\nabla_1} r$ and $p \not\approx_{\nabla_1} q$. Analogously we get for $\approx_{\nabla_1}$ via (2.6) that $p \not\approx_{\nabla_2} r$ and $p \approx_{\nabla_2} q$. Hence we have $\approx_{\nabla_1} \not\subseteq \approx_{\nabla_2}$ and $\approx_{\nabla_2} \not\subseteq \approx_{\nabla_1}$, as wanted. □

This also explains why we had to take the intersection of all operator-co-hyperintensionality relations as the co-hyperintensionality relation $\sim$ of the entire language. If there is a unique maximal element in the set of co-hyperintensionality relations $\mathbf{H}_S$ (and hence also in $(\Omega, \leqslant)$), then this element is $\sim$. However, as $(\Omega, \leqslant)$ and hence also $(\mathbf{H}_S, \leqslant)$ is not linear, there could be several maximal elements. Hence we have to take $\sim$ as their intersection which then, in general, is not in $\mathbf{H}_S$ anymore.

The operators used to show that $(\Omega_S, \leqslant)$ is not a linear order are formally speaking very simple and thus provide a quick and incontestable way of providing two incomparable operators. However, they are, in a (philosophical) sense, not very natural. So one might conjecture that the co-hyperintensionality relations of conceptually and philosophically interesting operators still can be ordered linearly by granularity. However, this is not the case either as we will see in the following example using the knowledge and the explanation operator.

Example 2.3.3. We consider the natural language English and the set S of (sufficiently regimented) declarative sentences. We consider the following two 1-ary operators:

$$K(\varphi) := \text{Tarski knows that } \varphi$$
$$C(\varphi) := \varphi \text{ because } A \text{ is at the conference,}$$

where 'A' is the name of a person acquainted with Tarski, and – unbeknownst to Tarski – this person also goes by the name 'B'. We want to find sentences $\varphi_1, \psi_1, \varphi_2, \psi_2$ such that

$$K(\varphi_1) \Leftrightarrow K(\psi_1) \text{ and } K(\varphi_2) \not\Leftrightarrow K(\psi_2), \text{ and} \tag{2.7}$$
$$C(\varphi_1) \not\Leftrightarrow C(\psi_1) \text{ and } C(\varphi_2) \Leftrightarrow C(\psi_2). \tag{2.8}$$

Because then we can reason as in the proof of proposition 2.3.2: Via proposition 2.2.7 we get from the above that $\varphi_1 \approx_K \psi_1$ and $\varphi_2 \not\approx_K \psi_2$ and $\varphi_1 \not\approx_C \psi_1$ and $\varphi_2 \approx_C \psi_2$. Hence $\approx_K \not\subseteq \approx_C$ and $\approx_C \not\subseteq \approx_K$. So $K \not\leqslant C$ and $C \not\leqslant K$, that is, the operators K and C are not comparable in terms of granularity.

There are indeed such sentences. We set

$$\varphi_1 := A \text{ is at the conference,} \qquad \varphi_2 := A \text{ is fascinated by logic,}$$
$$\psi_1 := \text{'}A \text{ is at the conference' is true,} \qquad \psi_2 := B \text{ is fascinated by logic.}$$

This gives us (2.7): Tarski knows that A is at the conference if and only if he knows that 'A is at the conference' is true—since Tarski knows how truth works.[42] Moreover, Tarski knows that A is fascinated by logic, say, yet he doesn't know that B is fascinated by logic because he doesn't know that 'B' and 'A' denote the same person.

[42] So the notion of validity we assume here is something like an "outside of a philosophy classroom" notion of validity since it assumes some background knowledge (that Tarski knows how truth works). If we want to avoid this and deal with "pure logical validity", we can define the operator $K(\varphi)$ as "Tarski knows how truth works and he knows that φ".

We also get (2.8): We have that 'A is at the conference because A is at the conference' is false because an empirical fact cannot explain itself, while it is commonly held that

> 'A is at the conference' is true because A is at the conference.

is true. Moreover, whenever it is true that

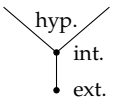
> A is fascinated by logic because A is at the conference,

it is arguably also true that

> B is fascinated by logic because A is at the conference,

and vice versa, because A and B are necessarily identical. ◁

Example 2.3.3 provided two operators that commonly are taken to be hyperintensional and showed that they are not comparable by granularity. It seems as if it is often implicitly assumed that we have extensional and intensional operators and only "above" them it is getting messy. That is, a common implicit assumption is that the order of operators looks roughly like this:



On this view, the definitions (I) and (II) of hyperintensional operators are still equivalent (which explains why these two conceptions of hyperintensional operators are usually not explicitly distinguished). However, we shall now see that this view is not tenable either: There are natural operators even on the level of intensionality that are not comparable.

Example 2.3.4. We still consider the natural language English and now the following two 1-ary operators:

$\nabla(\varphi) :=$ Frege inexactly makes φ true
$\Box(\varphi) :=$ It is necessary that φ.

For the notion of inexact truthmaking, see e.g. Fine (2016c, p. 3). As in example 2.3.3, we're done if we can find sentences $\varphi_1, \psi_1, \varphi_2, \psi_2$ such that

$$\nabla(\varphi_1) \Leftrightarrow \nabla(\psi_1) \text{ and } \nabla(\varphi_2) \not\Leftrightarrow \nabla(\psi_2) \text{ and} \tag{2.9}$$
$$\Box(\varphi_1) \not\Leftrightarrow \Box(\psi_1) \text{ and } \Box(\varphi_2) \Leftrightarrow \Box(\psi_2). \tag{2.10}$$

Because this will give us that $\nabla \not\leqslant \Box$ and $\Box \not\leqslant \nabla$, that is, the operators ∇ and $\Box$ are not comparable in terms of granularity. There are indeed such sentences:

$\varphi_1 :=$ Frege exists,
$\psi_1 :=$ Frege exists or Frege doesn't exist,
$\varphi_2 :=$ Frege exists or Frege doesn't exist,
$\psi_2 :=$ Russell exists or Russell doesn't exist.

This gives us (2.9): Frege makes 'Frege exists' exactly true, and hence he also makes it inexactly true, and Frege inexactly makes true 'Frege exists or Frege doesn't exist'. Moreover, Frege doesn't inexactly make true that 'Russell exists or Russell doesn't exist' because Frege is not the right truthmaker (he is neither wholly nor partially relevant to the statement).

We also get (2.10): It is not necessary that Frege exists while φ_2 and ψ_2 are both necessary as these are tautologies. ◁

This example explicitly provides an operator (namely inexact truthmaking) that is hyperintensional (I) but not strictly hyperintensional (II).

2.4 Further philosophical consequences and open questions

As already discussion in section 2.2.4, the problem of co-hyperintensionality asked us to find a co-hyperintensionality relation given a language $(S, \Omega, \Leftrightarrow)$. And in that sense, theorem 2.2.10 is the solution. In the last section, we've seen some applications of that result for the ordering of operators by granularity. In this section, we want to discuss some further philosophical implications of the above results and mention some questions that have to be tackled next.

2.4.1 Co-hyperintensionality, equipollence and cognitive adequacy: An impossibility result

We find that co-hyperintensionality and Fregean equipollence are jointly inconsistent criteria for sameness of content. This provides an interpretation of our findings as an impossibility result: Hyperintensionality (as commonly defined) cannot provide a cognitively adequate individuation of content. Since this was one of the reasons to go hyperintensional in the first place, this should make us rethink the notion of hyperintensionality.

One motivation for providing hyperintensional content is to find a notion of content that both avoids the problems of intensions – for example, that they collapse all necessary truths – and that is individuated by the co-hyperintensionality relation. If this succeeds, then having the same content coincides with being co-hyperintensional. In short (and justified by proposition 2.2.12): Synonymy coincides with substituition salva veritate.[43] A problem arises since this is not the only alleged principle for individuation of content. Famously, Frege (1891, p.14; 1979, p.197) put forward the equipollence criterion: Two sentences have the same content if and only if they are equipollent, that is, roughly, one could not rationally regard one as true and the other as false (or, better, as not true). This is commonly considered to be on the right track in spelling out a cognitively adequate individuation of content.

Bjerring and Schwarz (2017, sec. 5) argue that being equipollent is not transitive: Take a lengthy sequence of sentences $\varphi_1, \ldots, \varphi_n$ such that adjacent sentences are triv-

[43] I find the established terminology that takes 'co-hyperintensionality' to mean substitution salva veritate in any context a bit misleading. Taken literally, we should call two expressions 'co-hyperintensional' if they have the same hyperintension (whatever these might be). The hope is that these two criteria will coincide, but at least prima facie it is not clear that the correct individuation of hyperintenions coincides with the co-hyperintensionality relation, i.e., the relation that ensures substitution salva veritate in any context.

ially equivalent but the first and the last sentence aren't—for example, take a sequence of algebraic equations where adjacent ones are trivial equivalence-transformations of each other while it is a highly non-trivial step from the first to the last equation. Now, since adjacent sentences are trivially equivalent they are equipollent, and since φ_1 and φ_n are not trivially equivalent, they are not equipollent. Hence equipollence is not transitive.

It should be noted that Fregean equipollence – at least in an unmodified form – has its problems as a criterion for content identity. One immediate problem is that all trivial truths (like "apples are apples" and "bananas are bananas") are equipollent but intuitively can still differ in content. See Schellenberg (2012) for a discussion of further problems.[44]

Regardless of the problems of equipollence, it seems to be a commonplace that any cognitively adequate "individuation" of content has to be intransitive.[45] Because our rationality is bounded, we can cognize the content identity of sentences only within a certain bound of complexity. That is, there are sequences of sentences $\varphi_1, \ldots, \varphi_n$ such that neighboring sentences are regarded as cognitively synonymous while the first and the last are not. As examples of necessarily true sentences, we can again take the algebraic equations from above. Or a long logical deduction (cf. e.g. Jago 2014, p. 11). Or we can take the sentences $\varphi_i :=$"i in binary is i_2" for i ranging from 1 to a large natural number. We can also take John Conway's famous cellular automaton called *Game of Life* with a fixed simple initial state and consider the sentences $\varphi_i :=$ "A_i is the state of this game of life after i many steps". (Similarly for other complex systems where each update step is trivial.) Let's consider examples where neighboring sentences are not necessarily equivalent but "cognitively synonymous" in the sense that we usually take them to express the same thing. We can take a long list of synonymous adjectives – for example, some of the 2,259 alleged synonyms of "strong"[46] –, then build sentences by ascribing the adjectives to the same noun, and finally put sentences with very similar adjectives next to each other—thus the sentence at the beginning of the sequence and the one at the end won't be regarded as synonymous anymore.

So Fregean equipollence and – more generally – any cognitively adequate "individuation" of content cannot be transitive. However, our results show that for every language there is one unique co-hyperintensionality relation (which is the relation that ensures substitution salva veritate in any context). And we readily see that this relation is always transitive since validity (or truth-preservation) is (assumed to be) transitive. Thus, our results show that if validity is transitive, the substitution salva veritate criterion and the equipollence criterion are jointly inconsistent criteria for sameness of content.[47] And, more generally, we see that the individuation of content provided by co-hyperintensionality cannot be cognitively adequate.[48]

[44] The reconstruction of Fregean equipollence that she ends up with seems to be transitive. This avoids many problems but it also looses the appeal of Fregean equipollence that it describes a cognitively adequate individuation of content.

[45] For example, see – again – Bjerring and Schwarz (2017, sec. 5) who also provide more general arguments for the intransitivity of content that do not hinge on idiosyncrasies of Fregean equipollence.

[46] Provided by the online thesaurus "Power Thesaurus": `https://www.powerthesaurus.org/strong` (retrieved May 23, 2017).

[47] Here we won't discuss how this inconsistency can be resolved within Frege's doctrine.

[48] To wit, here is a cheap argument that co-hyperintensionality implies sameness of content (so that the

This has many consequences. Here we wish to stress four in particular. First, this provides a reading of our findings as an impossibility result: If we look for a co-hyperintensionality relation as specified in the problem, then we find that there is exactly one. However, the only one that exists cannot be a cognitively adequate one.

Second, we thus should rethink the purpose (and hence the notion) of co-hyperintensionality: There are various arguments for hyperintensional content, that is, arguments for a notion of content that is more fine-grained than the intensions of possible worlds semantics. Several of them stem from the fact that possible worlds intensions don't always capture the cognitive significance of sentences. If hyperintensional content is to address this issue, then, by the just mentioned impossibility interpretation, co-hyperintensionality as defined in the problem is a non-starter: it should individuate closer to cognitive significance as it currently does. This, however, would require re-defining co-hyperintensionality.[49]

Third, both co-hyperintensionality (i.e. substitution salva veritate) and Fregean equipollence (or, more generally, "cognitive synonymy") are independently motivated criteria for individuation of content. However, since they are mutually exclusive, a choice for either of them has to be well justified—for example, by the nature and purpose of the content that they individuate (also cf. section 5.4).[50]

Fourth, if the transitivity of validity entails the inconsistency of these criteria, this shifts the attention to non-transitive validity relations (or conditionals). This will be taken up in question 10 in section 2.4.4 below.

2.4.2 A universal co-hyperintensionality relation or a plurality of languages?

We discuss the possibility of a (genuine) plurality of co-hyperintensionality relations given a plurality of languages.

We've seen that for every language there is a unique co-hyperintensionality relation. But we've also seen that this relation depends very much on the language, that is, for different languages the co-hyperintensionality relations might look very differently. In a way, this is not surprising at all: For a language that contains some sort of quoting operator the co-hyperintensionality relation has to be the identity relation, and if a language contains no such operator, the co-hyperintensionality relation will be considerably more coarse-grained. However, especially if co-hyperintensionality is tightly linked with sameness of content (which is not unproblematic as seen in the previous subsection), there is the intuition that there really should be only one co-hyperintensionality relation, (quite) independent of the language, that deserves the name: Because – or so the intuition goes – there is a language-independent notion of content with a fixed granularity.

---

non-adequateness stems from co-hyperintensionality being too fine-grained). Consider the operator ∇ of the natural language English where $\nabla(\varphi, \psi) :=$ φ has the same content as ψ. Assume two sentences φ and ψ stand in the co-hyperintensionality relation of the language, i.e. $\varphi \sim \psi$. Hence $\varphi \approx_\nabla \psi$ and hence $\nabla(\varphi, \varphi) \Leftrightarrow \nabla(\varphi, \psi)$. As $\nabla(\varphi, \varphi)$ is true, we also have $\nabla(\varphi, \psi)$, i.e. φ and ψ have the same content, as wanted.

[49] This is – to the best of my knowledge – a novel criticism of co-hyperintensionality. Other critical assessments of the notion of co-hyperintensionality – or intermediate granularity of content – can be found, for example, in Faroldi (2016) and Bjerring and Schwarz (2017).

[50] In many discussions, this inconsistency always seems about to surface: e.g. in Schellenberg (2012) or in Penco (2013, esp. p. 57) discussing the debate between Kripke (2008) and Künne (2010). However, to the best of my knowledge, it hasn't been explicitly stated and/or argued for before.

There are at least four ways to respond to this. The first way defends the intuition by saying that there might be some artificial languages that come with their own (artificial) co-hyperintensionality relations, but that the proper co-hyperintensionality relation that deserves its name is that of the language of sufficiently regimented English (or any other natural language).

The second way aims to show that the intuition of a universal co-hyperintensionality relation is ill-conceived and embraces a plurality of logics and languages each coming with their own co-hyperintensionality relation which hence are all on a par. The main thrust for this view is to acknowledge that both our scientific and everyday reasoning is about all kinds of domains and that we conceptualize these different domains by different logics and languages. The fact that most of the reasoning happens superficially in one language, namely a natural language like English, obscures the fact that the underlying logical form (or our interpretation of the natural language) varies from domain to domain. (Also see section 2.4.3 for related ideas.)

The third way is a mediating position. It acknowledges a plurality of logics and languages but accounts for the intuition of a universal co-hyperintensionality relation as follows. In each domain, context, or discourse we seem to always find the right calibration of meaning (or at least have strong intuitions about what the right one should be). So what is constant is not the co-hyperintensionality relation itself but our ability to always judge what the right one is.

The fourth way asks us more neutrally to investigate how the co-hyperintensionality relations of different languages are related in order to understand what effect a slight change in the language has on the co-hyperintensionality relation. This amounts to searching for stability or continuity theorems: Look at the space of all languages and show that the co-hyperintensionality relations of languages in the neighborhood of an arbitrarily fixed language only vary to a limited degree from the co-hyperintensionality relation of the fixed language.[51]

2.4.3 Reduction of co-hyperintensionality to validity and logical pluralism

We argue that our results point to a plurality of logics for hyperintensional operators.

In section 2.2.4, we already argued that our results reduce co-hyperintensionality to validity. Thus, future research has to investigate the notion of validity for statements involving hyperintensional operators. It is far from clear that there is one single such notion of validity. For example, Stenning and van Lambalgen (2008) provide

[51] Here is a cheap such theorem. Let $\mathcal{S}$ be the set of all languages (up to isomorphism). Order $\mathcal{S}$ by $(S_1, \Omega_1, \Leftrightarrow_1) \leqslant (S_2, \Omega_2, \Leftrightarrow_2)$ iff S_2 extends S_1, i.e., there are injections $i : S_1 \to S_2$ and $j : \Omega_1 \to \Omega_2$ such that for all $\nabla \in \Omega_1$, the arity of ∇ and $j(\nabla)$ are the same, and for all $\bar{\varphi} \in S_1$

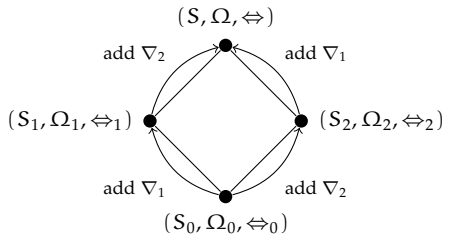
$$i\big(\nabla(\bar{\varphi})\big) = j(\nabla)\big(i(\bar{\varphi})\big),$$

and for all $\varphi, \psi \in S_1$, if $\varphi \Leftrightarrow_1 \psi$, then $i(\varphi) \Leftrightarrow_2 i(\psi)$. (This last condition might be debatable). In order to speak of "the neighborhood" of a language, define the tree topology $\tau_{\mathcal{S}}$ on the partial order $(\mathcal{S}, \leqslant)$, i.e. the open sets are $\leqslant$-upsets (sets that are closed under $\leqslant$). (For this topology see, e.g., Levy 2002, p. 201). Let $\mathcal{R} := \{\sim_S \mid S \in \mathcal{S}\}$ and order it by $\sim_{S_1} \leqslant \sim_{S_2}$ iff $\sim_{S_2}$ is at least as fine-grained as $\sim_{S_1}$, i.e., $S_1 \leqslant S_2$ (via injections i, j) and for all $\varphi, \psi \in S_1$, if $\varphi \not\sim_{S_1} \psi$, then $i(\varphi) \not\sim_{S_2} i(\psi)$. Again, to speak of the neighborhood of a co-hyperintensionality relation, define the tree topology $\tau_{\mathcal{R}}$ on the partial order $(\mathcal{R}, \leqslant)$.

Now, it is easy to observe that $S_1 \leqslant S_2$ iff $\sim_{S_1} \leqslant \sim_{S_2}$. Defining the function $f : \mathcal{S} \to \mathcal{R}$, $S \mapsto \sim_S$, we thus easily get stability in two directions: (1). If S is a language, and U a neighborhood of S (i.e., $S \in U$ and U is open), then $f(S)$ is a neighborhood of $\sim_S$. (2). If $\sim_S$ is a co-hyperintensionality relation, and V a neighborhood of $\sim_S$, then $f^{-1}(V)$ is a neighborhood of S (so particularly, f is continuous).

empirical (and conceptual) evidence that the right notion of validity depends very much on how we conceptualize the domain we talk about and how we interpret the discourse in which inferences take place. But there is also theoretical evidence for a plurality of validity, for example in three-valued logics (cf. e.g. Chemla et al. 2016). More generally, the thesis of logical pluralism – as famously defended by Beall and Restall (2005) – holds that there is more than one correct logic and hence more than one correct notion of validity. And if there is a plurality of notions of validity, there also is a plurality of notions of co-hyperintensionality.

But also our observation that the structure of (hyperintensional) operators is not linear points in the direction of a plurality of validity relations, because two incomparable operators can, in general, yield two incomparable logics which hence are on a par. This is because in general the following situation can occur. We have a language $(S, \Omega, \Leftrightarrow)$ with two incomparable operators ∇_1 and ∇_2. We obtained S as an extension from a base language S_0 via either the left or the right route:



Moreover, $\Leftrightarrow_1$ and $\Leftrightarrow_2$ are incomparable (i.e., neither $\Leftrightarrow_1 \subseteq \Leftrightarrow_2$, nor $\Leftrightarrow_2 \subseteq \Leftrightarrow_1$). That is, S_1 and S_2 are incomparable logics—they are not just different, but neither of them can be extended to the other.

2.4.4 Further ideas and open questions

We end this section by listing some further open questions and ideas that haven't been mentioned in the preceding sections (including possible developments of our formalism)

1. In section 1.2.3, we summarized approaches providing hyperintensional content. Evaluate these hyperintensional semantics by whether the relation of hyperintensional equivalence that they provide coincides with the co-hyperintensionality relation of the language.
2. Very broadly: How do our results affect answers to the other problems of hyperintensionality (cf. section 1.3.2)?
3. Precisely define language extensions (as in section 2.4.3 or footnote 51) and investigate their structure (cf. e.g. field extensions in algebra or Hodges (2001)). Characterize more precisely how incomparable operators correspond to incomparable logics.
4. What more can be said if we look at a particular language and take its further structure into account (e.g. first-order languages or contexts of utterances)?
5. Precisely formulate and prove stability theorems (as mentioned in the last paragraph of section 2.4.2).

6. Further explore the framework: Since a language is a universal algebra we may ask what are uncontroversial algebraic properties of the language that ensure nice properties of the co-hyperintensionality relation. For example, what algebraic properties ensure that in the definition of $\approx_\nabla$ (definition 2.2.4) it is enough to quantify over atomic sentences?[52] Or, what algebraic properties ensure that our language is dually equivalent to a topological space?

7. Use the results to define a (rough) "closeness" relation on the set of sentences: The more similar the meaning of two sentences is, the closer they are. For example, given two sentences φ and ψ, look for the $\leqslant$-minimal operator ∇ such that $\varphi \approx_\nabla \psi$ (circumvent issues of uniqueness). Then the height of ∇ in $(\Omega, \leqslant)$ measured from the identity operator is the distance between φ and ψ.

8. Having a quoting operator in the language trivializes the co-hyperintensionality relation of the language to the identity relation. Thus, we may ask: Is there a way to determine what the relevant operators of a language are? If there is no unique way to do so, does this point to pluralism as well?

9. Intuitively, hyperintensional operators can be divided into representational and non-representational ones. The representational ones include, for example, belief or conceiving, and they have a mental and cognitive aspect. The non-representational ones include, for example, grounding or truthmaking, and they have a metaphysical and extensional aspect. What of the above encountered negative results can be avoided if we restrict us to one of these two subclasses? For example, the impossibility interpretation of our results applies to the representational operators, but the substitution salva veritate criterion most plausibly applies to extensional operators—so can we get a possibility result via our finding for extensional operators? Or can an interesting subclass of representational operators be linearly ordered?[53] What about the claim that we have a plurality of logics for representational operators, and logical monism for non-representational operators?

10. We made the standard assumption that validity is transitive, which we essentially used only in the proof of lemma 2.2.5 (ii). In section 2.4.1, however, we saw that this assumption entails that equipollence and substitution salva veritate are jointly inconsistent criteria for sameness of content which, in turn, yielded an impossibility interpretation of our results. This fostered the question of what happens if the transitivity of validity is dropped. In particular, we may ask: For what interesting non-transitive validity relations can the substitution salva veritate and equipollence criteria be reconciled, and which of our results do still

[52] Note that in general this is not possible: Consider the operator $\nabla(\varphi, \psi) :=$ "φ or ψ is an atomic sentence". If we quantify only over atomic sentences, $\approx_\nabla$ will be the trivial relation, which is not the case if we quantify over all sentences.

[53] Example 2.3.3 provides a representational operator (knowledge) that is incomparable with a non-representational operator (because). Example 2.3.4 provides two non-representational operators (necessity and inexact truthmaking) which are incomparable.

hold then?[54, 55]

2.5 Conclusion

Let's end with a brief summary of what we've seen.

We saw that the most general notion of a language that still allows the problem of co-hyperintensionality to be formulated is that of a triple $(S, \Omega, \Leftrightarrow)$ where S is a set of sentences, Ω is a set of operators on S, and $\Leftrightarrow$ is an equivalence relation on S. Thus, we could define a co-hyperintensionality relation of a language $(S, \Omega, \Leftrightarrow)$ as a reflexive binary relation $\approx$ on S that is $\subseteq$-maximal with the property that for any operator $\nabla \in \Omega$, $\bar{\varphi} \approx \bar{\psi}$ implies $\nabla(\bar{\varphi}) \Leftrightarrow \nabla(\bar{\varphi})$. Then we've seen that for every language $(S, \Omega, \Leftrightarrow)$ there is exactly one co-hyperintensionality relation which can be defined in terms of $\Leftrightarrow$. So we reduced co-hyperintensionality to validity. Thus, our simple, yet well-motivated, general and neutral framework of languages as triples $(S, \Omega, \Leftrightarrow)$ is already powerful enough to define co-hyperintensionality relations and show their unique existence. But more can be done, as is seen in applying this framework to investigate the structure of hyperintensionality—which is an important task since most philosophical notions are considered to be hyperintensional.

The order of operators by granularity ($\nabla_1 \leqslant \nabla_2$ iff $\approx_{\nabla_1} \leqslant \approx_{\nabla_2}$) in general is a partial but not a linear order. There even are natural operators on the level of intensionality (necessity and inexact truthmaking) that are not comparable in terms of granularity. Thus, defining an operator ∇ to be hyperintensional iff $\nabla \not\leqslant \Box$ differs from defining it via $\nabla > \Box$. Moreover, despite the often tacitly neglected differences, we saw that being substitutable salve veritate and being co-hyperintensional coincides. Yet, if validity is transitive, equipollence and substitution salva veritate are jointly inconsistent criteria of sameness of content. This provided an impossibility interpretation of our results: Hyperintensionality (as commonly defined) cannot provide a cognitively adequate individuation of content. Since this was one of the reasons to go hyperintensional in the first place, this should make us rethink the notion of hyperintensionality. We mentioned the question whether there is only one natural co-hyperintensionality relation (namely that of the natural language we speak), or whether there is a plurality of languages and validity notions, and hence of co-hyperintensionality relations. The fact that there are incomparable operators points to a plurality of logics for hyperintensional operators.

This is were we stopped, but there is a lot more to be explored, for example: Stability theorems, the structure of language extensions, a plurality of hyperintensional logics, representational and non-representational operators, and re-defining co-hyperintensionality for representational operators via non-transitive validity relations.

However, since co-hyperintensionality fails to deliver the promised cognitively adequate individuation of content, it arguably is the most pressing problem to explore how else to capture cognitive synonymy. This is what we will do in the next chapter.

[54] Note that there is room to pursue this idea while avoiding the radical step of dropping the transitivity of validity (i.e. logical consequence): Instead of reading "If $\nabla(\bar{\varphi})$ (and some possible additional conditions), then $\nabla(\bar{\psi})$" as inference (and considering its validity), we can understand it as a conditional (and consider its correctness)—and thus we can consider appropriate non-transitive conditionals. We take up this idea in the next chapter.

[55] Note an interesting analogous situation: Moving from Lewisian subject matter to Yabloian subject matter also essentially involves dropping transitivity (cf. Yablo 2014a, p. 5, 2014b, pp. 1-3).

3rd *Chapter*

COGNITIVE SYNONYMY

Since co-hyperintensionality cannot provide the promised cognitively adequate individuation of content, we describe in this chapter when two sentences are regarded as cognitively synonymous.

We describe cognitive synonymy as likeness in cognitive role: Usually, what we (should) interpret, presuppose, understand, and – inductively, abductively, or deductively – conclude from one sentence is the same for the other. We can recognize that two sentences φ and ψ have the same cognitive role in this sense, if, roughly, we have the defeasible rules "If φ, then ψ" and "If ψ, then φ" in our knowledge base.

This is a "conceptual" analysis of our cognitive ability to recognize two sentences as playing the same cognitive role. We then provide this analysis on the other two Marrian levels: on the algorithmic level by using logic programming (precisely capturing the notion of a defeasible rule), and on the neural level by providing a neural network that explains why cognitive synonymy has both rule-like features but still is (contextually) flexible.

The framework thus developed will be the foundation for the notion of a scenario that we introduce in the next chapter.

In the previous chapter, we saw an impossibility result: If we follow the standard definition of co-hyperintensionality, then there is exactly one such co-hyperintensionality relation. However, this one relation is transitive and hence not a cognitively adequate individuation of content since it doesn't take our bounded rationality into account. This is particularly bad for "cognitive" operators (like belief, conceiving, etc.) where the hope of going hyperintensional precisely was to find a cognitively more adequate individuation of content.

The good thing about impossibility results is that they spur positive ones: In this chapter, we look at our *cognitive ability* to recognize two sentences as having the same meaning (in a given context)—if we do so, the two sentence have the same cognitive role for us. The idea is to thus find when two sentences are cognitively synonymous, that is, how a cognitively adequate individuation of content should look like. (And indeed, individuating sentences by their cognitive role will give us a hyperintensional relation that retains much of the substitutability properties that define the notion of co-hyperintensionality, but that still is cognitively adequate.)

The following three sections correspond to Marr's three levels on which one should analyze a cognitive ability. Section 3.1 describes on a conceptual (or informal) level what we do when we recognize two sentences as having the same cognitive role.

Section 3.2 describes this on an algorithmic level. And section 3.3 describes how this in principle can be implemented in the brain: That is, the section provides a neural network which determines whether or not two sentences have the same cognitive role.

3.1 Cognitive synonymy: Conceptually

In this section, we describe the concept of cognitive synonymy as likeness in cognitive role, we see what kind of content it is that is individuated in this way, and we analyze on a conceptual level how we can recognize that two sentences have the same cognitive role.

3.1.1 A first approximation of cognitive synonymy

What does it mean that two sentences are cognitively synonymous? That we *can* recognize that the two sentences play the same cognitive role. But what exactly is our *cognitive ability* to recognize two sentences have, cognitively speaking, the same role? As a first intuitive approximation, it is our ability to do the following: In a given context, we can take into account (i) our knowledge, (ii) general knowledge, and (iii) facts and rules that we perceive as obtaining in the context, in order to judge (deliberately or automatically) that two sentences play the same role which should imply things like:

- If someone takes one of the two sentences to be true, she will also take the other to be true.
- The inductive, abductive, and deductive conclusions drawn from the one sentence (in the context), will also be drawn from the other.
- The change to the intended discourse model (in the context) made by one sentence is the same made by the other.
- Usually, what will be understood by uttering one sentence, will be also be understood by uttering the other.

As an example, consider the two sentences "Bob has a dark-brown beard" and "Bob's facial hair is of dark color". In most contexts of utterance (which, e.g., don't draw attention to the very exact color of Bob's beard or whether there is some difference between "beard" and "facial hair"), these two sentences will have the same cognitive role: What I interpret, presuppose, understand, or conclude when I'm told the first will be the same – across these contexts – as when I'm told the second. (We will soon make precise the qualifier "usually".)

To develop this first approximation in more detail we will do the following: In section 3.1.2, we clarify what kind of content it is that we individuate when we individuate by cognitive role. In section 3.1.3, we describe both our agent-based knowledge and general knowledge. In section 3.1.4, we spell out a criterion for when two sentences have the same cognitive role and check that this criterion implies the above characteristics.

3.1.2 What kind of content does cognitive synonymy individuate?

We present five remarks on the concept of cognitive synonymy and on what kind of content it individuates.

First, individuating sentences by cognitive synonymy – that is, by their cognitive role – provides an individuation of a notion of content that we may call *cognitive content*: the content of a sentence that figures in our cognition.

Second, we take the following terms to all describe more or less the same notion that we're after: Cognitive synonymy, likeness of cognitive role, likeness of cognitive content, likeness of representation, likeness of conceptualization. Note that all of these are relations between sentences that are hyperintensional (in the sense of section 1.2.2): There are co-intensional sentences that are not cognitively synonymous (like "apples are apples" and "bachelors are unmarried"). However, cognitive synonymy probably isn't strictly hyperintensional. The two sentences describing Bob's beard in the above example arguably usually have the same cognitive role but not the same intension (there is a world separating "dark-brown" from "dark color"). However, the reason for the qualifier "probably" is that the two sentences only have the same cognitive role in *most* but not all contexts. Usually – i.e. in most contexts –, we interpret these two sentences as saying the same thing, but there are contexts (that, e.g., draw attention to the very exact color of Bob's beard) where we interpret them differently.

Third, as a rough distinction, there are two ways to look at cognitive synonymy which we may call the *cognitive* and the *semantic* perspective, respectively.[56] From the cognitive perspective, we want to understand *why* two sentences are cognitively synonymous—that is, play the same cognitive role. From the semantic perspective, we want to systematically understand *what* sentences are cognitively synonymous. The cognitive perspective – which we take in this chapter – is bottom-up: We will go from the rules that agents come to learn as governing the world, to when two sentences play the same cognitive role for them ("from rules to synonymy"). Last chapter we took the semantic perspective: Co-hyperintensionality promised to give a notion of cognitive synonymy (by providing an individuation of the content of cognitive operators like belief). This perspective is top-down ("from synonymy to rules"): We started from the common assumption that there should be a notion of co-hyperintensionality and we investigated the literature's suggestion that it is governed by the "substitutability rule"—only to find that if co-hyperintensionality is conceived of this way, it cannot provide a cognitively adequate individuation of content.

Fourth, one might wonder whether this cognitive content that cognitive synonymy individuates is (the or a notion of) meaning. An answer depends – non-surprisingly – on one's concept of meaning. Let's assume for the moment the most splendid conception of meaning that Chalmers (2006a, p. 55) describes as the "golden triangle" conception of meaning constructed by Kant, Frege, and Carnap. The three vertices of the triangle are reason, modality, and meaning; and reason is connected to modality (a priori coincides with necessary), modality is connected to meaning (intension is an important part of meaning), and meaning is connected back to reason (sense is an

[56] These two perspectives roughly parallel the two ways we can look at meaning: The foundational perspective that focuses on *why* a sentence has the meaning it has (e.g. grounded in a linguistic community), and the semantic perspective that focuses systematically on *what* meaning a sentence has (taking our linguistic and conceptual knowledge for granted). For this distinction see e.g. Lewis (1970, p. 19) or Speaks (2017).

important aspect of meaning that reflects cognitive significance). On this conception, cognitive content – which is located at the reason vertex – would be meaning. However, as already mentioned in chapter 1, the golden triangle was severely damaged by Saul Kripke, Hilary Putnam, and others. The result was an externalist conception of meaning which – since then – seems to be the received conception of meaning. On this conception, cognitive content is not meaning because meaning "ain't in the *head*" (Putnam 1973, p. 704).

Fifth, when we discussed the Lewis-Stalnaker objection in section 1.3.3, we remarked that the objection (merely) shifts the problem to explaining when we take two sentences to refer to the same intension. But this basically is explaining likeness of cognitive role: To say that two sentences play the same cognitive role is to say that we usually would assign them to the same intension.[57] So the Lewis-Stalnaker objection prompts us to explain cognitive synonymy.

3.1.3 The agent-based and generic knowledge base

In order to describe cognitive synonymy or likeness of cognitive role, we need the concepts of agent-based knowledge and general knowledge. We describe these concepts here.

The agent-based knowledge base. When we grow up we learn, and we keep this knowledge in a knowledge base. For example, we learn (empirical) facts – like that fire is hot or what the name of our family members are – and we store these facts in our knowledge base. Moreover, we also learn general rules (that allow exceptions) and keep them in our knowledge base, too. A simple (notorious) instance of such a rule is: "If it is raining, the street is wet". Of course, there might be situations where this rule doesn't apply (e.g., if there is a roof above the street), but in general – that is, if we don't have any reason to believe that something weird is going – we rely on this rule. Such rules are often called defeasible (since they can be "defeated" by exceptions). Thus, we can think of the knowledge base of an agent as a collection of facts and defeasible rules that can be updated: via learning we can add new facts and rules.[58]

Note that our knowledge base is closely connected to our beliefs: based on our knowledge base we form our beliefs. Roughly, in a given context (of everyday life) we believe that φ, if the following holds: After consulting both the facts we observed about the context we're in and the relevant facts and rules in our knowledge base, we conclude – based on these rules – that φ should be the case.[59]

Also note that cognitive synonymy should be grounded in the rules in our knowledge base. For, we don't first learn that two sentences are synonymous and then conclude that by believing one we also believe the other. Rather, we notice that the

[57] Note that we can take two sentences to refer to the same intension in a given context even though we know that, strictly speaking, they differ in intension. This is because of the same reason why likeness in cognitive role is not strictly hyperintensional: In most contexts, we take the two sentences about Bob's beard to refer to the same intension (because such contexts don't draw attention to the exact color of Bob's beard), though we know that there are worlds that can distinguish the two sentences.

[58] The dynamics of a knowledge base are far more complex: We can also delete items and perform many more operations that are dealt with in the theory of belief revision. But here we, fortunately, don't need more details about the dynamics of knowledge bases.

[59] Of course, not everything we believe is true: Mistakes can occur, for example, if we don't realize that one of the rules we applied was defeated in the context we're in.

rules we gathered in our knowledge base tell us that whenever we come to believe one sentence we will also believe the other.

The generic knowledge base. We not only have to compute our own beliefs but also those of others. For many social interactions we need a so-called *theory of mind*: the ability to correctly ascribe mental states (like believing, intending, desiring, knowing, pretending) not only to ourselves but also – potentially different ones – to others. Now, to compute our beliefs it is enough to only consider our knowledge base and the facts and rules we observe about the context. But to compute the beliefs (and other mental states) of others we additionally have to consider both *their* knowledge bases and the facts and rules about the context that *they* observe. However, their knowledge base is not directly accessible to us: we don't know exactly what others take for granted. To still be able to ascribe beliefs to others, we need to assume that their knowledge base at least contains some "common knowledge" rules and facts that we share with them. To illustrate this, let's consider a simple example:

> In Alice's village, they recently moved the local museum but they forgot to change the signs leading to it. Alice sees a touristy old man coming along, looking for the museum sign, finding it, and then moving on. She is deliberating whether she should tell him that the museum has moved, so she is wondering: Does he believe that the museum moved? Of course, Alice concludes that he doesn't believe so, based on the following reasoning: The tourist also has in his knowledge base the commonly shared rule "If there is a sign to a museum, the museum will be there" (this is part of the "common knowledge" mentioned above). And he also saw the sign to the museum (this is a fact about the context that Alice saw the tourist observing). As Alice cannot see any reason why he would have inhibited that rule (the signs are not rusty, etc.), Alice reasons that the tourist must have concluded that the museum is where the signs lead to and not anywhere else. So Alice concludes that the tourist doesn't believe that the museum moved.[60]

So we need to assume that there are some facts and rules that are present in every knowledge base of agents to which we want to ascribe a theory of mind. We call these facts and rules the *generic knowledge base*, and the set of relevant agents A.

In fact, the need for a theory of mind is only one reason – an "agent-based" reason – for assuming a general knowledge base and a set of relevant agents A. There also is the following second, "modeler-level" reason. We want to find out when two sentences have the same cognitive role and what laws govern this notion. However, having the same cognitive role is relative to an agent, but of course the laws governing it should be general: We're not interested in how one particular agent individuates her cognitive content and take this to be the laws. Rather, we want to know what laws hold universally over all relevant possible agents.[61] Once this set A of relevant

[60] Once we've introduced negation as failure in the next section, we could better explain this way of handling negation (concluding that something is not the case if there is no information that it is the case). However, here this is not necessary because this is not the point of the example. The example should just illustrate how we compute the belief of others and that this requires a shared knowledge base.

[61] The situation is analogous to laws of nature: Not every fact about our world is a law of nature, rather we're interested in finding the laws that hold across all relevant possible worlds. That is, there, too, has

possible agents is given, the generic knowledge base can be seen as the common ground of all the knowledge bases of the agents in A.

In the appendix (section A), we further discuss the notion of a set of relevant agents and a generic knowledge base. But for our purposes, the idea of these concepts should be clear enough.

3.1.4 An informal criterion for cognitive synonymy

We describe our ability to recognize that two sentences play the same cognitive role on a conceptual level (Marr's computational level). We then give an informal criterion for cognitive synonymy.

Our ability to recognize two sentences as playing the same cognitive role is our ability to compute the following function: Given two sentences as input we compute – using our agent-based and/or general knowledge – the output YES or NO depending on whether or not the two sentences play the same cognitive role. If we want to have a context-sensitive individuation of cognitive role, we add a context to the input and also take into account in the computation the facts and rules provided by the context. Note that we can either deliberately compute this function by "consciously" reflecting on whether the two sentences do play the same cognitive role, or we can automatically compute the function so that we are, if at all, only aware of the outcome.

But how should the computation proceed? A detailed algorithmic description will be the task for the next level (section 3.2), but the underlying idea is the following.

Likeness of cognitive role, conceptually Two sentences φ and ψ have the same cognitive role (in context c) iff there are the defeasible rules "If φ, then ψ" and "If ψ, then φ" in the generic knowledge base (or in the set of rules provided by the context).

Four remarks. First, of course, this characterization of individuation of cognitive role can also be done agent-relative ("φ and ψ have the same cognitive role for agent α"). To do so, we simply replace "the general knowledge base" by "the knowledge base of the agent".

Second, we claim that identity of cognitive role is a hyperintensional relation that finds a good compromise between providing the (mutually exclusive) full-fledged substitution properties and cognitive adequacy.

On the one hand, the substitutability under cognitive operators (that being co-hyperintensional allows for) remains in the following sense—taking the belief operator as an example. Assume φ and ψ have the same cognitive role. Then – according to the above characterization – there are the rules "If φ, then ψ" and "If ψ, then φ" in the generic knowledge base. Now, if I (resp. an agent α of the set A) believes φ, then I (resp. α) derived that φ should be the case based on the facts and rules from my (resp. the generic) knowledge base and those obtained from the context. But since I (resp. α) can go via the rules in the knowledge base from φ to ψ and vice versa, I (resp. α) can also derive ψ. Hence I (resp. α) also believes ψ.

On the other hand, individuating sentences by their cognitive role is cognitively adequate: it is grounded in how actual agents come to know that two sentences are

to be a set of possible entities (or rather: representations of possibilities) to which the laws should apply: namely the set of physically possible worlds.

cognitively synonymous. In particular, individuation by cognitive role (as spelled out in the above characterization) is not transitive: Let's go back to the examples of a long list of sentences where adjacent ones are trivially equivalent or very close synonyms, but where the first and the last sentences are not (section 2.4.1). Plausibly, the generic knowledge base contains rules allowing to go from one sentence to its neighbor, because the examples where constructed such that this transition is possible. (And of course, this rule can still be inhibited, for example, if the task of making the transition has been formulated – or framed – in a way that impedes thinking of it). This renders adjacent sentences to have the same cognitive role. However, the first and last sentence do not have the same cognitive role: The examples have been built such that a transition between them does not occur.

Third, similarly, likeness in cognitive role (as spelled out via defeasible rules) can be seen to have the properties mentioned in the beginning: For example, we usually conclude the same things from two sentences that are linked by defeasible rules. For reasons of space, we don't explain this in detail. Though, for the use of defeasible rules in human reasoning and discourse understanding see van Lambalgen and Hamm (2005), Stenning and van Lambalgen (2008), or Besold et al. (2017). Moreover, this will also become clearer once we spelled out precisely defeasible rules in the next two sections.

Fourth, taking up on section 3.1.3, this criterion indeed grounds likeness of cognitive role in the generic knowledge base, which in turn is grounded in the individual knowledge bases. That is, this criterion grounds likeness of cognitive role in the common knowledge of the group of all relevant possible cognitive agents. This criterion thus also reflects the position of the concept of likeness of cognitive role in the cognitive/conceptual priority order: It's not that we learn that two sentences have the same cognitive role and then deduce from believing one that we also believe the other.[62] Rather we learn as a rule that we can go from (believing) one to (believing) the other and thus conclude that they must have the same cognitive role. In particular, likeness of cognitive role is *not* a primitive notion.

3.2 Cognitive synonymy: Algorithmically

In this section, we describe our cognitive ability to recognize that two sentences have the same cognitive role on an algorithmic level. To do so, we first introduce logic programming in subsection 3.2.1: a formal logic that is used in artificial intelligence and cognitive science, and that is – unlike classical logic – cognitively plausible (it is, for instance, computationally tractable). Just how suitable and fruitful logic programming is to model cognitive and linguistic abilities and phenomena can be seen, for example, in van Lambalgen and Hamm (2005), Stenning and van Lambalgen (2008), or Besold et al. (2017). In subsection 3.2.2, we will use this framework to describe the ideas of the previous section on an algorithmic, formal-symbolic level.

[62] More care has to be put in distinguishing between the conceptual priority order and the cognitive priority order—and, respectively, between the claim that likeness in cognitive role is *conceptually* grounded in the knowledge bases of the relevant possible agents, and the claim that it is *cognitively* grounded therein. The claims I make here concern the cognitive ordering, but given that cognitive role is a cognitive concept, reading the claims as being about the conceptual ordering also might have some plausibility. For reasons of space, we won't further discuss this here.

3.2.1 Logic programming

We introduce the basics of logic programming that we will need later on. It will largely be a summary of existing work.[63] We will mainly follow Stenning and van Lambalgen (2008, sec. 7.2.1–2) with some fixes.[64] Other helpful references are e.g. Doets (1994) and Kencana Ramli (2009).

We start by defining what a general logic program is: a collection of rules (that are called clauses) over a propositional language.

Definition 3.2.1 (General logic program)**.** Fix a countably infinite set S of propositional variables, and fix the propositional constants $\bot$ and $\top$. A *general clause* is of the form

$$p_1 \wedge \ldots \wedge p_n \wedge {\sim}r_1 \wedge \ldots \wedge {\sim}r_m \to q$$

where $p_i, r_j \in S \cup \{\bot, \top\}$ and $n, m \in \{0, 1, 2, \ldots\}$. (We'll soon get to the exact meaning of the negation symbol $\sim$.) A *fact* is a clause of the form $\top \to q$ (often abbreviated to q). A *general logic program*[65] P is a finite set of general clauses. An *atom* of P is an element of S that occurs in a clause in P.

The negation $\sim$ is what really makes logic programming different from classical logic.[66] In classical logic, if we want to conclude $\neg\varphi$, we have to derive this with the rules for negation; thus, if we succeed, we "know" that it is not the case that φ. In logic programming, to conclude ${\sim}\varphi$, it is enough to not "know" that φ is the case. This kind of negation is called *negation as failure* and the assumption behind it is called the *closed-world assumption*: what is not known to be true is false. We apply this assumption to many sentences in everyday reasoning: For example, if we don't see a train scheduled at 2 pm, we conclude that there won't be a train.

Formally, negation as failure can either be described syntactically[67] or semantically—here we only need to consider the latter. Semantically, negation as failure can be described by spelling out a relation $P \mid\!\approx \varphi$ which is intended to say that according to the general logic program P, φ is the case. Then we can consider the sentence $\varphi :=\!{\sim}\psi$ to get a description of when a negated sentence is true according to a program P.

To spell out $P \mid\!\approx \varphi$, we first have to fix a language of which φ can be a sentence. We call this language the *query language* since it is the language in which we ask whether certain sentences are the case according to a given program.[68]

[63] But it arguably is too diversely discussed to be omitted or moved into an appendix.

[64] We add as a parameter a set of sentences to which the so-called negation as failure is applicable. This will fix problems with the definition of completion and its correspondence with the consequence operator. However, we also note that there still are problems with completion which we can avoid – for reasons of space – by just working with the consequence operator.

[65] The term "general" is to separate these programs from positive programs whose clauses aren't allowed to contain the negation symbol. Here, however, we will always work with general programs and never restrict us to positive ones.

[66] In addition to the more restricted language of logic programming and the intuitive reading of a clause *not* as a sentence that can have a truth-value, but as a "contentful" rule of inference.

[67] Describing ${\sim}\varphi$ as: the attempt to derive φ from a given program in the so-called resolution calculus fails in finitely many steps.

[68] It contains more operators (and thus is more expressive) than it is usually assumed to be in standard logic programming contexts. Though, this greater expressive power comes at no cost (since the additional operators can be given a straightforward semantic), and it will be needed later on. In the next chapter, we discuss how we can add even more operators to the query language so it also contains belief operators and others.

p	q	$\sim p$	$p \wedge q$	$p \vee q$	$p \to q$	$p \leftrightarrow q$
1	1	0	1	1	1	1
0	0	1	0	0	1	1
u	u	u	u	u	u	1
1	0		0	1	0	0
1	u		u	1	u	0
0	1		0	1	1	0
0	u		0	u	1	0
u	1		u	1	1	0
u	0		0	u	u	0

Figure 3.1: Truth-tables for strong Kleene logic with a classical $\leftrightarrow$ added.

Definition 3.2.2 (Query language)**.** The *query language* $\mathcal{L}_Q$ is the language obtained by the following syntax

$$\mathcal{L}_Q ::= p \in S \mid \bot \mid \top \mid \ \sim p \mid p \wedge q \mid p \vee q \mid p \to q \mid p \leftrightarrow q.^{69}$$

The other ingredient to spell out $P \mid\!\approx \varphi$ are so-called three-valued models. They are three-valued because they don't only assign the truth-values true (1) and false (0) to sentences but also undetermined (u). The value u is intuitively interpreted as: at the current stage it is not known yet whether this sentence will turn out to be true or false.

Definition 3.2.3 (Three-valued model)**.** A three-valued model M is a function $S \to \{0, 1, u\}$. This function can be extended to all $\mathcal{L}$ sentences by: $M(\bot) = 0$, $M(\top) = 1$ and the truth-tables[70] in figure 3.1 (note that thus $\varphi \to \psi$ is equivalent to $\sim\varphi \vee \psi$).

For an $\mathcal{L}_Q$-sentence φ, $M \models_3 \varphi$ is defined as $M(\varphi) = 1$. Given two $\mathcal{L}_Q$-sentences φ and ψ, we define $\varphi \models_3 \psi$ as: For all three-valued models M, if $M \models_3 \varphi$, then $M \models_3 \psi$. (We often skip the "3" in the subscript.)

Now we can define $P \mid\!\approx \varphi$. There are two ways to do so. Both capture the idea that to determine whether $P \mid\!\approx \varphi$ we consider "the intended model" of P and check whether in this model φ is true. The two ways correspond to two ways of spelling out "the intended model". The first way proceeds by what is called the *completion* of P, which captures the closed-world assumption by telling us how to make exactly those sentences true for which we have information. The second way provides – via the so-called *consequence operator* – a procedure that starts with a minimal model in which every sentence is undetermined and updates this model step by step making exactly those sentences true (or false) that need to be made true (or false) according to the program. One can show that both approaches amount to the same notion $\mid\!\approx$, yet both ways capture important aspects of the intended model of a program. Here we only need the second way, so we only sketch the first and then spell out the second in more detail.

[69] Actually we will only need the fragment of the language where $\sim, \to, \leftrightarrow$ are never nested, and $\wedge, \vee$ are only nested by forming finite disjunctions of finite conjunctions. But for our purposes here we can also work with the full language.

[70] The truth-tables for $\sim, \wedge, \vee, \to$ are the ones of strong Kleene three-valued logic. It was first formulated by Kleene (1952, § 64) – though other kinds of three-valued logics existed before that –, and a short summary is found, e.g., in Priest (2008, sec. 7.3). The connective $\leftrightarrow$ has a classical interpretation: being true iff both sentences have the same truth-value.

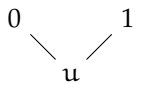
Completion. Fix a general logic program P and a finite set of sentences $N \subseteq S$ (this will be those to which we want negation as failure to apply). For $q \in N$, construct the $\mathcal{L}_Q$-sentence $def_{(P,N)}(q)$ by: Take all the (finitely many) clauses $\varphi_i \to q \in P$, and set $def_{(P,N)}(q) := \left(\bigvee_i \varphi_i\right) \leftrightarrow q$ (if there is no such clause set $def_{(P,N)}(q) := \bot \leftrightarrow q$). Define the *completion of* P *under* N as the $\mathcal{L}_Q$-sentence

$$comp(P,N) := \bigwedge_{q \in N} def_{(P,N)}(q) \wedge \bigwedge \{\varphi \to q \in P \mid q \notin N\}.$$

Note how $comp(P,N)$ captures the closed-world assumptions for sentences $q \in N$: A three-valued model of $comp(P,N)$ makes true q if *and only if* the antecedent of one of its clauses is true. And if $q \notin N$, then q will be made true if the antecedent of one of its clauses is true, but if all of them are false we cannot conclude yet that q has to be false, too.[71] Thus, the intended model of P (under N) should be a three-valued model of $comp(P,N)$. As an example, consider the program $P = \{p \to q, p \to r\}$ and $N = \{p, q\}$. Then $comp(P,N) = (p \leftrightarrow q) \wedge (\bot \leftrightarrow p) \wedge (p \to r)$, and the intended model makes p and q false but leaves r undecided. The idea of completion is on the right track in obtaining the intended model, but it still has its problems.[72] However, for reasons of space and because we don't need it in the remainder we don't discuss how to avoid these problems.

Consequence operator. To define the consequence operator, we first have to order the class of all models.

Definition 3.2.4 $((\mathcal{M}, \leqslant))$**.** Let $\mathcal{M}$ be the class of all three-valued models. To define a partial order $\leqslant$ on $\mathcal{M}$, we order the three truth-values $1, 0, u$ by



and define for $M, M' \in \mathcal{M}$

$$M \leqslant M' \text{ iff}_{df} \ \forall p \in S : M(p) \leqslant M'(p).$$

Definition 3.2.5 (Consequence operator $T_{(P,N)}$)**.** Given a general program P and $N \subseteq S$ finite, set $P_N := P \cup \{\bot \to q \mid q \in N\}$. Define the consequence operator

[71] There is not only the closed-world assumption for sentences but also for rules in the program. This becomes relevant if we also allow programs to be updated (i.e. adding and deleting rules and facts), and is then captured by so-called integrity constraints. However, for our purposes here we don't need them.

[72] We could define $(P,N) \approx \varphi$ as $comp(P,N) \models_3 \varphi$, and this would give us for "non-pathological" programs a good notion $\approx$ that also is equivalent to the one we define below with the consequence operator. However, for pathological programs this wouldn't give the right results. Here are three examples.
(1). For $P = \{\sim p \to p\}$ we have $comp(P) = \sim p \leftrightarrow p$, so P doesn't have a model. However, the better choice seems to say that the intended model leaves p undetermined—which is indeed what we'll get via the consequence operator.
(2). Let's require abnormability clauses for every clause (as we will demand in the next section). Then, for $P = \{p \wedge \sim ab \to q, q \to ab, \top \to p\}$ we have that $comp(P)$ is equivalent to $\sim ab \leftrightarrow ab$, so P again doesn't have a model. But via the consequence operator we will get $\langle p, q, ab\rangle = \langle 1, u, u\rangle$ as intended model which again seems to be the better choice.
(3). For $P = \{p \wedge \sim ab \to q, \bot \to ab, \top \to q\}$ and $N = \{q, ab\}$ we have that the $comp(P)$ is equivalent to $p \leftrightarrow q$, from which not much can be said. However, via the consequence operator we will get $\langle p, q, ab\rangle = \langle u, 1, 0\rangle$ which is inconsistent with $comp(P)$, but seems to be the right model.

$T_{(P,N)} : \mathcal{M} \to \mathcal{M}$ as follows. Given a model M define the model $T_{(P,N)}(M)$ by

$$T_{(P,N)}(M)(q) := \begin{cases} 1 & \text{if there is a clause } \varphi \to q \text{ in } P_N \text{ s.t. } M \models_3 \varphi \\ 0 & \text{if there is a clause } \varphi \to q \text{ in } P_N \text{ and for all such clauses } M \models_3 \neg\varphi \\ u & \text{otherwise.} \end{cases}$$

If we omit N, we assume N to be the set of atoms of P.

One can show that $T_{(P,N)}$ is monotone: If $M \leqslant M'$, then $T_{(P,N)}(M) \leqslant T_{(P,N)}(M')$. So by the Knaster-Tarski fixed point theorem, $T_{(P,N)}$ has a least fixed-point obtained as follows. Start with the base model M_0 where all $p \in S$ have value u, iterate $T_{(P,N)}$ $n+1$ times (if P_N contains n clauses)[73], and obtain the model M_{n+1} which is the $\leqslant$-least fixed-point of $T_{(P,N)}$. This model is called the *minimal model* of (P, N). It is the intended model of the program: By construction, it makes those sentences true that have to be true but nothing more.

As promised, we can now define what $P \mid\!\approx \varphi$ means.

Definition 3.2.6 ($P \mid\!\approx \varphi$)**.** Let P be a general logic program, let $N \subseteq S$ be finite, and let φ be an $\mathcal{L}_Q$-sentence. Define

$$(P, N) \mid\!\approx \varphi \;\text{ iff}_{df}\; M \models \varphi,$$

where M is the minimal model of (P, N).

For "well-behaved programs" (cf. footnote 72), models of the completion of the program are fixed-points of the consequence operator and vice versa, and $(P, N) \mid\!\approx \varphi$ iff $comp(P, N) \models_3 \varphi$. But, again, for reasons of space and because of the problems of completion, we will only work with the consequence operator.

3.2.2 A logico-algorithmic individuation of cognitive role

Let's use this logic programming framework to formally model the conceptual description of individuation by cognitive role from the previous section.

The general knowledge base KB and the knowledge base KB_a of an agent $a \in A$ simply are logic programs where clauses always come with their own abnormality clause.

Definition 3.2.7 (Knowledge base)**.** A *knowledge base* (no matter whether agent-based or generic) is general logic program where every clause is of the form

$$p_1 \wedge \ldots \wedge p_n \wedge \sim r_1 \wedge \ldots \wedge \sim r_m \wedge \sim ab \to q,$$

where ab is a propositional constant, and every clause has its own unique abnormality clause ab. Also, instead of q the abnormality clause ab' of another clause is allowed. Moreover, negation as failure is always applicable to all these abnormality clauses.

Next, we need to say what a context is. Roughly, it consists of the rules and facts that obtain in it, and the sentences it could potentially falsify.

[73] That is, set $M_1 := T_{(P,N)}(M_0)$, $M_2 := T_{(P,N)}(M_1)$, …, $M_{n+1} := T_{(P,N)}(M_n)$.

Definition 3.2.8 (Context)**.** A context c is a triple (P_c, F_c, N_c) where P_c is a finite set of clauses (all, too, with their own abnormability clauses), F_c is a finite set of propositional variables, and N_c contains all the abnormality clauses occurring in P_c and possibly further (finitely many) propositional variables.

The intended interpretation is that the clauses in P_c are the rules that apply in the context (in addition to the rules in KB).[74] The sentences in F_c are the facts that hold in the context, and the sentences in N_c are the sentences to which negation as failure is applicable in the context. Thus, it will be useful to define

$$c^{KB} := KB \cup P_c \cup \{\top \to q \mid q \in F_c\} \cup \{\bot \to q \mid q \in N_c\}.$$

To have a notion of a context that takes an agent a into account, we can define c_a just as c but replacing KB with KB_a.

Note that we explicitly left open whether a context (in the above sense) represents the rules, facts, and potentially falsifiable sentences that are *perceived* by an agent as obtaining or that *actually* obtain. Usually, we take the former interpretation, but the latter can be used, too, if the application demands it. In the next chapter, we will generalize the concepts of a knowledge base and a context to the notion of a scenario.

Now we can formulate the individuation of cognitive role logico-algorithmically.

Likeness of cognitive role, algorithmically Two atomic sentences $p, q \in S$ have the same cognitive role (in context c) iff the generic knowledge base KB (resp. the context-augmented generic knowledge base c^{KB}) contains the clauses $p \wedge \sim ab \to q$ and $q \wedge \sim ab' \to p$. The version relative to an agent $a \in A$ is obtained by replacing KB by KB_a (resp. c^{KB} by c_a^{KB}).

Three comments. First, in the next chapter, we generalize identity of cognitive role to complex sentences.

Second, the "algorithm" to check whether p and q have the same cognitive role (which consists of checking whether the respective rules are in the knowledge base) is a very simple one. It still makes sense to speak of the "algorithmic level" since the framework of logic programming is very computational in spirit (it is, for example, straightforwardly implemented in the programming language PROLOG). To give a flavor of this, we can give – as an example – a computational notion of when an agent $a \in A$ believes φ (a sentence in the query language) in a context c_a: namely, precisely if $c_a \mid\!\approx \varphi$. In the next chapter, we discuss this notion of belief (e.g. why it avoids logical omniscience).

Third, one might be tempted to choose this criterion: p and q have the same cognitive role iff $KB \mid\!\approx p \leftrightarrow q$.[75] However, if p and q are facts in KB, they would always satisfy this condition, but intuitively they don't always play the same cognitive role. So one should at least revise the condition by "$KB \setminus \{\top \to p, \top \to q\} \mid\!\approx p \leftrightarrow q$", or in words: regardless of whether or not p or q are known to be the case, the generic

[74] Also, P_c may add clauses triggering abnormalities in KB and thus effectively prevent some of the rules in KB to be applicable in the context.

[75] Note how this resembles Fregean equipollence (cf. section 2.4.1). Though, there also is a difference since KB can also contain contingent facts while Fregean equipollence only talks about what can "rationally be conceived".

agent can conclude that the two sentences must have the same truth-value.[76] This, then, arguably doesn't describe cognitive role anymore, but rather "synonymy that is discovered (or cognized) via reasoning":

Reasoned-to synonymy, algorithmically Two atomic sentences $p, q \in S$ are reasoned-to synonymous (in context c) iff

$$KB \setminus \{\top \to p, \top \to q\} \approx p \to q \wedge q \to p$$

(resp. $c^{KB} \setminus \{\top \to p, \top \to q\} \approx p \to q \wedge q \to p$).

So, p and q are reasoned-to synonymous if it can be deduced that they must have the same classical truth-value. Note that this yields a more coarse-grained individuation of sentences than individuation by cognitive role: If two sentences have the same cognitive role, the agent can conclude that they must have the same truth-value (regardless of whether or not they are facts). The other direction doesn't necessarily hold. However, if an agent reasons – and thus *learns* – that p and q must have the same truth-value by a long reasoning-chain from p to q (and vice versa), she might add to her knowledge base the direct connection from p to q (and vice versa). So, via learning, identity of cognitive role and reasoned-to synonymy become very similar (and identical in the limit of this learning process). We will come back to this relation (in section 5.1), but for now we concentrate on the finer one of identity of cognitive role.

3.3 Cognitive synonymy: Neurally

The last section used the algorithmic framework of logic programming to describe our cognitive ability to recognize that two sentences have the same cognitive role. In this section, we consider how this can be implemented in a neural network.[77] In section 3.3.1, we show how to provide a neural network for a given logic program such that the network "computes" the intended model of the program. In section 3.3.2, we comment on some further features of this implementation like its biological plausibility or its learning abilities. In section 3.3.3, we then characterize the individuation of cognitive role from the neural perspective.

3.3.1 Neural implementation of logic programming

We provide a neural implementation of the logic programming framework that is an extension of Stenning and van Lambalgen (2008, ch. 8). Since the 1990s there has been a lot of research on how to combine symbolic approaches (like logic programming)

[76] Note that this is a transitive relation, but it is immune to counterexamples: If $KB \approx p \leftrightarrow q$, then *usually* p and q have the same truth-value, but this may be inhibited in some contexts c. This may happen, for example, if KB contains $p \wedge \sim ab \to q$, but c adds the clause $r \to ab$ and the fact r. So this relation is not monotone: If more contextual information is provided, then previously same-valued sentences may be assigned different truth-values. Note that this seems indeed to be going when real agents deliberate whether or not two sentences have the same truth-value.

[77] We assume a basic familiarity with (artificial) neural networks (e.g. that they are weighted graphs, how they compute by spreading activation, and how the dynamics of their state space look like, etc.). Our definitions will be largely self-contained but won't elaborate much on the basics of neural networks. An older but good introduction to neural networks is Rojas (1996).

in artificial intelligence and computational cognitive science with neural network approaches.[78] The reason for basing our framework on Stenning and van Lambalgen (2008, sec. 8) is their focus on cognitive processing rather than machine intelligence, their aim for biological plausibility (including the use of inhibitory neurons), the good handling of negation as failure, and – as we shall see – the fruitfulness of the framework (but this is, by no means, the only option we could have chosen). So our exposition is as in Stenning and van Lambalgen (2008, sec. 8), except that we add (or change) the following.

- Instead of binary neurons[79], we allow neurons with continuous values in $[0, 1]$. If their weighted and summed input reaches a threshold θ, they fire, otherwise they are inactive. Their value can loosely be interpreted as the "strength" with which the neuron fires. If the neuron represents a proposition, then this strength can, in turn, be interpreted roughly as the evidence that the subject has for that proposition.[80]

- We show how we still can implement AND and NOT gates as well as abnormalities.

- We fix the implementation of sentences to which negation as failure is applicable.

- We can model that for each spread of activation from one neuron to another a small bit of the signal is lost. This shows how long reasoning chains eventually fade out, accounting for bounded rationality.

Now, in order to neurally implement the logic programming framework, we provide a procedure to translate any logic program (i.e., any knowledge base) into a recurrent neural network such that the minimal model of the logic program (i.e., the intended interpretation of the knowledge base) corresponds to the stable state of the neural network. To do so, we show:

(i) How to represent neurally the truth-values true (t), false (f), and undetermined (u); and, additionally, the inconsistent truth-value "both true and false" ($\perp$).

(ii) What the general shape of the final network is.

(iii) What the building blocks are out of which we build the network for a given program.

(iv) How to build the network for a given program, and that this indeed computes the minimal model of the program.

[78] A very incomplete list of references is the following. Early work includes Balkenius and Gärdenfors (1991) and Hölldobler and Kalinke (1994), a standard reference is D'Avila Garcez et al. (2009), and recent summary is Besold et al. (under review). A more general connection between non-monotonic logics, neural networks, and dynamical systems is due to Leitgeb (2001, 2003, 2005).

[79] We'll later more neutrally speak of a "unit" instead of a "neuron". The reason is that units (can also) stand for clusters of actual neurons (in the brain). They thus allow for a slightly more high-level modeling that is more explanatory for our purposes. Cf. footnote 82.

[80] Stenning and van Lambalgen (2008, p. 226) only mention that the framework can be extended in this direction, but don't show how. There are complications especially with the AND gates which we will tackle here.

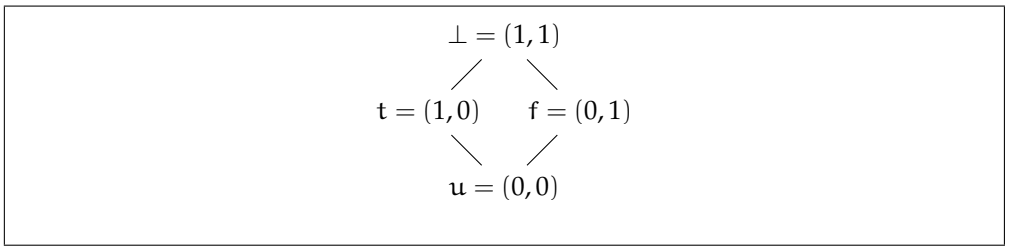


Figure 3.2: Two dimensional truth-values forming a lattice

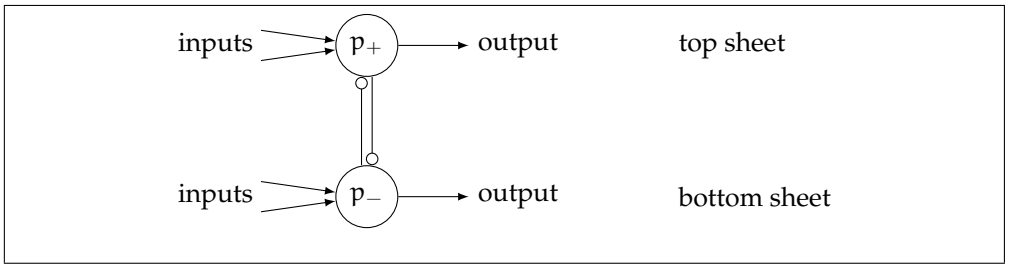


Figure 3.3: Implementation of a sentence p

Ad (i). We represent the three truth-values as $t = (1,0)$, $f = (0,1)$ and $u = (0,0)$. So a truth-value now has "two dimensions": the first specifying its positive (or truth) component, the second specifying its negative (or falsity) component.[81] As can be seen in figure 3.2, they keep the same ordering as before—now completed to a lattice by also adding as a supremum the inconsistent truth-value $\bot = (1,1)$.

Ad (ii). This idea of representing the positive component of a sentence separately from its negative component carries over to the general shape of neural networks that we are going to construct. They consist of two sheets: one on top of the other. Each sheet is a recurrent neural network, and for each unit in one sheet, there is a unique corresponding unit in the other layer. We call a pair of corresponding units a *node*. As will be seen, the top sheet computes the positive components of sentences and the bottom one computes the negative components. We call such networks *coupled recurrent neural network*. (We don't need a more precise formal definition at this point because we will construct for any given program P a concrete coupled recurrent neural network $\mathsf{CRN}(P)$, and this $\mathsf{CRN}(P)$ is formally defined below.) To already see an example of such a network, one can look ahead to figure 3.8. For definiteness, we fix the threshold θ of any unit to be 0.1, but of course this could be varied (also allowing different neurons to have different thresholds).

Ad (iii). As building blocks we need – to deal with clauses – subnetworks that implement: sentences, ANDs, NOTs, ABs, facts, and sentences to which negation as failure is applicable. And to deal with whole programs we also need ORs. We now show how they are constructed.

We implement a sentence p simply by a node (p_+, p_-) where the unit p_+ in the top sheet represents the positive component of the sentence and the bottom unit p_- represents the negative component. This is represented in figure 3.3:

[81] Note that this is very natural in a three-valued logic setting: Because there is a third value, truth and falsity become independent, that is, falsity is not the opposite of truth anymore. Thus, we represent independently both the truth-component of a sentence and its falsity-component.

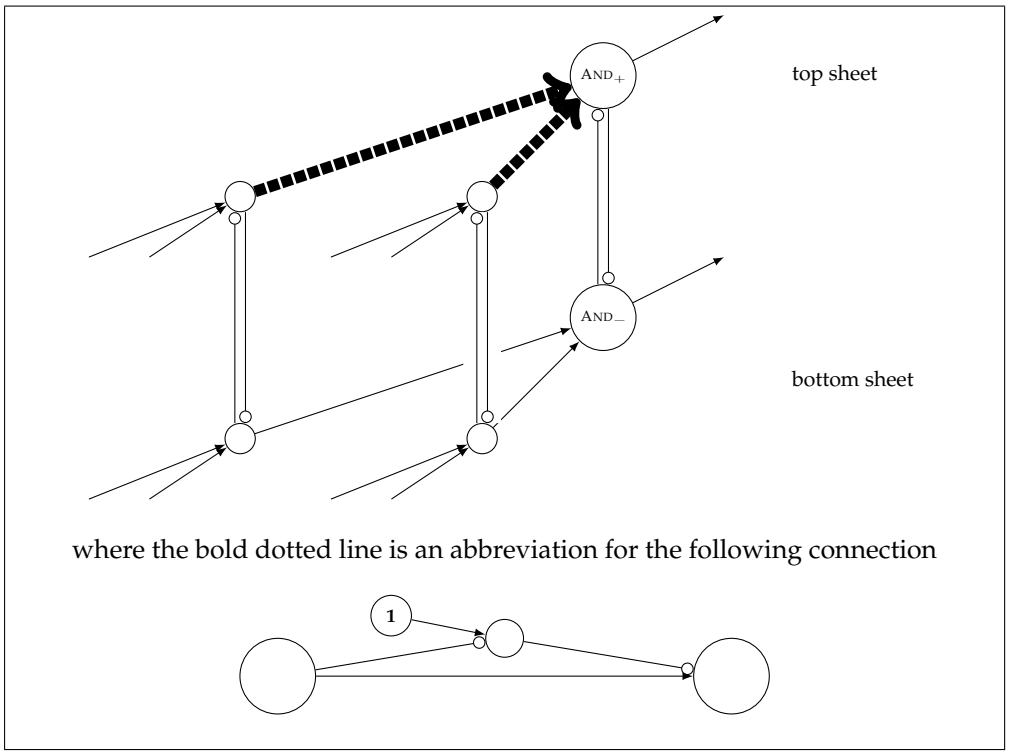
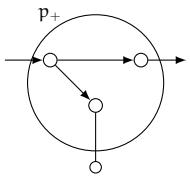


Figure 3.4: Implementation of AND

The input into p_+ and the output from p_+ are excitatory connections (denoted by the arrow $\to$). Similarly for p_-. Both from p_+ to p_- and from p_- to p_+ there is an inhibitory connection (denoted by the arrow $\multimap$). This avoids "blatant" inconsistencies: For example, if p_+ receives excitatory input and turns on, then this shuts off p_- (if it was on) and makes sure to not become active again as long as p_+ remains on. If both p_+ and p_- turn on simultaneously, then they shut each other off in the next step, rendering p undecided. The weights of the arrows are always 1 unless indicated otherwise.[82]

We implement the AND connective as shown in figure 3.4, where the unit denoted by **1** is a unit that continuously fires (i.e. always has value 1). Let's convince us that this indeed computes AND. Assume the input to the nodes p and q is such that $p_+ = 0.3$, $p_- = 0$ (so p is true) and $q_+ = 0$, $q_- = 0$ (so q is undecided). Then in the bottom layer the activation sums to 0, so the unit AND_- so doesn't fire (i.e. has value 0). In the top layer, we look at the bold dotted connection from q_+ to AND_+: Since

[82] Here we see one reason why we speak of "units" rather than "neurons". The unit p_+, for example, has both inhibitory and excitatory outgoing connections. In the neurons of our brain, this is not possible; so we technically should think of the unit p_+ as a cluster of neurons containing at least the following neurons:



As this "trick" works universally (and to reduce the complexity of our drawings) we omit this complication and allow our *units* to have both inhibitory and excitatory outgoing connections.

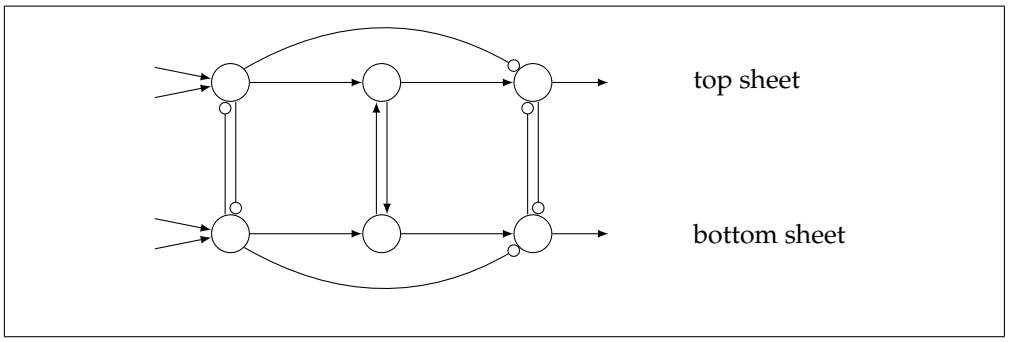


Figure 3.5: Implementation of NOT

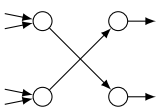
q_+ is off, the intermediate neuron of the connection is 1 (as it receives input 1) and hence inhibits the AND_+ unit (no matter what p_+ does). Thus, the AND node has value $(0, 0)$, that is, undetermined—just as the truth-table of figure 3.1 for conjunction in three-valued logic tells us. Notice why we have the bold dotted connection instead of regular excitatory connections: If we had regular excitatory connections, then the input would sum up to 0.3 making AND_+ fire—although not all its inputs are on. The bold dotted connection avoids this problem acting like a regular excitatory connection if the input is on, and like an inhibitory connection if the input is off. This more complicated connection is the price to pay for neurons with values in $[0, 1]$ instead of binary ones.

Note that by just adding more sentences with the same type of connections to the AND node, this implementation straightforwardly generalizes to n-ary (and not just 2-ary) conjunctions. Also, we could normalize the input that the AND node receives to 1 by putting a weight $\frac{1}{n}$ on the n-many connections feeding into it; but for our purposes this is not needed (if a unit receives an input > 1, its value stays at 1).

To implement OR we just take AND and swap the bold dotted lines in the top layer with their corresponding regular excitatory connection in the bottom layer (and write "$\text{OR}_{+/-}$" instead of "$\text{AND}_{+/-}$").

We implement NOT as shown in figure 3.5. Let's convince us that this indeed computes NOT: For example, if the input is true, that is, the leftmost node has, say, value $(0.7, 0)$, then the 0.7 propagates through to the top and bottom unit of the rightmost node, but the top unit gets shut off by the top left unit—so the output is $(0, 0.7)$, as wanted.[83] Again there is a price to pay to have continuous rather than binary units: In order to get the exact activation of the input unit on one sheet to its "negated" unit on the other sheet, we need to allow for either excitatory or inhibitory connections between the units of one node. For binary units we can implement NOT with inhibitory connections only (Stenning and van Lambalgen 2008, 228f.).

[83] One could implement NOT just by crossing wires:



However, then it would not be the case anymore that the only links between the sheets are those between units that form a node. Yet, we want to keep this feature for reasons of biological plausibility (cf. section 3.3.2).

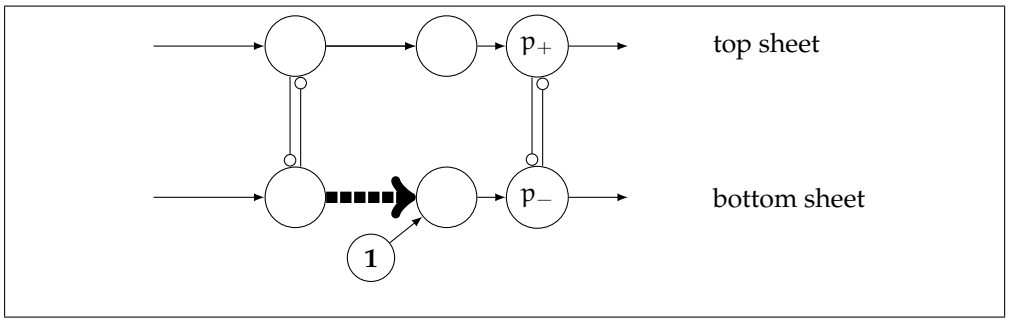


Figure 3.6: Implementation of a sentence p to which negation as failure applies

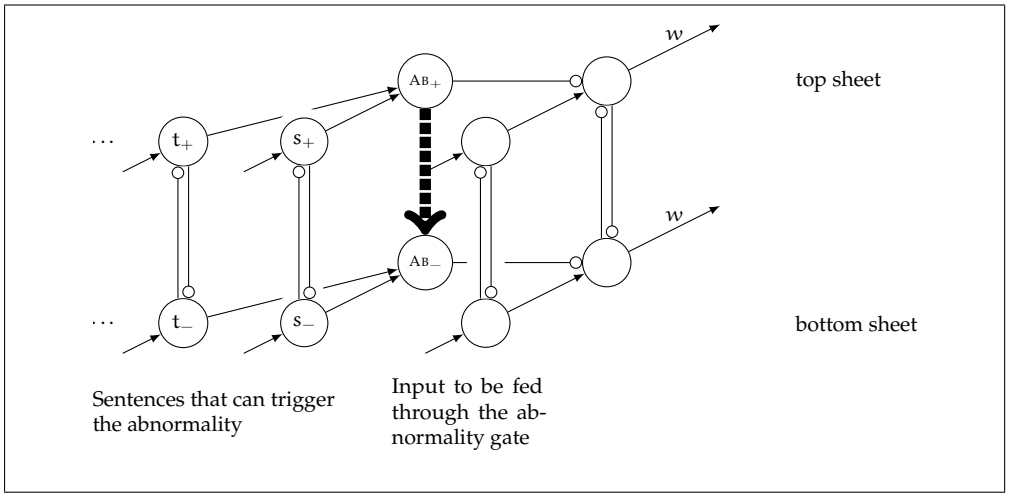


Figure 3.7: Implementation of AB

In figure 3.6, we show how to implement that negation as failure applies to a sentence p: Let's consider two cases. First, if the node p doesn't have any input (so the left node and its arrows are not present), then no information that p is the case can arrive, whence p is set to false. Second, if the node p has input (so the left node is present; which later will correspond to there being a clause $\varphi \to p$ in the program), then p will only be set to false if its input sets it to false (and p is not set to false by default via the **1**-unit). And if p receives excitatory input in the positive sheet, the **1**-unit in the bottom sheet is inhibited, and p becomes true.

Finally, we implement AB as shown in figure 3.7. For now, the weights w on the output node are – like all other weights – equal to 1, but we could set w to be a number slightly less than 1 to model the fading out of signals. We assume that the possible abnormalities are "gathered" in the top sheet: That is, if one of the sentences that can trigger the abnormality actually triggers the abnormality, then this happens because the *positive* unit of this sentence fires.[84] (This is not a real restriction, though, as we could stack a NOT after the sentence and then feed it into the AB node; also, if the abnormality is not specified by a single propositional letter but rather by a conjunction of several ones, we first AND them together.) Thus, on the top sheet, the units of the sentences that can trigger the abnormality are just summed into AB_+.[85]

[84] Syntactically that means that clauses specifying abnormalities are always of the form $s \to ab$.

[85] For reasons of symmetry we do the same on the bottom sheet, but – as is easily checked – this doesn't have much of an effect.

And if they reach the threshold of AB_+, then both AB_+ and AB_- will fire, whence inhibiting any input to pass the gate. So, if no abnormality is present, both the top and bottom unit of the gate is open, and if an abnormality is present, both units are shut—just as expected from an abnormality gate.[86, 87]

Ad (iv). Now that we have the building blocks, we can build the network for a given program. Let a general logic program P be given together with a finite $N \subseteq S$ (the set of sentences to which negation as failure is applicable). Without loss of generality we may assume for all clauses in P that they either only contain propositional variables (and no $\bot$ or $\top$), or are of the form $\bot \to q$ or $\top \to q$.[88] We construct the coupled recurrent neural network $CRN(P, N)$ implementing P under N as follows. For every atom p in P create a node (p_+, p_-), this forms the "leftmost" nodes (we call them input nodes). Copy them to form the "rightmost" nodes (we call them output nodes), and for each output node, put a connection back into its corresponding input node (i.e. a link from output-p_+ to input-p_+, and from output-p_- to input-p_-). If q is a proposition letter to which negation as failure is applicable (i.e. $q \in N$ and q is not an abnormality clause) stack the units shown in figure 3.6 in front of the output node p. For the middle nodes, go through all clauses

$$p_1 \wedge \ldots \wedge p_n \wedge \sim r_1 \wedge \ldots \wedge \sim r_k \wedge \sim ab \to q$$

in P (where q is a propositional variable and not an abnormality clause; the latter will be dealt with separately below) and for each do the following.

- If q is a fact in P (i.e. the clause is of the form $\top \to q$), add a link from the **1**-unit[89] into the plus component of the output node q; then go to the next clause. Otherwise continue.

- If the clause is of the form $\bot \to q$ and $q \notin N$,[90] stack the units shown in figure 3.6 in front of the output node p; then go to the next clause. Otherwise continue.

[86] Of course, there are other ways to implement AB: For example, we could choose not to have links from s_- to AB_- (analogously for t_-, etc.) and instead of the bold dotted line have a single dotted line. (Though, this would break the symmetry between the top and bottom sheet).

[87] Stenning and van Lambalgen (2008, sec. 8.5) implement the abnormality node differently. We have two reasons for departing.

First, they can't have a link from AB_- to the output (otherwise s_- could trigger the output), but then the falsity of the node p cannot be forwarded to the falsity of q which, however, is required in the program $\{p \wedge \sim ab \to q, \bot \to ab, \bot \to p\}$.

Second, in their implementation abnormalities have to come in the form $\sim s \to ab$. This, however, doesn't allow to apply negation as failure to abnormality clauses as intended: For example, take $\{p \wedge \sim ab \to q, \sim s \to ab, \top \to p\}$. Intuitively, since we know nothing about s and negation as failure doesn't apply to s, we should conclude by closed-world reasoning that there is no abnormality present (so that both p and q are the case according to the program). Formally, however, this cannot be done: If ab is described by $\sim s \to ab$, then to conclude $\sim ab$ we need to have – by completion – s as a fact in the program, which we don't.

They put forward two reasons for their implementation in their footnote 17 (p. 236): (i) the fact that inhibition plays a crucial role in abnormality nodes, and (ii) one can linearly add more sentences feeding into the abnormality node without needing to change the node itself. Both of these points are also realized in our implementation. What doesn't work (straightforwardly) anymore, though, is their explanation (p. 237) why – by learning – a link between s_+ and AB_+ forms if there both already is a link between s_- and AB_- and s_+ fires—however, see our section 3.3.2.

[88] Otherwise we can transform P into an equivalent program that has this property.

[89] In general, we could also choose a unit that constantly fires with a value $e \in [0, 1]$ representing the evidence we wish to assign to the fact p.

[90] If $q \in N$ we already did what we are about to do.

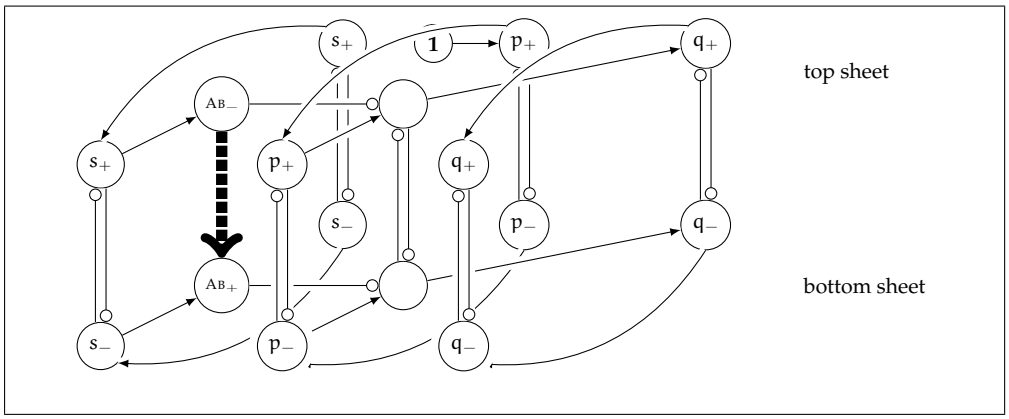


Figure 3.8: Implementation of $\{p \wedge \sim ab \to q, s \to ab, \top \to p\}$

(By our assumption now the clause has only propositional variables in its body).

- Feed the nodes corresponding to $p_1, \ldots, p_n$ directly into an $n+k$-ary AND. Feed the nodes corresponding to $r_1, \ldots, r_k$ first (separately) into NOTs, then into the remaining inputs of the $n+k$-ary AND.
- Feed the output of the AND through an AB node, and feed into this AB node all the sentences $s, t, \ldots$ that can trigger the abnormality clause ab of the given clause according to P (as shown in figure 3.7).[91]
- If q didn't occur as the head of a clause so far, feed the output of the AB directly into the output node for q. Otherwise, add the AB as a new input of an OR which then feeds into q.

This concludes the construction of $CRN(P, N)$.

As an example, consider the program $P = \{p \wedge \sim ab \to q, s \to ab, \top \to p\}$ with $N := \{p, q, s, ab\}$ (but the choice of N doesn't matter in this example). Its corresponding coupled recursive neural network $CRN(P, N)$ is represented in figure 3.8 (the implementation of the sentences to which negation as failure is applicable has been suppressed). This net computes as follows. At time $t = 0$, all nodes are inactive (in both components). At $t = 1$, the facts in P become $(1, 0)$: output-p_+ switches to 1 (and thus inhibits output-p_-) since it received input 1 from the **1**-unit. The output nodes to which negation as failure is applicable – and which have no other inputs – become $(0, 1)$: output-s_- is set to 1, output-p_- stays 0 since output-p_+ is 1, and output-q_- stays 0 since it has an input connection. All other nodes remain inactive in both units. For larger t the computation is as in recursive neural networks and determined by the AND, OR, NOT and AB nodes. For example, at $t = 2$, activation from the output nodes spreads to the input nodes. At $t = 3$, the top unit of the unnamed node in the middle turns on (since input-p_+ is 1). At $t = 4$, this activation spreads to q_+. Then, for any further point in time $t' > 4$ no further change occurs. We've reached the stable state of the network: p and q are true, s is false. So $CRN(P, N)$ indeed computes the minimal model of P under N: the activation of the output nodes in the stable state

[91] For a general implementation of logic programming (where not every clause is assumed to be inhibitable by an abnormality), this bullet point wouldn't apply (and if there still are some abnormality clauses we treat them simply as sentences).

of the network represents the minimal model. And this computation proceeds as the consequence operator (cf. definition 3.2.5). In general we have:

Theorem 3.3.1 (Neural implementation of logic programming[92])**.** *Let* P *be a general logic program,* $N \subseteq S$ *finite, and* $CRN(P, N)$ *the associated coupled recurrent neural network. The least fixed point of* $T_{(P,N)}$ *(i.e. the minimal model of* P *under* N*) corresponds exactly to the activation of the output layer in the stable state of* $CRN(P, N)$*.*

Note that in order for the theorem to make sense we have to consider the networks where the facts get assigned 1 as a value and all weights are 1 (because the consequence operator is defined only on the values $1, 0, u$). But, of course, the networks also generalize the consequence operator because they also cover any type of reasoning with real-valued evidence.

3.3.2 Some further comments

We comment on how the networks implementing logic programming can be further investigated and developed if space would allow.

First, a lot could and should be said on the biological plausibility of the networks implementing logic programming. For example, how the links between the units are formed by the short- and long-term memory, how abnormalities are implemented via so-called interneurons, or how the two layers are formed by a process that weeds out all cross-connections and only leaves the connection between the units of a node. (This process is a fundamental operation of the brain and also occurs, for example, in vision.) For reasons of space, we cannot go into this here, yet Stenning and van Lambalgen (2008, sec. 8.4) provide a good overview of the relevant research.

Second, much more could and should be said on learning—which is *the* advantage of neural networks over symbolic approaches. For example, assume we have a conflict between what we reason to be the case and what we observe to be the case. That is, assume that we have evidence that p is true (turning on the top unit of the p-node) but we reason with the network that from the other facts and rules it should follow that p is false (turning on the bottom unit of the p-node). Then a learning process should add a new abnormality to the network that blocks the reasoning from the other facts to p being false (or the evidence should be reconsidered). To explain such a learning process, not only the well-known Hebbian learning is important but also especially the Anti-Hebbian learning to provide inhibitory connections (for an overview, see Choe 2013).

Third, if space would allow, we could compare our framework to other neural implementations of logic programming: for example, to those mentioned in footnote 78, and especially to the one of Leitgeb (2003, sec. 5).

3.3.3 Neural individuation of cognitive role

We can now use this neural implementation of the logic programming framework to neurally describe individuation by cognitive role.

The general knowledge base KB and the knowledge base KB_a of an agent $a \in A$ are neurally implemented as the networks $CRN(KB)$ and $CRN(KB_a)$ respectively (for

[92] Due to Stenning and van Lambalgen (2008, p. 230).

definiteness choose $N := \emptyset$). The neural implementation of a context $c = (P_c, F_c, N_c)$ is the network $CRN(c^{KB})$ where we also specify the values $a \in [0, 1]$ of the facts in F_c. Thus, we can formulate:

Likeness of cognitive role, neurally Two atomic sentences $p, q \in S$ have the same cognitive role (in context c) iff the in the implementation of the generic knowledge base $CRN(KB)$ (resp. in $CRN(c^{KB})$) there is a connection from the input node p via an abnormality node to the output node q, and there is the same kind of connection from q to p. The version relative to an agent $a \in A$ is again obtained by adding the index 'a' to KB (resp. c^{KB}).

For now, this concludes our reconstruction of the concept of cognitive synonymy as likeness in cognitive role. In the next chapter, we generalize the framework developed here to also apply to logically complex sentences—which we do by developing the notion of a scenario. In the final chapter, we will locate cognitive synonymy in the zoo of notions of synonymy that we will encounter there.

4th Chapter

SCENARIOS

In the previous chapter, we saw how from a cognitive perspective the content of sentences should be individuated. In this chapter, we see how this framework gives rise to the notion of a scenario. These scenarios can be understood as possible representations of parts of the world (or of possible worlds).

We define these scenarios formally. We see that they can be grounded in states of (idealized) neural networks with which a (possible) agent conceptualizes and reasons about the part of the (possible) world she perceives.

We observe that the class of scenarios is rich in structure: it has, for example, a constructable notion of distance and various modal accessibility relations. We use this to give an (im-) possible world type semantics (including cognitive operators and a counterfactual). These scenarios thus provide hyperintensional content.

This scenario semantics gives rise to various notions of validity which will be characterized. We use the framework in the next chapter to describe notions of synonymy.

4.1 Describing scenarios

We formally define scenarios, we describe some of the structure of the class of scenarios, we show how scenarios can be grounded, and we compare them to other notions of "worlds" like (im-) possible worlds, situations, or states.

4.1.1 Defining scenarios

Let's start by formally defining what a scenario is. Afterward, we will motivate the definition.

Definition 4.1.1 (Scenario)**.** A *scenario* (or state) s is a quadruple (S_s, P_s, I_s, E_s), where

- $S_s \subseteq S$ is a finite set of atomic sentences,
- P_s is a general logic program[93] whose atoms are a subset of S_s,
- $I_s : S_s \to [0,1] \times [0,1]$ is an *interpretation* of the sentences, and
- $E_s : S_s \to [0,1] \times [0,1]$, where we call the a pair (a, b) assigned to an $p \in S_s$ the *evidence* for p, and E_s the *evidence valuation* of S_s.

[93] This (and corresponding terminology) was defined in definition 3.2.1 in chapter 3.

We write

- $Ab(P_s)$ for the set of clauses that have an ab in the head,
- $N(P_s)$ for the set of atoms in P_s to which negation as failure is applicable (i.e. those p with $\bot \to p \in P_s$).
- $F(P_s)$ for the set of facts in P_s (i.e. those p with $\top \to p \in P_s$),
- $C(P_s)$ for the set of all clauses in P_s not in the previous sets (the proper clauses).

We fix a $\theta \in [0,1]$ slightly bigger than 0 (for definiteness, say $\theta := 0.1$).[94] Then, each interpretation (and the same for evidence) of a sentence $I(p) = (a,b) \in [0,1] \times [0,1]$ corresponds to a (four-valued) truth-value given by the function $T : [0,1] \times [0,1] \to \{t, f, u, \bot\}$ defined by

$$T(a,b) := \begin{cases} u & \text{, if } a, b < \theta \\ t & \text{, if } a \geqslant \theta \text{ and } b < \theta \\ f & \text{, if } a < \theta \text{ and } b \geqslant \theta \\ \bot & \text{, if } a, b \geqslant \theta. \end{cases}$$

(We also write 1 for t and 0 for f.) Thus, an interpretation I_s of S_s corresponds (many-to-one) to a four-valued model $M_{I_s} : S \to \{u, t, f, \bot\}$ defined by $M_{I_s}(p) := T\big(I_s(p)\big)$ for all $p \in S_s$, and $(0,0)$ for $p \in S \setminus S_s$. Analogously, we define the model M_{E_s} corresponding to an evidence E_s. We call s *well-behaved* if S_s is the set of atoms of P_s, $M_{E_s} \leqslant M_{I_s}$, and M_{I_s} is the minimal model of P_s (and hence is three-valued). Well-behaved scenarios are the "normal" ones: the interpretation makes true or false exactly those sentences for which the program provides a reason for their truth or falsity, there is no contradiction in the interpretation, and the interpretation is consistent with the evidence. As the program is the core part of a scenario (the minimal model can tractably be computed and the evidence is only really relevant if it contradicts the minimal model), we often identify a scenario with its program—the context makes clear whether we mean the full-fledged scenario or its core. We denote the class of all scenarios by Σ.

Three remarks. First, intuitively a scenario represents a "part" of the world as conceptualized by an agent (so it could be an inconsistent conceptualization).[95] The sentences in S_s are the sentences that are *about* this part of the world. The program

[94] This is the threshold of a neuron (or, rather, unit) in the neural implementation of logic programming described in section 3.3.

[95] There is one terminological subtlety here: In a more narrow sense of "representation", one could claim that an inconsistent representation (like a scenario with an inconsistent interpretation) cannot represent something possible (like a part of a possible world). On this view, a representation both aims at representing something and it succeeds to a satisfying degree—i.e., a representation has a descriptive and a normative aspect. In a wider sense of "representation", one could hold that an inconsistent representation still can represent something possible. On this view, a representation aims at representing something (cf. the descriptive aspect) but the link between the representation and the represented might not be close enough for the intended use of the representation (the representation doesn't meet the normative demands). For example, naive set theory with the unrestricted comprehension axiom can be said to be *a* representation of the notion of a set—just not a very good one because it is inconsistent. We'll use the latter, wider sense, but those uncomfortable with this use may replace it by "conceptualization" or "description" where the normative aspect is separated more clearly from the descriptive one.

P_s describes the (non-logical) rules and facts obtaining in that part of the world, and the sentences in $N(P_s)$ are those that can potentially be falsified by that part of the world. Next, E_s measures how much evidence there is for a sentence $p \in S_s$: If $E_s(\varphi) = (a, b)$, then a describes how much positive evidence there is for p and b how much negative evidence there is for p. Finally, the interpretation I_s describes – in the usual cases – the model of that part of the world obtained by reasoning according to the program starting from the evidence. Thus, a scenario is a possible representation of a part of a possible world. So scenarios are – in a sense – one level above possible worlds: They are possible representations of possible worlds. This is one way to "ground" scenarios, and we'll get back to this in section 4.1.5 below where we also mention other ways.

Second, on the one hand, scenarios have a representational aspect like Fregean senses: the program of a scenario can be interpreted as a cognitively significant mode of presentation of the world. On the other hand, scenarios have an interpretational aspect like a possible world: the interpretation of a scenario describes what sentences are true (or $f, u, \bot$) according to the scenario.

Third, there is a close connection between scenarios and the notion of context introduced in the last chapter (definition 3.2.8). Scenarios correspond many-to-one to the contexts $\big(C(P_s) \cup Ab(P_s)\ ,\ F(P_s)\ ,\ N(P_s)\big)$. What is added to get a scenario is the evidence and interpretation.

4.1.2 The structure of the class of scenarios: Accessibility relations

The class Σ of all scenarios carries many natural accessibility relations that we describe here.

1. $sR_S s'$ iff $S_s \subseteq S_{s'}$, $P_s = P_{s'}$, $I_s = I_{s'}$, and $E_s = E_{s'}$. In words: s' is obtained from s by adding "new subject matter" – i.e., new propositional letters – and keeping the rest the same.

2. $sR_{(Ab,\subseteq)}s'$ iff $S_s = S_{s'}$, $Ab(P_s) \subseteq Ab(P_{s'})$, $I_s = I_{s'}$, and $E_s = E_{s'}$. In words: s' is obtained from s by adding new abnormality-specifying clauses and keeping the rest the same.

   Of course, not only the abnormality clauses could be added. We could add clauses to $N(P_s)$ (resp. $F(P_s)$ or $C(P_s)$) while keeping the rest fixed. And we get yet more accessibility relations by deleting (instead of adding) clauses. Each of these eight accessibility relations is interesting, but for reasons of space, we'll only consider $R_{(Ab,\subseteq)}$.

3. $sR_{E_1}s'$ iff $S_s = S_{s'}$, $P_s = P_{s'}$, $I_s = I_{s'}$, but possibly $E_s \neq E_{s'}$. In words: s' is obtained from s by changing some of the evidence for sentences in S_s.

   $sR_{E_2}s'$ iff $S_s = S_{s'}$, $P_s = P_{s'}$, $E_s = E_{s'}$, and $I_{s'}(p) = max\big(E_s(p), I_s(p)\big)$.[96] In words: s' is obtained from s by merging the evidence with the interpretation.

4. $sR_U s'$ (U for *u*pdate) iff $S_s = S_{s'}$, $P_s = P_{s'}$, $E_s = E_{s'}$, and if we set the activation in the output-nodes of the network $CRP(P_s)$ as given by I_s, then after n-many

[96] The maximum is taken pairwise: $max\big((a, b), (c, d)\big) := \big(max(a, c), max(b, d)\big)$.

computation steps of the network (for some $n \in \{0, 1, 2, 3, \ldots\}$), the output nodes are as given by $I_{s'}$. In words: s' is obtained from s after some computation steps in the underlying neural network.

And there are more: For example those changes to the program that can occur through a specific kind of learning in the network. Or the change in the program that moves from an agent-based knowledge base to the generic version (i.e. taking the intersection of the given program with the generic knowledge base).

4.1.3 The structure of the class of scenarios: Pseudometric

In this section, we show how to construct a pseudometric on the class of scenarios. This will give us a notion of distance between scenarios.

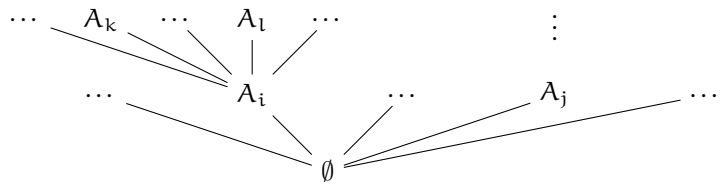
We start by observing that, given a scenario s, there are clauses that are more likely to extend s than other clauses. For example, consider the scenario that describes me sitting at a desk and writing these lines. Now consider two possible extensions of that scenario: One extending it by the fact that outside it is raining, and one by the fact that outside there is no gravity anymore. Then, clearly, the former extension is more common, similar, plausible, or likely than the latter. (We deliberately leave open the exact interpretation of this ordering.) We capture this ordering by assigning those clauses that are very common small numbers and those that are less common bigger numbers—and clauses that are either equally common or cannot be compared in terms of commonness get the same number.

To do so, we fix for the whole chapter a bijective enumeration e of all the clauses: so, writing Cls for the set of all general logic programming clauses, $e : \mathsf{Cls} \to \mathbb{N}$ is a bijection. To simplify notion, we often identify clauses and their corresponding numbers. Then we define:

Definition 4.1.2 (Clause extension ordering ρ)**.** A *clause extension ordering* ρ is a function $\rho : \Sigma \to \left(\mathbb{R}_{\geqslant 1}\right)^{\mathbb{N}}$ that assigns each scenario s a sequence $\rho(s)$ of real numbers $\geqslant 1$ (i.e. $\rho(s) : \mathbb{N} \to \mathbb{R}_{\geqslant 1}$).

Thus, all the possible extensions of s by one clause are of the form $s \cup \{A\}$ where $A \in \mathsf{Cls} \setminus s$ (and we identify A with its corresponding number), and they are ordered by their "plausibility" as given by $\rho(s)$.

Now, we describe the tree structure of Σ. This is obtained by identifying scenarios by their program ($s \equiv s'$ iff $P_s = P_{s'}$), and then consider the tree of program extensions:



For example, the program consisting of the clauses A_i and A_l can either be realized by the sequence of extensions $\langle A_i, A_l \rangle$ or $\langle A_l, A_i \rangle$. So now we also take the order into account in which the clauses have been added to a program—which so far wasn't

necessary.[97] This is welcomed because it might well be that the plausibility of a program depends on the order in which it was constructed: A program that was constructed by adding clauses each being a bit less plausible than the previous might in sum be more plausible, than obtaining the same program by first adding the last and very implausible clause, then a more plausible one, then again an implausible one, and so on. Thus, with our simplifications we now regard a program as a finite sequence of natural numbers: $s = \langle s(1), s(2), \dots, s(n)\rangle$.

Definition 4.1.3 (Tree-structure of (Σ, ρ))**.** Given a clause extension ordering ρ, the tree structure of (Σ, ρ) is given by the tree T_ρ which is constructed as follows:

- The nodes in the tree are finite injective sequences of natural numbers.[98]
- The root of T is the empty sequence $\langle\,\rangle$.
- Given a node s, its successors are the nodes $s^\frown n$ for $n \in \mathbb{N} \setminus s$,[99] and they are ordered from left to right by

$$s^\frown m < s^\frown n \text{ iff } \begin{cases} \rho(s)(m) < \rho(s)(n) & \text{, or} \\ \rho(s)(m) = \rho(s)(n) \text{ and } m < n & . \end{cases}$$

So the further we go to the top of the program extension tree T_ρ, the more implausible the programs become (since this corresponds to making more and more assumptions). And the further we go to the right in the tree, the more implausible the programs become (since we add more implausible clauses). Thus, we define the plausibility or *weight* of a program by

$$\mu^*_\rho(s) := \sum_{i=1}^{\mathtt{length}(s)} \rho\Big(s \restriction i-1\Big)\big(s(i)\big),$$

where $s \restriction 0$ is the empty sequence. So $\mu^*_\rho(s)$ just adds up the ρ-values of the nodes along the path to the program s.

We say ρ is *symmetric*, if for all $s = \langle s_1, \dots, s_n\rangle$ and all permutations π of $\{1, \dots, n\}$ we have $\mu^*_\rho(\langle s_1, \dots, s_n\rangle) = \mu^*_\rho(\langle s_{\pi(1)}, \dots, s_{\pi(n)}\rangle)$. If we speak of ρ in a context where we think of the programs of scenarios as sets (and not as finite sequences) we tacitly assume that ρ is symmetric.

Finally, we can define a metric on Σ given a clause extension ordering ρ.

Definition 4.1.4 (Induced metric d_ρ)**.** Fix a clause extension ordering ρ and define the pseudometric $d_\rho : \Sigma \times \Sigma \to \mathbb{R}_{\geqslant 0}$ by $d_\rho(s, s') := |\mu^*(s) - \mu^*(s')|$.

This is indeed a pseudometric because the properties of non-negativity, symmetry, and triangle inequality[100] are directly inherited from the properties of the absolute

[97] We could have done that from the start (and then identify programs with the same clauses but different order on every occasion we used them so far), but that would have resulted in an unnecessarily complicated notation.

[98] Where a sequence is injective if no number occurs twice.

[99] Where $s^\frown n$ is the sequence obtained from s by adding n as the last number. And $\mathbb{N} \setminus s$ is short for $\mathbb{N} \setminus \{s(i) \mid 1 \leqslant i \leqslant \mathtt{length}(s)\}$.

[100] Formally, these properties are, respectively, $\forall s, s' \in \Sigma : d_\rho(s, s') \geqslant 0$, $\forall s, s' \in \Sigma : d_\rho(s, s') = d_\rho(s', s)$, and $\forall s, s', s'' \in \Sigma : d_\rho(s, s'') \leqslant d_\rho(s, s') + d_\rho(s', s'')$.

value function $|\cdot|$. In general, it is not a metric (proper) because it doesn't satisfy $d_\rho(s,s') = 0 \Leftrightarrow s = s'$ since different scenarios might have the same weight.

Moreover, this pseudometric induces a system of spheres in the sense of Lewis (1973, 13f.): Given a scenario $i \in \Sigma$, the set $\$_i$ of spheres around i is given by $\$_i := \left\{B^\rho_\epsilon(i) \mid \epsilon \in \mathbb{R}_{\geqslant 0}\right\}$ where $B^\rho_\epsilon(i)$ are the ϵ-balls around i, that is, $B^\rho_\epsilon(i) := \{s \in \Sigma_0 \mid d_\rho(i,s) \leqslant \epsilon\}$.[101]

4.1.4 The structure of the class of scenarios: Negligible sets

In this section, we provide a precise way of saying that a set of scenarios is negligible. We will use this in section 5.1 to say that two sentences are synonymous if the set of scenarios where they differ in truth-value is negligible.

There are many ways to provide a formal notion for a "negligible" set. For example: as a null set under a given measure, as a meager set (or one of first category) under a given topology, as a finite set, or as a bounded set under a given partial order. For all of these notions, the class of negligible sets forms an ideal: it is non-empty, closed under subsets, and closed under finite unions. Here we do the following.

Definition 4.1.5 (Negligible set)**.** Given a clause extension ordering ρ, define $\mu_\rho : \mathcal{P}(\Sigma) \to [0,\infty]$ by

$$\mu_\rho(D) := \sum_{s\in D} \frac{1}{\mu^*_\rho(s)},$$

where $\mu^*_\rho(s)$ is the weight of program s, as defined in definition 4.1.3. Then say

D is negligible iff $\mu_\rho(D) < \infty$.

The class $\mathcal{D}$ of negligible sets of scenarios is indeed an ideal. It clearly is non-empty (since $\emptyset \in \mathcal{D}$). It is closed under subsets: If $D' \subseteq D \in \mathcal{D}$, then

$$\mu_\rho(D') = \sum_{s\in D'} \frac{1}{\mu^*_\rho(s)} \leqslant \sum_{s\in D} \frac{1}{\mu^*_\rho(s)} < \infty.$$

And $\mathcal{D}$ is closed under finite unions: If $D, D' \in \mathcal{D}$, then

$$\mu_\rho(D \cup D') = \sum_{s\in D\cup D'} \frac{1}{\mu^*_\rho(s)} \leqslant \sum_{s\in D} \frac{1}{\mu^*_\rho(s)} + \sum_{s\in D'} \frac{1}{\mu^*_\rho(s)} < \infty.$$

4.1.5 Grounding scenarios

In this section, we briefly mention three different philosophical interpretations of the formal concept of a scenario (or, conversely, three philosophical concepts that can be rationally reconstructed by scenarios).

In short, scenarios can either be regarded as describing agent-independently how the world actually is, or how an agent either actually conceptualizes the world

[101] It is easily checked that $\$_i$ fulfills the axioms demanded by Lewis (1973, p. 14): it is nested, closed under unions, and closed under nonempty intersections. However, it is not necessarily centered since we only have a pseudometric and not a proper metric, so we don't necessarily have $\{i\} \in \$_i$ since $B_0(i)$ could contain more scenarios than just i.

(descriptively) or how she should conceptualize the world (normatively). This is summarized in figure 4.1.5, and we'll now discuss it in a bit more detail.

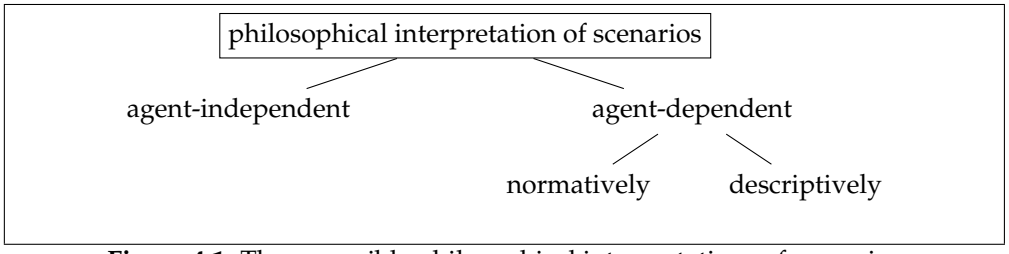


Figure 4.1: Three possible philosophical interpretations of scenarios

The agent-dependent interpretation. Scenarios are grounded in our cognition: They are grounded in the neural network with which an agent reasons about the part of the world she perceives (cf. section 3.3). They describe how a possible agent conceptualizes a part of a possible world.

This again can be understood in two ways: Either *descriptively* as how a given (actual) agent in fact does conceptualize the world, or *normatively* as how an agent should conceptualize the world (or, arguably equivalently, how the generic agent would conceptualize the world).

It is on the agent-dependent interpretation that the capabilities of logic programming to model cognitive abilities straightforwardly carries over to our scenario framework. Examples include: how logic programming can describe the ways actual agents handle inconsistencies (cf. Stenning and van Lambalgen 2016, 361f.), or how logic programming can describe framing, that is, how the formulation of a reasoning task affects its result (cf. Stenning and van Lambalgen 2008, esp. ch. 7).

The agent-independent interpretation. But, in principle, we can also see scenarios as (largely) independent of cognition: The program just describes how the world *actually* is (and not how it is conceptualized by an agent). This might be plausible, for example, in the case of causality (then $p \wedge \neg ab \rightarrow q$ would say that p causes q).

Note that a choice in these interpretations corresponds to a natural choice for the philosophical interpretation of the clause extension ordering ρ from section 4.1.3: The agent-independent interpretation goes well with a similarity interpretation of ρ, and the descriptive (resp. normative) interpretation goes well with a subjective (resp. normative) plausibility interpretation of ρ.

4.1.6 Comparing scenarios

For reasons of space, we cannot provide a detailed comparison of our scenarios to other notions of "worlds" like situations, possible worlds, impossible worlds or states. So we only briefly sketch some such comparisons.

Most obviously, scenarios can be compared to the situations of Barwise and Perry (1983).[102] Both are, roughly, incomplete (representations of) parts of the world. A difference is that situations are object-based entities (they represent objects and their relations), while scenarios are rule-based entities (they represent the rules that are taken to govern that part of the world).

[102] They are used to provide situation semantics starting with Barwise (1981).

Next, scenarios can be compared to possible worlds as follows. Any possible world is the limit of a sequence of appropriate scenarios: The possible world is the union of the minimal models of the programs of the scenarios in the sequence. Similarly for impossible worlds, if an impossible world is taken to be (isomorphic to) a set of literals.[103] (Also see Lindström (1991, sec. 6) for a similar idea on the relation between situations and possible worlds.) There is more to be discovered: For example, under what conditions the truth-value of an (atomic) sentence stabilizes after finitely many scenarios in the sequence.

Moreover, scenarios can be compared to the "meaning as algorithm" approach (from section 1.2.3): The program of a scenario can be seen as an instruction on how to build the minimal model. Though, scenarios are more than mere procedures since they also have an interpretational component.

Finally, in the next chapter (section 5.3), we'll see how scenarios can be related to states in truthmaker semantics.

4.2 Semantics with scenarios

In this section, we define a semantics based on scenarios (section 4.2.1), we investigate the various notions of validity it gives rise to (section 4.2.2), and we discuss the counterfactual that the semantics provides (section 4.2.3).

4.2.1 Defining the semantics

To define the scenario semantics we first fix a language for which it provides a semantics.

Definition 4.2.1 (Languages $\mathcal{L}$ and $\mathcal{L}_p$)**.** We fix the language $\mathcal{L}$ obtained by the following grammar:[104]

$$\mathcal{L} ::= p \in S \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \rightarrow \psi \mid B\varphi \mid C_{Ab}\varphi \mid C_e\varphi \mid \varphi \boxrightarrow \psi$$

where B is the belief operator, C_{Ab} and C_e are conceivability operators, and $\boxrightarrow$ is a counterfactual conditional. (We could introduce more operators for the accessibility relations that we mentioned in section 4.1.2.)

The language $\mathcal{L}_p$ is the propositional fragment of $\mathcal{L}$, that is, $\mathcal{L}_p ::= p \in S \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi$. (As already observed, we can view $\varphi \rightarrow \psi$ as an abbreviation for $\neg\varphi \vee \psi$.)

Now we can use the scenarios to give a semantics for this language. Since we use a four-valued logic (interpretations can not only be true and false but also undetermined or inconsistent), we have to specify both the truth- and falsity-conditions.

Definition 4.2.2 (Semantics for $\mathcal{L}$)**.** Let $s \in \Sigma$ and let ρ be a clause extension ordering (only needed for $\boxrightarrow$). We recursively define $s \Vdash^{+}_{(\Sigma,\rho)} \varphi$ and $s \Vdash^{-}_{(\Sigma,\rho)} \varphi$ for $\mathcal{L}$-sentences φ:

[103] This correspondence doesn't work (unrestrictedly) anymore if an impossible world is taken to be (isomorphic to) to an arbitrary function from sentences to truth-values. Such functions don't need to commute with the truth-functions of the operators (e.g. a conjunction could be false according to the impossible world while both conjuncts are true).

[104] In the last chapter, we used "$\sim$" as the negation symbol to stress that it is not classical negation. We take this to be understood now and go back – for better readability – to using "$\neg$".

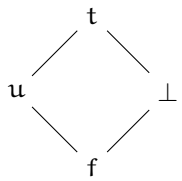
- $s \Vdash^{+}_{(\Sigma,\rho)} p$ iff $T(I_s(p)) \in \{t, \bot\}$.
  $s \Vdash^{-}_{(\Sigma,\rho)} p$ iff $T(I_s(p)) \in \{f, \bot\}$.
- $s \Vdash^{+}_{(\Sigma,\rho)} \neg\varphi$ iff $s \Vdash^{-}_{(\Sigma,\rho)} \varphi$.
  $s \Vdash^{-}_{(\Sigma,\rho)} \neg\varphi$ iff $s \Vdash^{+}_{(\Sigma,\rho)} \varphi$.
- $s \Vdash^{+}_{(\Sigma,\rho)} \varphi \wedge \psi$ iff $s \Vdash^{+}_{(\Sigma,\rho)} \varphi$ and $s \Vdash^{+}_{(\Sigma,\rho)} \psi$.
  $s \Vdash^{-}_{(\Sigma,\rho)} \varphi \wedge \psi$ iff $s \Vdash^{-}_{(\Sigma,\rho)} \varphi$ or $s \Vdash^{-}_{(\Sigma,\rho)} \psi$.
- $s \Vdash^{+}_{(\Sigma,\rho)} \varphi \vee \psi$ iff $s \Vdash^{+}_{(\Sigma,\rho)} \varphi$ or $s \Vdash^{+}_{(\Sigma,\rho)} \psi$.
  $s \Vdash^{-}_{(\Sigma,\rho)} \varphi \vee \psi$ iff $s \Vdash^{-}_{(\Sigma,\rho)} \varphi$ and $s \Vdash^{-}_{(\Sigma,\rho)} \psi$.
- $\varphi \to \psi$ is abbreviated to $\neg\varphi \vee \psi$.
- $s \Vdash^{+}_{(\Sigma,\rho)} B\varphi$ iff there is an R_U-endpoint in Σ that is R_U-reachable from s, and for all such endpoints $s' \in \Sigma$, $s' \Vdash^{+}_{(\Sigma,\rho)} \varphi$. (If in s all sentence-interpretations and evidence valuations are $(0,0)$, then this is sufficient – but not necessary – for there always being exactly one endpoint.)
  $s \Vdash^{-}_{(\Sigma,\rho)} B_a\varphi$ iff there is an R_U-endpoint in Σ that is R_U-reachable from s, and for all such endpoints $s' \in \Sigma$, $s' \Vdash^{-}_{(\Sigma,\rho)} \varphi$.
- $s \Vdash^{+}_{(\Sigma,\rho)} C_e\varphi$ iff there are scenarios $s', s'' \in \Sigma$ such that $sR_{E_1}s'$ and $s'R_{E_2}s''$ and $s'' \Vdash^{+}_{(\Sigma,\rho)} B\varphi$.
  $s \Vdash^{-}_{(\Sigma,\rho)} C_e\varphi$ iff for all scenarios $s', s'' \in \Sigma$, if $sR_{E_1}s'$ and $s'R_{E_2}s''$, then $s'' \Vdash^{-}_{(\Sigma,\rho)} B\varphi$.
- $s \Vdash^{+}_{(\Sigma,\rho)} C_{Ab}\varphi$ iff there is a scenario $s' \in \Sigma$ such that $sR_{(Ab,\subseteq)}s'$ and $s' \Vdash^{+}_{(\Sigma,\rho)} B\varphi$.
  $s \Vdash^{-}_{(\Sigma,\rho)} C_{Ab}\varphi$ iff for all scenarios $s' \in \Sigma$, if $sR_{(Ab,\subseteq)}s'$, then $s' \Vdash^{-}_{(\Sigma,\rho)} B\varphi$.
- $s \Vdash^{+}_{(\Sigma,\rho)} \varphi \boxdot\!\!\to \psi$ iff
  (a) for all scenarios $s \in \Sigma$, if $s \Vdash^{+}_{(\Sigma,\rho)} \varphi$, then $s \Vdash^{+}_{(\Sigma,\rho)} \psi$, or
  (b) there is an $\epsilon \in \mathbb{R}_{\geqslant 0}$ such that
    * there is a well-behaved scenario $s' \in \Sigma$ such that $d_\rho(s, s') \leqslant \epsilon$ and $s' \Vdash^{+}_{(\Sigma,\rho)} \varphi$, and
    * for all well-behaved scenarios $s' \in \Sigma$, if $s' \Vdash^{+}_{(\Sigma,\rho)} \varphi$ and $d_\rho(s, s') \leqslant \epsilon$, then $s' \Vdash^{+}_{(\Sigma,\rho)} \psi$.

  $s \Vdash^{-}_{(\Sigma,\rho)} \varphi \boxdot\!\!\to \psi$ iff for all $\epsilon \in \mathbb{R}_{\geqslant 0}$, if there is a well-behaved scenario $s' \in \Sigma$ such that $d_\rho(s, s') \leqslant \epsilon$, then there is a well-behaved scenario $s'' \in \Sigma$ such that $d_\rho(s, s'') \leqslant \epsilon$ and $s'' \Vdash^{+}_{(\Sigma,\rho)} \varphi$ and $s'' \Vdash^{-}_{(\Sigma,\rho)} \psi$.

The (Σ, ρ)-truth-value of a formula in a scenario is given by

$$T^s_{\Sigma,\rho}(\varphi) := \begin{cases} t \ (= \langle 1,0\rangle) & \text{, if } s \Vdash^{+}_{(\Sigma,\rho)} \varphi \text{ and } s \nVdash^{-}_{(\Sigma,\rho)} \varphi \\ f \ (= \langle 0,1\rangle) & \text{, if } s \nVdash^{+}_{(\Sigma,\rho)} \varphi \text{ and } s \Vdash^{-}_{(\Sigma,\rho)} \varphi \\ u \ (= \langle 0,0\rangle) & \text{, if } s \nVdash^{+}_{(\Sigma,\rho)} \varphi \text{ and } s \nVdash^{-}_{(\Sigma,\rho)} \varphi \\ \bot \ (= \langle 1,1\rangle) & \text{, if } s \Vdash^{+}_{(\Sigma,\rho)} \varphi \text{ and } s \Vdash^{-}_{(\Sigma,\rho)} \varphi \end{cases}$$

and we order truth-values by the diamond lattice $\preceq$ given thus:



(So this order of truth-values $\preceq$ is different from their ordering by definiteness $\leqslant$ seen in definition 3.2.4.) Note that the truth-function of $\wedge$ (resp. $\vee$) is taking the minimum (resp. maximum) in the lattice, and the truth-function of $\neg$ swaps $\mathtt{t}$ and $\mathtt{f}$ and keeps $\mathtt{u}$ and $\bot$ fixed—so the lattice is not a Boolean algebra.[105] A truth-value is *classical*, if it is $\mathtt{t}$ or $\mathtt{f}$. For a set of $\mathcal{L}$-sentences Γ, we write $\mathsf{T}^s_{\Sigma,\rho}(\Gamma) := \inf_{\preceq}\big(\{\mathsf{T}^s_{\Sigma,\rho}(\psi) \mid \psi \in \Gamma)\}\big)$, where $\inf_{\preceq}(X)$ is the greatest lower bound of $X \subseteq \{0, 1, \mathtt{u}, \bot\}$ (so $\inf_{\preceq}(\emptyset) = 1$).

There are many notions of logical consequence for many-valued logics (see e.g. Chemla et al. 2016). Here we work with the following three (the first are well-known and the last is – to the best of my knowledge – new). Given a set of $\mathcal{L}$-sentences Γ and an $\mathcal{L}$-sentence φ define:

- *Truth-preserving*: $\Gamma \models^{\mathtt{tr}}_{(\Sigma,\rho)} \varphi$ iff$_{\mathtt{df}}$ for all $s \in \Sigma$, if $\mathsf{T}^s_{(\Sigma,\rho)}(\Gamma) \in \{1, \bot\}$, then $\mathsf{T}^s_{(\Sigma,\rho)}(\varphi) \in \{1, \bot\}$.
- Instead of choosing $\{1, \bot\}$ as designated values (i.e. the values preserved by the validity relation) we could also choose $\{1, \mathtt{u}\}$ or $\{1\}$. We denote the resulting validity relations $\models^{\bar{\mathtt{tr}}}_{(\Sigma,\rho)}$ and $\models^{\mathtt{tr}^1}_{(\Sigma,\rho)}$, respectively.
- *Order-preserving*: $\Gamma \models^{\mathtt{or}}_{(\Sigma,\rho)} \varphi$ iff$_{\mathtt{df}}$ $\forall s \in \Sigma : \mathsf{T}^s_{\Sigma,\rho}(\Gamma) \preceq \mathsf{T}^s_{\Sigma,\rho}(\varphi)$.
- *Exclusion-preserving*: $\Gamma \models^{\mathtt{ex}}_{(\Sigma,\rho)} \varphi$ iff$_{\mathtt{df}}$ for all $s \in \Sigma$, if $\mathsf{T}^s_{(\Sigma,\rho)}(\Gamma)$ is classical, then $\mathsf{T}^s_{(\Sigma,\rho)}(\varphi)$ is classical and $\mathsf{T}^s_{(\Sigma,\rho)}(\Gamma) \preceq \mathsf{T}^s_{(\Sigma,\rho)}(\varphi)$.[106]

As usual, we write $\models_{(\Sigma,\rho)} \varphi$ for $\emptyset \models_{(\Sigma,\rho)} \varphi$.

We define the (Σ, ρ)-*content* of a $\mathcal{L}$-sentence φ as

$$[\![\varphi]\!]_{(\Sigma,\rho)} := \big\langle [\![\varphi]\!]^+_{(\Sigma,\rho)}, [\![\varphi]\!]^-_{(\Sigma,\rho)} \big\rangle,$$

where $[\![\varphi]\!]^+_{(\Sigma,\rho)} := \big\{s \in \Sigma \mid s \Vdash^+_{(\Sigma,\rho)} \varphi\big\}$ and $[\![\varphi]\!]^-_{(\Sigma,\rho)} := \big\{s \in \Sigma \mid s \Vdash^-_{(\Sigma,\rho)} \varphi\big\}$. We call the elements of $[\![\varphi]\!]^+_{(\Sigma,\rho)}$ the (Σ, ρ)-*truthmakers* of φ, and the elements of $[\![\varphi]\!]^-_{(\Sigma,\rho)}$ the (Σ, ρ)-*falsemakers* of φ.

If we're only concerned with the propositional fragment $\mathcal{L}_{\mathtt{p}}$, we don't need ρ, and thus drop it as a subscript.

Finally, note that all of these definitions also work for subsets $\Sigma_0 \subseteq \Sigma$ by replacing Σ with Σ_0, yielding $\Vdash^+_{(\Sigma_0,\rho)}, \Vdash^-_{(\Sigma_0,\rho)}, \models_{(\Sigma_0,\rho)}, [\![\cdot]\!]_{(\Sigma_0,\rho)}$. We call subsets of Σ (including Σ itself) *frames*. We drop the sub- or superscripts if they are clear from the context.

[105] Instead, it is a so-called De Morgan algebra. They have the same signature $(2, 2, 1)$ as Boolean algebras and they satisfy the same axioms except for $\neg x \vee x = 1$ and $\neg x \wedge x = 0$.

[106] Arguably, "classicality-preserving" would be a more fitting name, but then it would be to close to "classical validity"—and we will see that $\mathtt{ex}$-validity is very different to classical validity.

Six remarks. First, the following fact relates truth in a scenario ($\Vdash$) to being the case according to the program of the scenario ($\models\!\!\approx$). Let φ be an $\mathcal{L}_p$-sentence Let $s \in \Sigma$ be a well-behaved scenario. Then $s \Vdash^+_{(\Sigma,\rho)} \varphi$ iff $P_s \models\!\!\approx \varphi$ and $s \Vdash^-_{(\Sigma,\rho)} \varphi$ iff $P_s \models\!\!\approx \neg\varphi$. (This is immediate by induction on φ).

Second, if we consider all scenarios, we have a four-valued logic: Interpretations assigning $(1,1)$ to atomic sentences are allowed. If we restrict us to well-behaved scenarios, we have a three-valued logic: The minimal model of a program is at most three-valued (with values in $\{t, f, u\}$). The four-valued logic on the propositional fragment $\mathcal{L}_p$ provided by our semantics is that of *first-degree entailment*.[107] The three-valued logic on $\mathcal{L}_p$ provided by our semantics is *strong Kleene logic* (as already seen in the last chapter).

Third, the notions of validity can be understood by what truth-values they think are good (and better than others) and hence should be preserved when moving from premisses to conclusion:

- tr-validity takes truth (i.e. 1 and $\bot$) to be good, $\bar{\text{tr}}$-validity takes non-falsity (i.e. 1 and u) to be good, and tr^1-validity takes truth and nothing but the truth (i.e. 1) to be good (cf. Pietz and Rivieccio 2013).
- or-validity takes 1 to be better than u and $\bot$, and 0 worse than these two.
- ex-validity takes classical truth-values (i.e. 0 and 1) to be good – because they are in a sense definite –, and it takes 1 to be better than 0.

Fourth, scenarios have both aspects of (model-theoretic) models and of (modal logic) worlds: They are like models in the sense that they are not just structureless objects like worlds, but they rather have a structure that determines the truth or falsity of a formula. They are like worlds in the sense that the class of scenarios has very natural accessibility relations so that it makes sense to define some sort of modal logic on the class of scenarios where the scenarios act like worlds. (This is similar to, for example, directed systems in model theory, first order intuitionistic logic, or the modal logic of forcing in set theory).

On the modal logic perspective, any frame $\Sigma_0 \subseteq \Sigma$ together with the relations specified in section 4.1.2 (restricted to Σ_0) forms a Kripke frame. (And if we add a ρ we get a sphere model on top of the frame.) The valuation of the frame Σ_0 doesn't need to be specified as it is given by $\Vdash^+_{\Sigma_0}$ and $\Vdash^-_{\Sigma_0}$. Of course, the frame dependent notions $\Vdash^+_{(\Sigma_0,\rho)}, \Vdash^-_{(\Sigma_0,\rho)}, \models_{(\Sigma_0,\rho)}, [\![\cdot]\!]_{(\Sigma_0,\rho)}$ differ from frame to frame, but there is one salient frame – which thus can be taken as a reference frame –, namely the maximal frame Σ.

Fifth, for a well-behaved scenario s (where the interpretation corresponds to the minimal model of the program of the scenario) we have that $s \Vdash^+ BBp$ iff $s \Vdash^+ Bp$ iff $s \Vdash^+ p$ iff $P_s \models\!\!\approx p$. (This is the example of how to model belief in the logic programming framework given in section 3.2.2). Thus, one might wonder: Doesn't this just build in logical omniscience since agents believe exactly those sentences that are deducible? After all, if we were to say in a classical setting "an agent believes φ iff φ follows classically from a background theory", then this agent is

[107] It was developed in the 1950s by Anderson and Belnap—published in Anderson and Belnap (1975) and Anderson et al. (1992). For an overview, see Priest (2008, ch. 8).

logically omniscient. However, this worry doesn't apply here: The background logic of well-behaved scenarios is three-valued logic programming (resp., its neural implementation), and this logic is build to be cognitively adequate. This logic doesn't have (propositional) logical truths (so no sentences that an agent always needs to believe), and it is computationally tractable: If $P \mid\!\approx p$, then the agent can also do the reasoning that $P \mid\!\approx p$, so she comes to believe it. Deducibility in the logic, and computation in the neural network implementing the logic, is cognitively adequate computation of belief (also cf. Leitgeb 2001, 2003). In section 4.2.2, we will state this more precisely. Thus, the fact that in a well-behaved scenario belief coincides with truth (and deducibility) is not a surprising side-effect but deliberately wanted and built in.[108]

So we have a "construction based" notion of belief: belief is truth in the intended model constructed from one's representation of the world. This is quite different in spirit from the "quantifier based" notion of belief of standard epistemic logic: here belief is truth in all worlds that one considers to be (epistemically) possible.

Sixth, the ideas behind the semantic clauses for the counterfactual are the following. The two clauses (a) and (b) that make the counterfactual true correspond, respectively, to a "logical" tie between the antecedence and the consequence and to a "worldly or conceptual" tie. If a counterfactual is true because clause (a) is met, then there is a "logical" reason (which, in our multi-valued setting, is much stronger than in a classical setting) that "causes" the consequence to be true if the antecedence is true. If a counterfactual is true because clause (b) is met, then there is a (deep) conceptual reason present in the current scenario that "causes" in all relevantly similar scenarios the consequence to be true if the antecedence is true.

Note that these two clauses reflect the two clauses for the truth of a counterfactual given by Lewis (1973, p. 16). However, our first clause is different because we don't allow that all counterpossibles – that is, counterfactuals with impossible antecedents – are vacuously true. For counterpossibles to be true we demand that there still is a strong logical tie between the impossible antecedence and the consequence (just how strong will soon become clear when further investigate the kind of truth-preservation demanded in this case). For example, the counterpossible "$p \wedge \neg p \boxrightarrow p$" is valid, but the counterpossible "$p \wedge \neg p \boxrightarrow q$" is not.

There is a huge debate on whether or not counterpossibles should be vacuously true,[109] and we cannot enter it here for reasons of space. We restrict us to two comments. First, in our framework we also could have given the counterfactual a semantics that renders all counterpossibles vacuously true (just replace clause (a) by "no well-behaved scenario makes φ true"). And second, with our semantics we wish to capture certain aspects of counterfactuals that are relevant in the discussion of synonymy, and for those aspects a semantics on which not all counterpossibles are vacuously true is preferable. (For the use of counterfactuals in metaphysics this might

[108] Of course, we allow arbitrarily complex sentences to be embedded behind a belief operator, so checking whether they are true under the given interpretation of the scenario can also become arbitrarily demanding. However, this idealization may be excused since, in practice, sentences after a belief operator usually are of very low complexity.

[109] That counterpossibles are vacuously true is built into the original semantics for counterfactuals by Stalnaker (1968) and Lewis (1973). Defendants of a vacuous treatment of counterpossibles include Lewis, Stalnaker, and Williamson (2016). Critics include Nolan (1997), Krakauer (2012), Brogaard and Salerno (2013), Kment (2014), and Bjerring (2014).

be different.)

Finally, note that the "philosophical meaning" of the counterfactual depends on the philosophical interpretation of the clause extension ordering ρ which provides the metric. If ρ captures the similarity of scenarios, we get the prototypical counterfactual. And if ρ captures the subjective plausibility of scenarios, the counterfactual gets closer to a notion of conditional belief—indicating a link to the fields of belief revision and dynamic epistemic logic. Again, if space would permit, much more could be said here.

We get back to more properties of the counterfactual in section 4.2.3 below.

4.2.2 Describing the notions of validity

In this section, we describe some of the properties of the notions of validity introduced above, and we show how they are interrelated. Our treatment is entirely semantic, only in section 5.2.1 of the next chapter we will introduce proof systems capturing some of the notions of validity.

We start by some straightforward facts about what truth-values sentences can receive in the semantics.

Lemma 4.2.3 (Truth- and falsemakers)**.** *(i) For propositional sentences $\varphi \in \mathcal{L}_p$: If all the atoms of φ have a classical value (resp. only value u, only value $\bot$) in a scenario s, then φ has a classical value (resp. value u, value $\bot$), too.*

(ii) In particular, every propositional sentence has a truth-maker (the scenario $s_\bot$ where every atom is $\bot$) and a false-maker (scenario $s_\bot$, too).

(iii) In a well-behaved scenario no propositional sentence has value $\bot$.

(iv) Deciding more sentences preserves $\bot$: For all $s, s' \in \Sigma_0$, if $\forall p : s \Vdash^+_{(\Sigma_0,\rho)} p \Rightarrow s' \Vdash^+_{(\Sigma_0,\rho)} p$ and $\forall p : s \Vdash^-_{(\Sigma_0,\rho)} p \Rightarrow s' \Vdash^-_{(\Sigma_0,\rho)} p$, then

$$\forall \varphi \in \mathcal{L}_p : s \Vdash^+_{(\Sigma_0,\rho)} \varphi \Rightarrow s' \Vdash^+_{(\Sigma_0,\rho)} \varphi, \text{ and}$$
$$\forall \varphi \in \mathcal{L}_p : s \Vdash^-_{(\Sigma_0,\rho)} \varphi \Rightarrow s' \Vdash^-_{(\Sigma_0,\rho)} \varphi,$$

and in particular: $\forall \varphi \in \mathcal{L}_p : T^s(\varphi) = \bot \Rightarrow T^{s'}(\varphi) = \bot$.

Proof of 4.2.4. Immediate (by induction). □

Next we see that in the four-valued case truth-preserving and order-preserving validity coincide.

Theorem 4.2.4 ($\models^{tr}$ iff $\models^{or}$)**.** *For propositional sentences (i.e. $\Gamma \subseteq \mathcal{L}_p$, $\varphi \in \mathcal{L}_p$) we have*

$$\Gamma \models^{tr}_{(\Sigma_0,\rho)} \varphi \overset{(i)}{\Leftrightarrow} \Gamma \models^{\bar{tr}}_{(\Sigma_0,\rho)} \varphi \overset{(ii)}{\Leftrightarrow} \Gamma \models^{or}_{(\Sigma_0,\rho)} \varphi$$

where $\bar{tr}$-validity has $\{1, u\}$ as designated values instead of $\{1, \bot\}$.

Proof of 4.2.4. Ad (i). We consider the following mapping $\bar{\cdot} : \Sigma_0 \to \Sigma_0$ where $s = (S_s, P_s, I_s, E_s)$

is mapped to $\bar{s} := (S_s, P_s, \bar{I}_s, \bar{E}_s)$ with

$$\bar{I}_s(p) := \begin{cases} (1,0) & \text{, if } T(I_s(p)) = t \\ (0,1) & \text{, if } T(I_s(p)) = f \\ (1,1) & \text{, if } T(I_s(p)) = u \\ (0,0) & \text{, if } T(I_s(p)) = \bot. \end{cases}$$

(So, essentially, $\bar{\cdot}$ mirrors the truth-value diamond on the vertical axis.) We define $\bar{E}_s$ analogously (but, in fact, this is not needed as we only look at propositional sentences).

Then we show that for all propositional φ and all $s \in \Sigma_0$

$$\begin{aligned} T^s_{(\Sigma,\rho)}(\varphi) = u \quad &\text{iff} \quad T^{\bar{s}}_{(\Sigma,\rho)}(\varphi) = \bot \\ T^s_{(\Sigma,\rho)}(\varphi) = \bot \quad &\text{iff} \quad T^{\bar{s}}_{(\Sigma,\rho)}(\varphi) = u \\ T^s_{(\Sigma,\rho)}(\varphi) = 0 \quad &\text{iff} \quad T^{\bar{s}}_{(\Sigma,\rho)}(\varphi) = 0 \\ T^s_{(\Sigma,\rho)}(\varphi) = 1 \quad &\text{iff} \quad T^{\bar{s}}_{(\Sigma,\rho)}(\varphi) = 1 \end{aligned} \tag{4.1}$$

This can be seen by a straightforward induction on φ.

Now we can show (i). If $\Gamma \not\models^{tr}_{(\Sigma_0,\rho)} \varphi$, then there is a $s \in \Sigma_0$ such that $T^s_{(\Sigma,\rho)}(\Gamma) \in \{1, \bot\}$ but $T^s_{(\Sigma,\rho)}(\varphi) \notin \{1, \bot\}$. Then, by (4.1), $T^{\bar{s}}_{(\Sigma,\rho)}(\Gamma) \in \{1, u\}$ but $T^s_{(\Sigma,\rho)}(\varphi) \notin \{1, u\}$. Hence $\Gamma \not\models^{\bar{tr}}_{(\Sigma_0,\rho)} \varphi$. We show the other direction analogously.

Ad (ii). We first show that

$$\forall a, b \in \{t, f, u, \bot\} : a \preceq b \quad \text{iff} \quad \begin{cases} a \in \{u, 1\} \Rightarrow b \in \{u, 1\} & \text{, and} \\ a \in \{\bot, 1\} \Rightarrow b \in \{\bot, 1\} & . \end{cases} \tag{4.2}$$

Indeed, the left-to-right direction is immediate, and for the other direction we consider all the seven cases where $a \not\preceq b$ and observe that in each case we find that either $a \in \{\bot, 1\}$ but $b \notin \{\bot, 1\}$ or $a \in \{u, 1\}$ but $b \notin \{u, 1\}$. Then we have

$$\begin{aligned} \Gamma \models^{or}_{(\Sigma_0,\rho)} \varphi &\Leftrightarrow \forall s \in \Sigma_0 : T^s_{(\Sigma_0,\rho)}(\Gamma) \preceq T^s_{(\Sigma_0,\rho)}(\varphi) \\ &\overset{(4.2)}{\Leftrightarrow} \forall s \in \Sigma_0 : \begin{cases} T^s_{(\Sigma_0,\rho)}(\Gamma) \in \{u, 1\} \Rightarrow T^s_{(\Sigma_0,\rho)}(\varphi) \in \{u, 1\} & \text{, and} \\ T^s_{(\Sigma_0,\rho)}(\Gamma) \in \{\bot, 1\} \Rightarrow T^s_{(\Sigma_0,\rho)}(\varphi) \in \{\bot, 1\} \end{cases} \\ &\Leftrightarrow \Gamma \models^{tr^*}_{(\Sigma_0,\rho)} \varphi \text{ and } \Gamma \models^{tr}_{(\Sigma_0,\rho)} \varphi \\ &\overset{(i)}{\Leftrightarrow} \Gamma \models^{tr}_{(\Sigma_0,\rho)} \varphi, \end{aligned}$$

as wanted.[110, 111] □

Now we can show how the various notions of validity are interrelated—both in the four-valued and in the three-valued setting.

Proposition 4.2.5 (Relation between the notions of validity)**.** *If we consider all scenarios (both well-behaved and not well-behaved ones) we have the following.*

[110] Chemla et al. (2016, theorem 2.11) observe that in a general multi-valued setting, order-preserving validity is the conjunction of all truth-preservation validity relations whose set of designated values is an upset in the truth-value lattice that moreover contains its infimum. We have shown here that in the four-valued case something stronger is the case: First, it is enough to only consider the sets of designated values $\{1, \bot\}$ and $\{1, u\}$ (4.2), so in particular the set $\{1\}$ doesn't need to be considered. Second, the conjunction trivializes since both conjuncts are the equivalent.

[111] After I proved this, I noticed that those results are also shown in Font (1997, prop. 2.2 and 2.5). (I would have been surprised if they weren't already known.) However, we think that our proof is more elementary as it doesn't require the machinery of algebraic logic used in Font (1997).

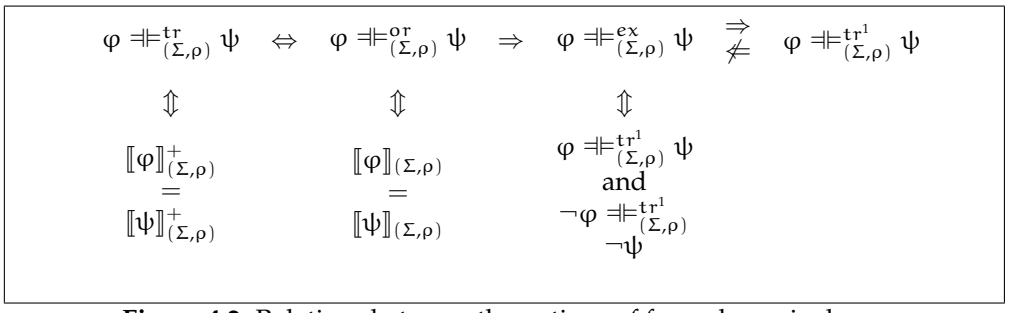


Figure 4.2: Relations between the notions of formula equivalence

(i) For propositional sentences (i.e. $\Gamma \subseteq \mathcal{L}_p$, $\varphi \in \mathcal{L}_p$) we have

$$\Gamma \models^{\mathtt{ex}}_{(\Sigma,\rho)} \varphi \not\Leftrightarrow \Gamma \models^{\mathtt{or}}_{(\Sigma_0,\rho)} \varphi \Leftrightarrow \Gamma \models^{\mathtt{tr}}_{(\Sigma_0,\rho)} \varphi.$$

(ii) If we consider all sentences (i.e. any Γ and φ) we have

$$\begin{array}{ccc} \Gamma \models^{\mathtt{ex}}_{(\Sigma,\rho)} \varphi & \not\Leftrightarrow & \Gamma \models^{\mathtt{or}}_{(\Sigma,\rho)} \varphi \\ & \searrow \quad \swarrow & \\ & \Gamma \models^{\mathtt{tr}}_{(\Sigma,\rho)} \varphi & \end{array}$$

where the (uncrossed) implications in fact hold for any $\Sigma_0 \subseteq \Sigma$.

(iii) For formula-validity of $\mathcal{L}$-sentences φ we have:

$$\models^{\mathtt{ex}}_{(\Sigma_0,\rho)} \varphi \quad \textit{iff} \quad \models^{\mathtt{or}}_{(\Sigma_0,\rho)} \varphi \quad \textit{iff} \quad \forall s \in \Sigma_0 : \mathsf{T}^s_{(\Sigma_0,\rho)}(\varphi) = 1.$$

(iv) For formula-equivalence – that is, $\varphi \models^*_{(\Sigma_0,\rho)} \psi$ and $\varphi \models^*_{(\Sigma_0,\rho)} \psi$ (where $* \in \{\mathtt{tr}, \mathtt{or}, \mathtt{ex}\}$) – we write $\varphi \dashv\models^*_{(\Sigma_0,\rho)} \psi$. Then we have for all $\varphi, \psi \in \mathcal{L}_p$ the (non-) implications shown in figure 4.2.5.

That is, for formula-equivalence the notions of $\mathtt{tr}$-validity, $\mathtt{or}$-validity, content-identity, and positive-content-identity collapse.[112]

If we restrict us to well behaved scenarios *we have the following.*

(v) The more general picture of (ii) already occurs for propositional sentences (so (i) isn't true anymore).

(vi) We get (iii) also with $\models^{\mathtt{tr}} \varphi$ as part of the equivalences.

(vii) The picture of (iv) now changes to the one given in figure 4.2.5 (still only for propositional sentences).

[112] For reasons of time and since we won't need it below, we leave open the (interesting) question whether also $\mathtt{ex} \Rightarrow \mathtt{or}$ holds. This amounts to asking whether or not there are formulas φ and ψ that agree on their classical value under any interpretation but that there also is a scenario in which one is $\bot$ while the other is $\mathfrak{u}$. Algebraically speaking, the question is the following: Are there two elements φ and ψ of the free algebra of propositional sentences such that (i) for all homomorphisms $\mathfrak{h}$ into the De Morgan algebra of the four truth-values, $\mathfrak{h}(\varphi) = 1$ iff $\mathfrak{h}(\psi) = 1$ and the same for 0, and (ii) there is a homomorphism $\mathfrak{h}$ such that $\mathfrak{h}(\varphi) = \mathfrak{u}$ and $\mathfrak{h}(\psi) = \bot$?

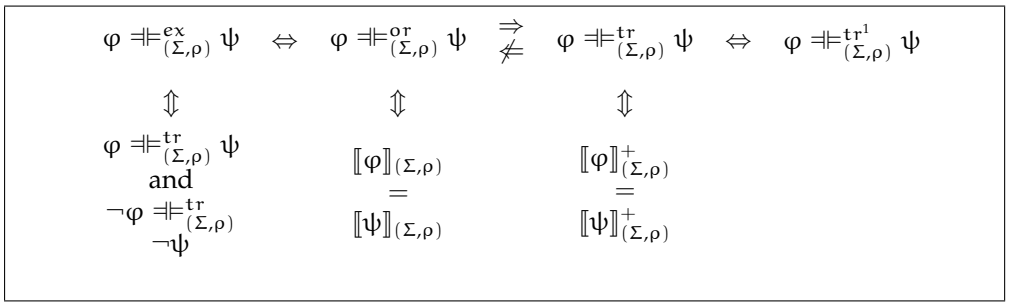


Figure 4.3: Formula equivalence restricted to well-behaved scenarios

Proof of 4.2.5. Fix the sentences

$$\varphi_0 := p \wedge \neg p \qquad \psi_0 := (p \wedge \neg p) \wedge (q \wedge \neg q)$$
$$\varphi_1 := p \qquad \psi_1 := p \vee q$$
$$\varphi_2 := B(p \wedge \neg p) \qquad \psi_2 := q.$$

Ad (i). The equivalence is proposition 4.2.4, and for the inequivalences observe that $\varphi_0 \models^{ex}_{(\Sigma_0)} \psi_0$ but $\varphi_0 \not\models^{or}_{(\Sigma_0)} \psi_0$. And $\varphi_1 \models^{or}_{(\Sigma_0)} \psi_1$ but $\varphi_1 \not\models^{ex}_{(\Sigma_0)} \psi_1$.

Ad (ii). The (uncrossed) implications are immediate from the definitions. The top two inequalities are (i). For the others observe that $\varphi_2 \models^{tr}_{(\Sigma_0)} \psi_2$ (since the belief operator takes one to a well-behaved scenario and there $p \wedge \neg p$ can never be true). However, both $\varphi_2 \not\models^{or}_{(\Sigma_0)} \psi_2$ and $\varphi_2 \not\models^{ex}_{(\Sigma_0)} \psi_2$.

Ad (iii). Immediate from $T^s_{(\Sigma_0,\rho)}(\emptyset) = 1$.

Ad (iv). The vertical equivalences are immediate. The leftmost horizontal equivalence is proposition 4.2.4.

Concerning "or $\Rightarrow$ ex": If $T^s(\varphi) = 1$, then $1 = T^s(\varphi) \leqslant T^s(\psi)$, so $T^s(\psi) = 1$. And if $T^s(\varphi) = 0$, then $T^s(\psi) \leqslant T^s(\varphi) = 0$, so $T^s(\psi) = 0$. Hence $\varphi \models^{ex} \psi$. The other direction is analogous.

The rightmost implication is immediate, and the a counterexample to the crossed arrow is $\varphi := p \wedge \neg p$ and $\psi := q \wedge \neg q$.

Ad (v). The top inequalities still persist. The (uncrossed) implications are immediate, and their reverses now fail because: $p \wedge \neg p \models^{tr}_{\Sigma,\rho} q$ but $p \wedge \neg p \not\models^{or}_{\Sigma,\rho} q$ and $p \wedge \neg p \not\models^{ex}_{\Sigma,\rho} q$.

Ad (vi). All statements are immediately equivalent to the last one.

Ad (vii). Concerning tr $\not\Rightarrow$ or: $p \wedge \neg p$ is tr-equivalent to $q \wedge \neg q$ but they are not or-equivalent.

Concerning ex $\Rightarrow$ or: If $\varphi \dashv\models^{ex}_{(\Sigma,\rho)} \psi$, then if one sentence has value 1, the other must have value 1, too, and if one sentence has value 0, the other must have value 0, and hence if one sentence has value u, the other must have value u. Hence $\varphi \dashv\models^{or}_{(\Sigma,\rho)} \psi$.

All other implications are immediate. □

Now that the interrelations of the notions of validity have been settled, we look at their individual properties.

Proposition 4.2.6 (Validities)**.** *(i) Both in the three- and four-valued setting, there are no $\mathcal{L}_p$-validities: That is, neither in the three-valued nor in the four-valued setting there is a $\mathcal{L}_p$-sentence φ such that $\models^{*}_{\Sigma,\rho} \varphi$ (for any $* \in \{tr, or, ex\}$).*

(ii) This is not the case anymore for all $\mathcal{L}$-sentences. For example, $\models^{tr}_{(\Sigma,\rho)} p \boxrightarrow p$.

(iii) Both in the three- and four-valued setting, the notions of validity are reflexive: That is, for $* \in \{tr, or, cl\}$ we have $\varphi \models^{*}_{\Sigma,\rho} \varphi$.

(iv) *Both in the three- and four-valued setting,* tr- *and* or*-validity is transitive: That is, for* $* \in \{\mathtt{tr}, \mathtt{or}\}$ *we have (for all* $\Gamma, \varphi, \psi, \chi, \Sigma_0, \rho$*)*

$$\text{If } \Gamma, \varphi \models^*_{(\Sigma_0,\rho)} \psi \text{ and } \Gamma, \psi \models^*_{(\Sigma_0,\rho)} \chi, \text{ then } \Gamma, \varphi \models^*_{(\Sigma_0,\rho)} \chi. \tag{4.3}$$

(v) *In the three-valued case* ex*-validity is not necessarily transitive (i.e. for* $* = \mathtt{ex}$ *(4.3) doesn't hold in general).*

Proof of 4.2.6. Ad (i). The empty scenario doesn't make any $\mathcal{L}_\mathrm{p}$-sentence true. Ad (ii)–(iv). Immediate. Ad (v). Take $\Gamma := \{q \vee \neg q\}$, $\varphi := p \wedge \neg p$, $\psi := p$, $\chi := p \wedge (q \vee \neg q)$. □

Finally, we consider to what degree these notions of validity meet the two main characteristica of relevant logics: The no-explosion condition and the relevance condition. The former says that a contradiction doesn't entail everything, and the second says that entailment has to preserve a link in subject matter between premisses and conclusion.

Proposition 4.2.7 (No-explosion condition)**.** *For* $* \in \{\mathtt{tr}, \mathtt{or}, \mathtt{ex}, \mathtt{tr}^1\}$*, we say* $*$*-validity satisfies the* no-explosion condition *if the following holds: If there is no truthmaker for* Γ*, it doesn't follow that* $\Gamma \models^*_{(\Sigma,\rho)} \varphi$ *for all* $\mathcal{L}$*-sentences* φ*.*

(i) *In the three-valued setting:* or*- and* ex*-validity satisfy the no-explosion condition, but* $\mathtt{tr}$ *(*$= \mathtt{tr}^1$*)-validity does not.*

(ii) *In the four-valued setting:* tr*-,* or*-, and* ex*-validity satisfy the no-explosion condition, but* $\mathtt{tr}^1$*-validity does not.*

Proof of 4.2.7. As counterexample take $\Gamma := \{p \wedge \neg p\}$ (which doesn't have a truthmaker) and $\varphi := q$. □

Proposition 4.2.8 (Relevance condition)**.** *We say that a binary relation* R *between (sets of)* $\mathcal{L}$*- or* $\mathcal{L}_\mathrm{p}$*-sentences meets the* relevance condition*, if being in that relation implies that the sentences (in the two sets) share at least one atomic sentence (i.e.* $\Gamma \mathrm{R} \Delta$ *implies* $\mathtt{At}(\Gamma) \cap \mathtt{At}(\Delta) \neq \emptyset$ *where* Γ *and* Δ *are either sentences or sets of sentences). We have for propositional sentences in the three-valued setting*

(i) *Neither* tr*-validity nor* or*-validity have the relevance condition.*

(ii) *Both* cl*-validity and content identity (*$[\![\cdot]\!] = [\![\cdot]\!]$*) have the relevance condition.*

And for propositional sentences in the four-valued setting tr*-,* or*-,* cl*-validity, and content identity have the relevance condition, but* $\mathtt{tr}^1$*-validity does not.*

Proof of 4.2.8. For the counterexamples (showing that the given relation doesn't have the relevance condition) use the sentences $p \wedge \neg p$ and $q \vee \neg q$.

To show that a given relation R has the relevance condition, assume for contradiction that not and get $\Gamma \mathrm{R} \Delta$ and two disjoint sets $\mathtt{At}(\Gamma)$ and $\mathtt{At}(\Delta)$ of atomic sentences. Then consider a scenario that makes all sentences in $\mathtt{At}(\Gamma)$ true (or $\bot$) and leaves those in $\mathtt{At}(\Delta)$ undecided. Then Γ is classical (or $\bot$) while Δ is undecided—in contradiction to $\Gamma \mathrm{R} \Delta$. □

Here is a way to look at the three notions of validity in the three-valued setting: Mere truth-preservation is too weak of a tie between premisses and conclusion. The two stronger ones are order-preservation and exclusion-preservation, but they have different, mutually exclusive features: or-validity captures content-identity but

doesn't have the relevance condition, and *ex*-validity has the relevance condition but doesn't capture content-identity. Both in the three- and four-valued setting, the – to the best of my knowledge – novel *ex*-validity provides a natural and interesting notion of validity.[113]

4.2.3 The counterfactual in the semantics

In this section, we discuss some of the features of the counterfactual of our semantics. First, we discuss the so-called limit assumption, and second, we see that our counterfactual satisfies the principles that a "good" counterfactual should satisfy.

In our semantics for the counterfactual, we don't build in the notorious *limit assumption*: the assumption that for all scenarios and all antecedents that can be made true, there is a smallest ϵ-sphere in which the antecedence is true (cf. Lewis 1973, pp. 19-21). Again, our framework is compatible with this assumption, but it is not built in. To be more precise, we say a clause-extension ordering ρ has the limit assumption, if for all scenarios s and for all φ, if there is a well-behaved scenario making φ true, then there is a smallest $\epsilon \geqslant 0$ such that there is a well-behaved scenario in $B^{\rho}_{\epsilon}(s)$ that makes φ true.

Proposition 4.2.9 (Limit assumption)**.** *(i) Under the limit assumption the semantic clause for the counterfactual simplifies:*

$s \Vdash^{+}_{(\Sigma_0,\rho)} \varphi \boxdot\!\!\rightarrow \psi$ *iff either* $\varphi \models^{\mathrm{tr}}_{(\Sigma_0,\rho)} \psi$ *or* φ *is true at a well-behaved scenario and in the* ϵ*-smallest* $B^{\rho}_{\epsilon}(s)$ *containing a well-behaved* φ*-scenario, all well-behaved* φ*-scenarios are also* ψ*-scenarios.*

$s \Vdash^{-}_{(\Sigma_0,\rho)} \varphi \boxdot\!\!\rightarrow \psi$ *iff* φ *is true at a well-behaved scenario and the* ϵ*-smallest* $B^{\rho}_{\epsilon}(s)$ *contains a well-behaved* φ*-scenario that makes* ψ *false.*

(ii) ρ *does* not *have the limit assumption if and only if there is a* φ *and a sequence of well-behaved scenarios* $(s_i)_{i\in\mathbb{N}_{\geqslant 0}}$ *each making* φ *true such that the sequence* $(d_\rho(s_0, s_i))_{i\geqslant 1}$ *is strictly decreasing.*

(iii) In particular, if ρ *only assigns sequences of natural numbers, then it has the limit assumption.*

Proof of 4.2.9. Ad (i). This is already observed by Lewis (1973, p. 20) and straightforwardly carries over to our setting.

Ad (ii). If ρ doesn't have the limit assumption, then there is a scenario s_0 and a φ such that there is an $\epsilon \geqslant 0$ with $B^{\rho}_{\epsilon}(s_0)$ containing a well-behaved φ-scenario, but there is no smallest such ϵ. Hence there is a strictly decreasing sequence $(\epsilon_i)_{i\geqslant 1}$ with

$$\emptyset \neq B^{\rho}_{\epsilon_1}(s_0) \supsetneq B^{\rho}_{\epsilon_2}(s_0) \supsetneq B^{\rho}_{\epsilon_3}(s_0) \supsetneq \ldots,$$

where each $B^{\rho}_{\epsilon_i}(s_0)$ contains a well-behaved scenario s_i making φ true that is not contained in any of the subsequent $B^{\rho}_{\epsilon_j}(s_0)$ $(j > i)$. Thus, we have a sequence $(s_i)_{i\geqslant 0}$ of well-

[113] Of course, more can be investigated. For example, we can ask about compactness: Which properties of the class of scenarios Σ are already reflected on finite frames $\Sigma_0 \subseteq \Sigma$? Or we can ask about decidability: Which of the notions of validity are decidable, and if so, what is their complexity? Hähnle (1993, p. 270) states strong Kleene logic to be NP-complete, and Bimbó (2007, p. 730) states first-degree entailment to be co-NP-complete. Also note that truth of an atomic sentence at a well-behaved scenario is tractably decidable (cf. Dantsin et al. 2001, thm. 5.5). And, see our section 5.3.1 for an axiomatization of content identity. However, for reasons of space, we cannot pursue these questions here.

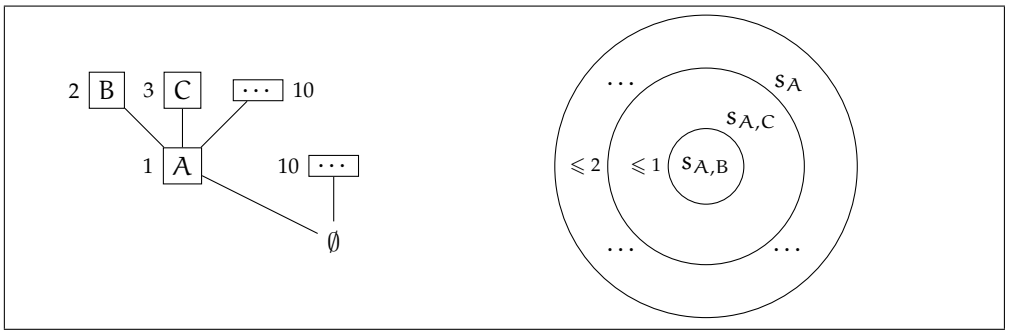


Figure 4.4: A clause extension ordering (left) and the resulting sphere model (right)

behaved scenarios each making φ true and such that the sequence $(d_\rho(s_0, s_i))_{i\geqslant 1}$ is strictly decreasing. For the other direction take $\epsilon_i := d_\rho(s_0, s_i)$.

Ad (iii). If ρ only assigns sequences of natural numbers, then $\mu^*_\rho(s)$ is a natural number for every scenario s, and hence the distance d_ρ between any two scenarios is a natural number, and hence no sequence $(d_\rho(s_0, s_i))_{i\geqslant 1}$ can be strictly decreasing. □

Let's see that our counterfactual is worth its name—that is, that it validates those axioms that are commonly taken to govern any "good" notion of a counterfactual, and that it invalidates those principles that a good notion should invalidate.

Proposition 4.2.10 (Counterfactuals)**.** *(i) For all frames (Σ_0, ρ) and for all $\varphi \in \mathcal{L}$ we have $\models^{tr}_{(\Sigma_0,\rho)} \varphi \mathrel{\Box\!\!\rightarrow} \varphi$.*

(ii) For all frames (Σ_0, ρ) and for all $\varphi, \psi \in \mathcal{L}$ we have $\varphi, \varphi \mathrel{\Box\!\!\rightarrow} \psi \models^{tr}_{(\Sigma_0,\rho)} \psi$.

(iii) There are frames (Σ_0, ρ) and $\varphi, \psi, \chi \in \mathcal{L}$ such that $\varphi \mathrel{\Box\!\!\rightarrow} \psi \not\models^{tr}_{(\Sigma_0,\rho)} \varphi \wedge \chi \mathrel{\Box\!\!\rightarrow} \psi$.

(iv) There are frames (Σ_0, ρ) and $\varphi, \psi, \chi \in \mathcal{L}$ such that $\varphi \mathrel{\Box\!\!\rightarrow} \psi, \psi \mathrel{\Box\!\!\rightarrow} \chi \not\models^{tr}_{(\Sigma_0,\rho)} \varphi \mathrel{\Box\!\!\rightarrow} \chi$.

(v) There are frames (Σ_0, ρ) and $\varphi, \psi \in \mathcal{L}$ such that $\varphi \mathrel{\Box\!\!\rightarrow} \psi \not\models^{tr}_{(\Sigma_0,\rho)} \neg\psi \mathrel{\Box\!\!\rightarrow} \neg\varphi$.

Proof of 4.2.10. Ad (i). We trivially have that for all $s \in \Sigma_0$, if $s \Vdash^+_{(\Sigma_0,\rho)} \varphi$, then $s \Vdash^+_{(\Sigma_0,\rho)} \varphi$. Hence by clause (a) of the counterfactual, we have for all $s \in \Sigma_0$ that $s \Vdash^+_{(\Sigma_0,\rho)} \varphi \mathrel{\Box\!\!\rightarrow} \varphi$, and the claim follows.

Ad (ii). Let $s \in \Sigma_0$ such that $s \Vdash^+_{(\Sigma_0,\rho)} \varphi$ and $s \Vdash^+_{(\Sigma_0,\rho)} \varphi \mathrel{\Box\!\!\rightarrow} \psi$. We have to show $s \Vdash^+_{(\Sigma_0,\rho)} \psi$. If the counterfactual is true because of clause (a), we immediately have $s \Vdash^+_{(\Sigma_0,\rho)} \psi$, so let the counterfactual be true because of clause (b). Then there particularly is an $\epsilon \geqslant 0$ such that all φ-scenarios in ϵ-distance from s are also ψ-scenarios. Since s itself is a φ-scenario and $d_\rho(s, s) = 0 \leqslant \epsilon$, s is also a ψ-scenario, that is, $s \Vdash^+_{(\Sigma_0,\rho)} \psi$.

Ad (iii). Take the sentences p, q, r and the frame (Σ, ρ) where ρ is given in figure 4.4 on the left where all scenarios not shown receive weight 10 (apart from those required to make ρ symmetric), and we choose the clauses A, B, C to be, respectively, p, q, r. The corresponding sphere system around the scenario containing only the facts p and q is shown on the right (the scenarios consist of the clauses indicated in their subscripts).

We claim that $s_{p,q} \Vdash^+_{(\Sigma,\rho)} p \mathrel{\Box\!\!\rightarrow} q$ but $s_{p,q} \nVdash^+_{(\Sigma,\rho)} p \wedge r \mathrel{\Box\!\!\rightarrow} q$. By lemma 4.2.9 (iii), ρ has the limit assumption, so to the check the truth of a counterfactual with satisfiable antecedent it is enough to look at the closest sphere containing an antecedent-scenario (cf. part (i) of the lemma).

Indeed, the sphere closest to $s_{p,q}$ containing a p-scenario is $\{s_{p,q}\}$, and there all (well-behaved) p-scenarios are q-scenarios. Thus, $s_{p,q} \Vdash^{+}_{(\Sigma,\rho)} p \boxdot\!\!\rightarrow q$. The sphere closest to $s_{p,q}$ containing a $p \wedge r$-scenario is $\{s_{p,r}, s_{p,q}\}$, but $s_{p,r}$ doesn't make the consequence q true. Thus, $s_{p,q} \nVdash^{+}_{(\Sigma,\rho)} p \wedge r \boxdot\!\!\rightarrow q$.

Ad (iv). Take the sentences p, q, r and the frame (Σ, ρ) where ρ is again given in figure 4.4, now with A, B, C being, respectively, q, r, p.

Then we have, again by virtue of the limit assumption, that $s_{q,r} \Vdash^{+}_{(\Sigma,\rho)} p \boxdot\!\!\rightarrow q$ (the closest p-scenario – which is $s_{A,C}$ – is a q-scenario), and $s_{q,r} \Vdash^{+}_{(\Sigma,\rho)} q \boxdot\!\!\rightarrow r$ (the closest q-scenario – which is $s_{A,B}$ – is a r-scenario), but $s_{q,r} \nVdash^{+}_{(\Sigma,\rho)} p \boxdot\!\!\rightarrow r$ (the closest p-scenario – which is $s_{A,C}$ – is not a r-scenario).

Ad (v). Take the sentences p, q and the frame (Σ, ρ) where ρ is again given in figure 4.4, now with A, B, C being, respectively, $p, q, \bot \rightarrow q$. There we have, again by the limit assumption, $s_{p,q} \Vdash^{+}_{(\Sigma,\rho)} p \boxdot\!\!\rightarrow q$ (the closest p-scenario – which is $s_{A,B}$ – also is a q-scenario) but not $s_{p,q} \nVdash^{+}_{(\Sigma,\rho)} \neg q \boxdot\!\!\rightarrow \neg p$ (the closest $\neg$q-scenario – which is $s_{A,C}$ – is not a $\neg$p-scenario). □

For reasons of space, we have to leave open the following question: Is there a way to connect the logic programming clauses "$p \wedge \neg ab \rightarrow q$" with the counterfactual $p \boxdot\!\!\rightarrow q$? For example, is there a ρ such that for all $s \in \Sigma$, $p \wedge \neg ab \rightarrow q \in P_s$ iff $s \Vdash^{+}_{(\Sigma,\rho)} p \boxdot\!\!\rightarrow q$?

5th *Chapter*

NOTIONS OF SYNONYMY

In the last chapter, we have developed the framework of scenarios. In this chapter, we use it to describe and characterize various notions of synonymy. These notions range from "world based" ones that only take the current scenario into account, over those that take surrounding scenarios into account, to "logical" ones that only take logical relations between sentences into account.

We provide a logic where $\varphi \equiv \psi$ is derviable if and only if φ and ψ have the same hyperintensional content according to the scenario semantics.

We observe something paradoxical about the notion of content identity (or absolute synonymy). The following two principles are jointly inconsistent: (i) If no scenario distinguishes two sentences, they are identical in content, and (ii) content identity entails identity in subject matter. Content identity according to the scenario semantics satisfies (i) but not (ii). Looking for notions of synonymy that satisfy (ii), we find that the logic of analytic containment of Fine (2016a) provides a – though not the first – notion of content identity above scenario semantics that satisfies (ii). It is required to move from scenarios to sets of scenarios to get a semantics that individuates content according to analytic synonymy.

We conclude that our investigation leads us to a pluralistic conception of synonymy: This is not only because of the sheer number of independently well-motivated notions of synonymy, but also because of the many opposing features of synonymy that can only be reconciled by acknowledging a plurality of synonymy.

5.1 A zoo of notions of synonymy

In this section, we describe various notions of synonymy and how they can be modeled by the scenario framework.

Let's start by recalling the two notions we've already seen in chapter 3.

Cognitive-role synonymy Two atomic sentences p and q are cognitive-role synonymous in a scenario s if $p \wedge \neg ab \to q \in P_s$ and $q \wedge \neg ab \to p \in P_s$.

Reasoned-to synonymy Two atomic sentences p and q are reasoned-to synonymous in a scenario s if $P_s \setminus \{p, q\} \approx p \to q \wedge q \to p$.

Reasoned-to synonymy is transitive[114], but cognitive-role synonymy is not. How-

[114] It is easily checked that we have for all programs P, if $P \approx (p \to q) \wedge (q \to p)$ and $P \approx (q \to r) \wedge (r \to q)$, then $P \approx (p \to r) \wedge (r \to p)$.

ever, we saw that via learning the two notions converge to being equivalent: Being cognitive-role synonymous implies being reasoned-to synonymous. And if two sentences p and q are reasoned-to synonymous in a scenario, then we can extend the scenario – via learning – by the two rules $p \wedge \neg ab \to q$, $q \wedge \neg ab \to p$.

Both notions only describe synonymy in a given scenario without taking any other (close-by) scenarios into account. Thus, heuristically speaking, we can call them *maximally local*. Moreover, they only work for atomic sentences.

We will now consider notions of synonymy that work for all $\mathcal{L}$-sentences, and that either (i) take the whole class of scenarios into account (so, heuristically speaking, they are *global*) or (ii) at least take their surrounding scenarios into account (so, heuristically speaking, they are *extendedly local*).

The first (and most obvious) more general notion of synonymy is that of content identity:

Strict synonymy Two $\mathcal{L}$-sentences φ and ψ are strictly synonymous on (Σ_0, ρ) ($\varphi \approx_{str} \psi$) if they have the same content, that is, $[\![\varphi]\!]_{(\Sigma_0,\rho)} = [\![\psi]\!]_{(\Sigma_0,\rho)}$.

(A variant would be "well-behaved strict synonymy": On well-behaved scenarios the two sentences have exactly the same truth-value.)

We've already encountered several properties of this relation: Strict synonymy is transitive, it coincides with order-preserving equivalence, and strict synonymy meets the relevance condition (that is, if two $\mathcal{L}_p$-sentences are strictly synonymous, they share subject matter in the sense of sharing an atomic sentence).[115] Also, strict synonymy is the co-hyperintensionality relation for the language $\mathcal{L}$ (with $\Leftrightarrow$ chosen as or-equivalence). Moreover, strict synonymy is global: To determine that two sentences are strictly synonymous we need to consider the whole space of scenarios, that is, at *every* scenario the two sentences have to coincide in truth-value.

However, strict synonymy is a very fine-grained notion of synonymy: It reflects the idea mentioned in the introduction that for nearly every pair of sentences we can cook up a context where the two sentence come apart. Even for the sentences "Peter is a bachelor" and "Peter is an unmarried man" we can find contexts where they have different truth-values: For example, in an (imaginative) country where marriage comes in degrees; and in such a scenario the fact that marriage comes in degrees would trigger, say, an abnormality inhibiting the rule "If someone is a bachelor, he is an unmarried man". Thus, it is just as hard for two sentences to be strictly synonymous as it is to ensure that no context can be cooked up that makes the two sentences come apart.

On a scale of notions of synonymy from "entirely worldly and conceptual" to "entirely logical", we have so far seen the two extremes. Roughly, cognitive-role synonymy and reasoned-to synonymy ask: Given what I know about the world and how I conceptualize it, when are two atomic sentences synonymous? Note that it makes sense to restrict this question to atomic sentences: This notion of synonymy is concerned with the *worldly and conceptual* content of the sentences (which is provided by the logic program of the scenario). It is not concerned with the *purely logical*

[115] In section 5.2 below we'll observe that being strictly synonymous doesn't entail *identity* in subject matter: it could be that two sentences are strictly synonymous but that their respective sets of atoms are not identical (though they are not disjoint).

relationship between the sentences. Strict synonymy, on the other hand, asks: Is there some logical relationship between the sentences that makes them synonymous? Note that here it makes sense to consider logical complex sentences since – as just seen – no two atomic sentences can be strictly synonymous.

We'll now explore the middle-ground of these two extremes. After all, the notion(s) of synonymy that we usually employ – be it in everyday discourse or in philosophical debates – is more tolerant and coarse-grained than the notion of synonymy claiming that no context whatsoever can make the sentences come apart. However, especially for philosophical purposes, we also don't want to restrict us to atomic sentences and one single fixed scenario: For example, in thought experiments (or simply by thinking about what others might think) we also have to consider scenarios other than ours.

There are at least two ways to more tolerant notions of synonymy. The first – and global – way is to say that the exceptions (that is the scenarios where the sentences come apart) are negligible. This will be captured by exception-tolerant synonymy and similarity synonymy below. The second – and extendedly local – way is to say that in similar or close-enough scenarios there are no exceptions. This will be captured by stable synonymy and counterfactual synonymy.

We start with the first notion of synonymy of the global way to more tolerant notions of synonymy.

Exception-tolerant synonymy Two $\mathcal{L}$-sentences φ and ψ are exception-tolerantly synonymous on (Σ_0, ρ) ($\varphi \approx_{ex} \psi$) if the set of exceptions

$$\Delta_{(\Sigma_0,\rho)}(\varphi,\psi) := \left\{ s \in \Sigma_0 \mid \mathsf{T}^s_{(\Sigma_0,\rho)}(\varphi) \neq \mathsf{T}^s_{(\Sigma_0,\rho)}(\psi) \right\}$$

is negligible, that is, $\mu_\rho\big(\Delta_{(\Sigma_0,\rho)}(\varphi,\psi)\big) < \infty$ where μ_ρ is as defined in definition 4.1.5.

Exception-tolerant synonymy is transitive: If $\Delta_{(\Sigma_0,\rho)}(\varphi,\psi)$ and $\Delta_{(\Sigma_0,\rho)}(\psi,\chi)$ are negligible, then $\Delta_{(\Sigma_0,\rho)}(\varphi,\chi)$ is negligible because the class of negligible sets is an ideal and

$$\Delta_{(\Sigma_0,\rho)}(\varphi,\chi) \subseteq \Delta_{(\Sigma_0,\rho)}(\varphi,\psi) \cup \Delta_{(\Sigma_0,\rho)}(\psi,\chi).^{116}$$

This renders exception-tolerant synonymy an equivalence relation. So given two sentences φ and ψ, the cluster of sentences that are exception-tolerantly synonymous to φ is always either identical to or disjoint from the cluster corresponding to ψ (because these clusters are in fact equivalence classes).

However, in many cases where we call two sentences synonymous, we want to express that they are very similar in meaning. And such a meaning similarity relation should be a proper (i.e. non-transitive) similarity relation—that is, we want the synonymy clusters to overlap. One way to spell out such a notion is to allow the "weight" of the exceptions to only rise up to a certain degree $r \in \mathbb{R}_{\geqslant 0}$ (and not arbitrarily high)—this is the second version of the global way to more tolerant notions of synonymy.

[116] Proof: Assume $s \in \Delta_{(\Sigma_0,\rho)}(\varphi,\chi)$. So, without loss of generality, $s \Vdash^+_{(\Sigma_0,\rho)} \varphi$ and $s \nVdash^+_{(\Sigma_0,\rho)} \chi$ (the other cases are analogous). Now, either $s \Vdash^+_{(\Sigma_0,\rho)} \psi$ or $s \nVdash^+_{(\Sigma_0,\rho)} \psi$. If the former, $s \in \Delta_{(\Sigma_0,\rho)}(\psi,\chi)$, and if the latter, $s \in \Delta_{(\Sigma_0,\rho)}(\varphi,\psi)$.

Similarity synonymy Two $\mathcal{L}$-sentences φ and ψ are similarity synonymous on (Σ_0, ρ) up to degree $r \in \mathbb{R}_{\geqslant 0}$ ($\varphi \approx_{\mathtt{sim}(r)} \psi$) if $\mu_\rho\big(\Delta_{(\Sigma_0,\rho)}(\varphi,\psi)\big) \leqslant r$.

(We can think of r being fixed by the context of utterance.) This is, in general, indeed a non-transitive similarity relation. Leitgeb (2008) proves a fundamental result about meaning-similarity relations: It is impossible for them to satisfy a set of finitely many, prima facie plausible axioms. It is suggested that the culprits are the compositionality axioms demanding that meaning-similarity is preserved under negation, conjunction, and disjunction. For our meaning-similarity relation we find that this suggestion is indeed right: In general, there are sentences $\varphi, \varphi', \psi, \psi'$ such that $\varphi \approx_{\mathtt{sim}(r)} \psi$ and $\varphi' \approx_{\mathtt{sim}(r)} \psi'$, but $\varphi \wedge \varphi' \not\approx_{\mathtt{sim}(r)} \psi \wedge \psi'$.

This concludes the global way to more tolerant notions of synonymy. Now we'll look at the local way. It aims to strike a balance between the two extremes of being maximally global (considering all scenarios) and of being maximally local (considering only a single scenario)—that is, we look for notions of synonymy that are extendedly local.

The motivation for avoiding the globality extreme is to get a notion of synonymy such that it can feasibly and relevantly be checked whether two sentences are synonymous—if we have to check all scenarios we cannot ensure feasibility and relevance anymore. To be sure, that doesn't mean that global notions of synonymy are useless: To the contrary, they are very insightful for someone looking at the logic from outside, but for an agent who needs to establish the synonymy of two sentences it might be unnecessarily demanding or not relevant in her situation to take all scenarios into account. The motivation for avoiding the maximal-locality extreme is to get a notion of synonymy that has a certain stability: If two sentences are synonymous at a scenario, they still are synonymous in close-by scenarios. If such a stability weren't given, then being synonymous would be an utterly contingent and accidental fact about the current scenario, and it would not indicate a "deeper" tie between the two synonymous sentences.

The first extendedly local notion of synonymy is counterfactual synonymy.

Counterfactual synonymy Two $\mathcal{L}$-sentences φ and ψ are counterfactually synonymous on (Σ_0, ρ) in scenario $s \in \Sigma_0$ ($\varphi \approx_{\boxright} \psi$) if $s \Vdash^{+}_{(\Sigma_0,\rho)} \varphi \boxright \psi \wedge \psi \boxright \varphi$.

This notion of synonymy is transitive.[117] (So counterfactual equivalence is transitive, although the single counterfactual is not.) The underlying idea of counterfactual

[117] *Proof.* Assume $\varphi \boxright \psi$, $\psi \boxright \varphi$, $\psi \boxright \chi$, and $\chi \boxright \psi$ are true at a scenario s. Show that $\varphi \boxright \chi$ and $\chi \boxright \varphi$ are true at s, too.

We consider two cases. Case 1: The assumed counterfactuals are all true because of clause (b). Then for each of the four, we find $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ such that within the respective ϵ-balls there is a well-behaved scenario making the antecedent true and if the antecedent is true at a scenario in that ball, also the consequent is true. Thus, in the $\mathtt{min}(\epsilon_1, \epsilon_2)$-ball, φ and ψ are $\mathtt{tr}^1$-equivalent and there are well-behaved scenarios making both true. The same for the $\mathtt{min}(\epsilon_3, \epsilon_4)$-ball and ψ and χ. Thus, taking $\epsilon := \mathtt{min}(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ we get that φ and χ are $\mathtt{tr}^1$-equivalent and that there are well-behaved scenarios making both true, which implies the claim.

Case 2: One of the assumed counterfactuals doesn't satisfy clause (b) but (a). Without loss of generality we may assume $\varphi \boxright \psi$ satisfies (a) – i.e., $\varphi \models^{\mathtt{tr}} \psi$ – but not (b). (The other cases will be seen to be analogous.) Hence there is no well-behaved s' making φ true. But then $\psi \models^{\mathtt{tr}} \varphi$, because if $\psi \boxright \varphi$ were true because of (b), there would be an ϵ-ball containing a well-behaved scenario making ψ true and hence also φ. So φ and ψ are $\mathtt{tr}$-equivalent. Hence there also is no well-behaved s' making ψ true. But then $\psi \boxright \chi$ can only be true because (a) holds, so $\psi \models^{\mathtt{tr}} \chi$. With the same reasoning as above, hence also $\chi \models^{\mathtt{tr}} \psi$, so ψ and χ are $\mathtt{tr}$-equivalent. But since $\mathtt{tr}$-equivalence is transitive, φ and χ are $\mathtt{tr}$-equivalent, which implies the claim. □

synonymy is that we take two sentences as synonymous just in case it holds that if (we imagine that) one were true, also the other would be true, and vice versa.

The second extendedly local notion of synonymy is stable synonymy.

Stable synonymy For an $\epsilon \in \mathbb{R}_{\geqslant 0}$, two $\mathcal{L}$-sentences φ and ψ are ϵ-stably synonymous on (Σ_0, ρ) in scenario s ($\varphi \approx_{stab(\epsilon)} \psi$) if $[\![\varphi]\!]_{(\Sigma_0,\rho)} \cap B^{\rho}_{\epsilon}(s) = [\![\psi]\!]_{(\Sigma_0,\rho)} \cap B^{\rho}_{\epsilon}(s)$, where $B^{\rho}_{\epsilon}(s)$ is the ϵ-ball around s.

This notion of synonymy is immediately seen to be transitive, too. The underlying idea of stable synonymy is that the context of discourse in which we're in determines a range only within which scenarios are considered—any scenario outside that range is not considered to be relevant to the discourse. Stable synonymy then says that within this range the sentences have the same content.

For completeness, we already mention two notions of synonymy that we will introduce later when we have the relevant concepts available. They will be even more fine-grained than strict synonymy, and hence they also are located at the "purely logical" extreme.

Analytic synonymy See section 5.3.1 below.

Super strict synonymy See section 5.3.1 below.

In section 5.4, we will discuss how these notions of synonymy are related (see especially figure 5.4).

5.2 Characterizing content identity

In this section, we want to provide a logic for content identity in order to better understand when two sentences have the same content according to our semantics. That is, we want to provide a proof system in which $\varphi \equiv \psi$ is derivable if and only if φ and ψ have the same content in our semantics. At the end, we mention some related proof systems.

There are two ways to achieve this. Either we go an indirect route and take an existing proof system capturing validity and turn it into one capturing content identity. Or we go a direct route and build a system having $\equiv$ as a primitive. We first sketch the indirect route, then dismiss it, and finally pursue the direct route in detail.

According to the indirect route, we take an existing sound and complete proof system $\mathcal{C}$ for first degree entailment. (There are many such systems: see, for example, Priest (2008, ch. 8), or Font (1997).) Then we have for propositional Γ and φ that $\Gamma \vdash_{\mathcal{C}} \varphi$ (i.e. φ is derivable from Γ with the rules in $\mathcal{C}$) if and only if $\Gamma \models_{FDE} \varphi$ (i.e. φ is a consequence from Γ according to first-degree entailment) if and only if $\Gamma \models^{tr}_{(\Sigma,\rho)} \varphi$. Then we can take the proof system for content identity according to which $\varphi \equiv \psi$ is derivable if $\varphi \vdash_{\mathcal{C}} \psi$ and $\psi \vdash_{\mathcal{C}} \varphi$ (since content identity coincides with tr-equivalence according to proposition 4.2.5).

A problem with this approach is that $\equiv$ is not a primitive symbol in the system: Its axioms and rules don't (directly) describe content identity but rather validity, which runs counter our aim of using the axioms and rules to better understand content identity. This, of course, is not a knock-down argument against this approach, but it

(A1)	$\varphi \equiv \neg\neg\varphi$	(A7)	$(\varphi \vee \psi) \vee \chi \equiv \varphi \vee (\psi \vee \chi)$
(A2)	$\varphi \equiv \varphi \wedge \varphi$	(A8)	$\neg(\varphi \wedge \psi) \equiv (\neg\varphi \vee \neg\psi)$
(A3)	$\varphi \wedge \psi \equiv \psi \wedge \varphi$	(A9)	$\neg(\varphi \vee \psi) \equiv (\neg\varphi \wedge \neg\psi)$
(A4)	$(\varphi \wedge \psi) \wedge \chi \equiv \varphi \wedge (\psi \wedge \chi)$	(A10)	$\varphi \wedge (\psi \vee \chi) \equiv (\varphi \wedge \psi) \vee (\varphi \wedge \chi)$
(A5)	$\varphi \equiv \varphi \vee \varphi$	(A11)	$\varphi \vee (\psi \wedge \chi) \equiv (\varphi \vee \psi) \wedge (\varphi \vee \chi)$
(A6)	$\varphi \vee \psi \equiv \psi \vee \varphi$	(A12)	$\varphi \vee (\varphi \wedge \psi) \equiv \varphi$
(R1)	$\dfrac{\varphi \equiv \psi}{\psi \equiv \varphi}$	(R3)	$\dfrac{\varphi \equiv \psi}{\varphi \wedge \chi \equiv \psi \wedge \chi}$
(R2)	$\dfrac{\varphi \equiv \psi \qquad \psi \equiv \chi}{\varphi \equiv \chi}$	(R4)	$\dfrac{\varphi \equiv \psi}{\varphi \vee \chi \equiv \psi \vee \chi}$

Figure 5.1: The system of strict synonymy

suggests to first try a direct route—which we now do. And, looking ahead, this turns out to be fruitful because we thus will get one single axiom making the difference between other systems for other notions of content identity—which provides a solid starting point for a philosophical discussion taken up afterward (section 5.3).

5.2.1 A sound and complete logic for strict synonymy

In this section, we introduce a proof system and show that it is sound and complete with respect to content identity. In the next section, we (philosophically) discuss the system.

Definition 5.2.1 (System of strict synonymy $\mathcal{S}$)**.** The system $\mathcal{S}$ of *strict synonymy* is given as follows. The basic statements are of the form $\varphi \equiv \psi$ where φ and ψ are $\mathcal{L}_p$-sentences (i.e. sentences build from atomic sentences via the connectives $\neg, \wedge, \vee$). The axioms and rules of the system then describe when such statements may be derived from others (or are axioms). These axioms and rules are given in figure 5.2.1. We write $\vdash_{\mathcal{S}} \varphi \equiv \psi$ if $\varphi \equiv \psi$ can be derived in $\mathcal{S}$.

One reason why we have chosen this system is that it extends the system of analytic containment ($\mathcal{AC}$) of Fine (2016a): $\mathcal{AC}$ is $\mathcal{S}$ minus axiom (A12). This axiom will be the starting point for the discussion of strict synonymy in the next section.

Working towards the completeness theorem, we start with a useful observation about replacement rules in the system.

Lemma 5.2.2 (Replacement)**.** *The following replacement rule is admissible, that is, if its premise is $\mathcal{S}$-derivable, then the conclusion is $\mathcal{S}$-derivable.*

$$\frac{\varphi \equiv \psi}{\chi[\varphi] \equiv \chi[\psi]}\ (FR)$$

(When $\chi[\varphi]$ is a formula containing occurrences of φ, then $\chi[\psi]$ is the result of replacing the occurrences of φ by ψ.)

Proof of 5.2.2. The proof is similar to Fine (2016a, pp. 202-204) and we move it to the appendix (section B.1). □

Now we will show that every sentence is S-provably equivalent to one in a strong version of a disjunctive normal form, which we now specify.[118]

Definition 5.2.3 (Standard minimal disjunctive form)**.** A sentence φ is in *conjunctive form* if it is a conjunction of literals[119]. We write $L(\varphi)$ for the set of literals of the conjunctive form φ. A sentence ψ is in *disjunctive form* if it is a disjunction of conjunctive forms.

We fix a bijective enumeration e of the atomic sentences, and we say that a conjunctive form is *standard* if its atomic sentences occur from left to right in e-increasing order (without repeats). We fix a bijective enumeration f of standard conjunctive forms, and we say that a disjunctive form is *standard* if its disjuncts are standard and f-increasing from left to right (without repeats).

We say a disjunctive form $\varphi = \varphi_1 \vee \ldots \vee \varphi_n$ is *minimal* if

$$\forall i \neq j : L(\varphi_i) \not\subseteq L(\varphi_j) \text{ and } L(\varphi_j) \not\subseteq L(\varphi_i).$$

Finally, we write $C(\varphi) := \{L(\varphi_1), \ldots, L(\varphi_n)\}$.

Lemma 5.2.4 (Normal form)**.** *For each formula φ there is a formula φ_{NF} in standard minimal disjunctive form (or short* normal form*) such that $\vdash_S \varphi \equiv \varphi_{NF}$. (In corollary 5.2.7 below we will see that this φ_{NF} in fact even is unique.)*

Proof of 5.2.4. In the standard way, we find a formula φ' provably equivalent to φ and in disjunctive form by using DeMorgan, double negation, distributivity, positive replacement, symmetry, and transitivity (cf. Fine 2016a, theorem 15, p. 214).

Next, we can delete – while preserving provability – any disjunct φ_j occurring in φ' if there already is a disjunct φ_i in φ' with $L(\varphi_i) \subseteq L(\varphi_j)$. This is because if there are such φ_j and φ_i, then, without loss of generality, $\varphi_j = \varphi_i \wedge \chi$ and (using the underlining to increase readability)

$$\begin{aligned}
\varphi' &= \varphi_1 \vee \ldots \vee \underline{\varphi_i} \vee \ldots \vee \underline{\varphi_j} \vee \ldots \vee \varphi_n \\
&= \varphi_1 \vee \ldots \vee \varphi_{i-1} \vee \underline{\varphi_i} \vee \varphi_{i+1} \vee \ldots \vee \varphi_{j-1} \vee \underline{(\varphi_i \wedge \chi)} \vee \varphi_{j+1} \vee \ldots \vee \varphi_n \\
&\equiv \underline{\varphi_i \vee (\varphi_i \wedge \chi)} \vee \varphi_1 \vee \ldots \vee \varphi_{i-1} \vee \varphi_{i+1} \ldots \vee \varphi_{j-1} \vee \varphi_{j+1} \vee \ldots \vee \varphi_n \\
&\equiv \underline{\varphi_i} \vee \varphi_1 \vee \ldots \vee \varphi_{i-1} \vee \varphi_{i+1} \ldots \vee \varphi_{j-1} \vee \varphi_{j+1} \vee \ldots \vee \varphi_n \\
&\equiv \varphi_1 \vee \ldots \vee \underline{\varphi_i} \vee \ldots \vee \varphi_{j-1} \vee \varphi_{j+1} \vee \ldots \vee \varphi_n,
\end{aligned}$$

where we essentially used commutativity and axiom (A12). Thus, we can reduce φ' to a provably equivalent formula φ^* in minimal disjunctive form.

Finally, by commutativity, associativity, and idempotence we can reorder φ^* to make it standard (without changing minimality). Thus, we get a formula φ_{NF} that is provably equivalent to φ and in standard minimal disjunctive form. □

The next lemma shows that for minimal disjunctive forms we already have completeness. Together with the just shown normal form lemma, this will give us the completeness theorem.

[118] The observation that we need *minimal* normal forms is the key to the completeness theorem. Interestingly, this is dual to the case of Fine (2016a) who needs *maximal* normal forms to prove completeness for his system AC.

[119] A literal is an atomic sentence or a negated atomic sentence.

Lemma 5.2.5. *For all sentences φ and ψ in standard minimal disjunctive form we have*

$$[\![\varphi]\!]_\Sigma = [\![\psi]\!]_\Sigma \textit{ iff } \varphi = \psi. \tag{5.1}$$

Proof of 5.2.5. The right-to-left direction is trivial, so we only consider the other one. We write

$$\varphi = \varphi_1 \vee \ldots \vee \varphi_r \vee \varphi'_1 \vee \ldots \vee \varphi'_s$$
$$\psi = \psi_1 \vee \ldots \vee \psi_u \vee \psi'_1 \vee \ldots \vee \psi'_v,$$

where $\varphi_1, \ldots, \varphi_r$ are exactly those disjuncts of φ that are – modulo ordering – also disjuncts of ψ (so the remaining $\varphi'_1, \ldots, \varphi'_s$ aren't disjuncts of ψ). Analogously, the unprimed disjuncts of ψ occur in φ, and the primed ones don't.

Next, we claim that primed disjuncts are extensions of unprimed ones, that is, for all $j \leqslant s$, $\varphi'_j = \varphi_i \wedge L$ (modulo ordering) for an $i \leqslant r$ and a set of literals L (L is a subset of the literals occurring in φ). Analogously for ψ.

Because: Assume for contradiction that there is a φ'_j that is not of this form. For a set of literals L we write s_L for a (any) scenario with the interpretation

$$I_s(p) := \begin{cases} (1,0) & p \in L \text{ and } \neg p \notin L \\ (0,1) & p \notin L \text{ and } \neg p \in L \\ (1,1) & p \in L \text{ and } \neg p \in L \\ (0,0) & p \notin L \text{ and } \neg p \notin L. \end{cases}$$

So $s \Vdash^+_\Sigma l$ for all $l \in L$ and $s \nVdash^+_\Sigma l$ for all literals l not in L. We consider two cases.

Case 1. There is no disjunct ψ_0 of ψ (primed or unprimed) such that $L(\psi_0) \subseteq L(\varphi'_j)$. Then $s_{L(\varphi'_j)} \Vdash^+_\Sigma \varphi$, but for any disjunct ψ_0 of ψ, there is a literal $l \in L(\psi_0)$ such that $l \notin L(\varphi'_j)$, so $s_{L(\varphi'_j)} \nVdash^+_\Sigma \psi_0$, hence $s_{L(\varphi'_j)} \nVdash^+_\Sigma \psi$, in contradiction to $[\![\varphi]\!]_\Sigma = [\![\psi]\!]_\Sigma$.

Case 2. There is a disjunct ψ_0 of ψ (primed or unprimed) such that $L(\psi_0) \subseteq L(\varphi'_j)$. Without loss of generality, we can choose a minimal such ψ_0, that is,

$$\text{For all disjuncts } \psi_* \text{ of } \psi, \text{ if } L(\psi_*) \subseteq L(\psi_0),^{120} \text{ then } L(\psi_*) = L(\psi_0). \tag{5.2}$$

If ψ_0 also is a disjunct of φ, then $\varphi'_j = \psi_0 \wedge (L(\varphi'_j) \setminus L(\psi_0))$ in contradiction to our assumption that φ'_j is not of this form. So assume ψ_0 is not a disjunct of φ.

We now argue by contradiction that for all disjuncts φ_0 of φ we have $L(\varphi_0) \not\subseteq L(\psi_0)$. If this were not the case, then there is a disjunct φ_0 of φ with $L(\varphi_0) \subseteq L(\psi_0)$. We again consider two cases. If there is no disjunct ψ_* of ψ such that $L(\psi_*) \subseteq L(\varphi_0)$ we have – as in case (1) – that $s_{L(\varphi_0)} \Vdash^+_\Sigma \varphi$, but $s_{L(\varphi_0)} \nVdash^+_\Sigma \psi$ (since any disjunct of ψ contains a literal that is not in $L(\varphi_0)$ and thus not true)—in contradiction to $[\![\varphi]\!]_\Sigma = [\![\psi]\!]_\Sigma$. If there is a disjunct ψ_* of ψ such that $L(\psi_*) \subseteq L(\varphi_0)$ we have $L(\psi_*) \subseteq L(\varphi_0) \subseteq L(\psi_0) \subseteq L(\varphi'_j)$, and hence, by minimality of ψ_0 (5.2), $L(\psi_*) = L(\psi_0)$, so (modulo ordering) $\psi_* = \varphi_0 = \psi_0$, so ψ_0 ($= \varphi_0$) is a disjunct of φ—contradiction.

Now, since for all disjuncts φ_0 of φ we have $L(\varphi_0) \not\subseteq L(\psi_0)$, we can again proceed as in case (1): We have $s_{L(\psi_0)} \Vdash^+_\Sigma \psi$, but $s_{L(\psi_0)} \nVdash^+_\Sigma \varphi$, in contradiction to $[\![\varphi]\!]_\Sigma = [\![\psi]\!]_\Sigma$.

Thus, the two cases both end in a contradiction, so all φ'_j's must be of the above form. The case for ψ'_j's is analogous. #

Now, since φ is minimal, no disjunct can be the extension of another one, hence the set of

[120] So, in particular, $L(\psi_*) \subseteq L(\varphi'_j)$.

primed disjuncts is empty. The same goes for ψ. Thus, φ and ψ really look like this:

$$\varphi = \varphi_1 \vee \ldots \vee \varphi_r$$
$$\psi = \psi_1 \vee \ldots \vee \psi_u,$$

and recall that the φ_i's also occur as disjuncts in ψ and vice versa. Hence

$$C(\varphi) = \{L(\varphi_1), \ldots, L(\varphi_r)\} = \{L(\psi_1), \ldots, L(\psi_u)\} = C(\psi).$$

Since φ and ψ are standard, their order is fixed, so $\varphi = \psi$, as wanted. □

As promised, now we can prove the completeness theorem.

Theorem 5.2.6 (Soundness and completeness)**.** *For all sentences φ and ψ we have*

$$[\![\varphi]\!]_\Sigma = [\![\psi]\!]_\Sigma \textit{ iff } \vdash_{\mathcal{S}} \varphi \equiv \psi.$$

Proof of 5.2.6. *Soundness (right to left).* It is readily checked that if $\varphi \equiv \psi$ is an axiom, then $[\![\varphi]\!]_\Sigma = [\![\psi]\!]_\Sigma$. Moreover, it is also readily checked that if $\varphi' \equiv \psi'$ is the result of applying one of the rules to $\varphi \equiv \psi$, and if $[\![\varphi]\!]_\Sigma = [\![\psi]\!]_\Sigma$, that then also $[\![\varphi']\!]_\Sigma = [\![\psi']\!]_\Sigma$.

Completeness (left to right). Assume $[\![\varphi]\!]_\Sigma = [\![\psi]\!]_\Sigma$, then by lemma 5.2.4 and soundness we have

$$[\![\varphi_{NF}]\!]_\Sigma = [\![\varphi]\!]_\Sigma = [\![\psi]\!]_\Sigma = [\![\psi_{NF}]\!]_\Sigma,$$

so by lemma 5.2.5 $\varphi_{NF} = \psi_{NF}$, so we can prove in $\mathcal{S}$ that

$$\varphi \equiv \varphi_{NF} \equiv \psi_{NF} \equiv \psi,$$

so $\vdash_{\mathcal{S}} \varphi \equiv \psi$. □

And as already announced, this enables us to show that the minimal normal form of a formula is unique.

Corollary 5.2.7 (Uniqueness of normal form)**.** *For each formula φ there is a unique formula φ_{NF} in standard minimal disjunctive form such that $\vdash_{\mathcal{S}} \varphi \equiv \varphi_{NF}$.*

Proof of 5.2.7. If φ_{NF} and φ'_{NF} are normal forms of φ, then provably $\varphi_{NF} \equiv \varphi \equiv \varphi'_{NF}$, so by soundness $[\![\varphi_{NF}]\!]_\Sigma = [\![\varphi'_{NF}]\!]_\Sigma$, so by lemma 5.2.5 $\varphi_{NF} = \varphi'_{NF}$. □

5.2.2 Related systems

We mention some systems related to $\mathcal{S}$.

The subsystem AC will be discussed in the next section. There we'll also introduce a system $\mathcal{S}^*$ that lies in between $\mathcal{S}$ and AC. French (2017) recently provided a sequent calculus for AC, and Ferguson (2014) provides a good discussion of systems similar to AC. As already mentioned, our system $\mathcal{S}$ is equivalent to equivalence in proof systems for first-degree entailment as, for example, provided by Priest (2008, ch. 8), or Font (1997). Moreover, it would be interesting to axiomatize strict synonymy when restricted to well-behaved scenarios. This could again be done by taking equivalence in a proof system for strong Kleene logic (which can be found, for example, in Font (1997), too). But it would also be interesting to see what has to be added to our system $\mathcal{S}$ to achieve this.[121] Concerning the other notions of validity that we defined, Pietz

[121] For reasons of time and since it is not needed here, I didn't check. But a guess would be to add

and Rivieccio (2013) provide a sound and complete logic for $\models^{tr^1}$ (which hence has explosion). And it would be interesting to define a proof system for our *e*x-validity, because in the usual systems for validity in first-degree entailment one has the rule $\varphi \vdash \varphi \vee \psi$ which *e*x-validity doesn't satisfy.

5.3 Something paradoxical about synonymy

In this section, we describe something paradoxical about the nature of synonymy or content identity and show how the paradox can be resolved. In the course of which, we characterize AC, we show that AC is not the first notion of synonymy above strict synonymy where content identity entails identity in subject matter, and we see that individuating according to AC-equivalence requires doing semantics with sets of scenarios.

In chapter 2, we've already seen something paradoxical about synonymy: cognitive synonymy and substitution salva veritate synonymy seem like two prima facie plausible reconstructions of the concept of synonymy, but they are inconsistent with each other. Now, the logic of strict synonymy from the preceding section makes apparent another somewhat paradoxical feature of synonymy or content identity. There are two intuitive principles that prima facie seem to be true about synonymy or content identity:

(P1) *Scenario respecting*: If there is no scenario – no matter how incomplete or how inconsistent – in which two propositional sentences come apart[122], then they are absolutely synonymous (or identical in content).

(P2) *Subject matter preserving*: If two propositional sentences are absolutely synonymous (or identical in content), then they have the same subject matter, that is, the sets of their atomic sentences are identical.

(An incidental remark on the notion of subject matter: Intuitively, the subject matter of a sentence is what it is about. In (P2), we find a syntactic conception of subject matter: The subject matter of a logically complex sentence is the set of atomic sentences occurring in the complex sentence. For our purposes here, this is a good enough *approximation* to the concept of subject matter of logically complex sentences. There are much more sophisticated notions of subject matter that are semantic instead of syntactic and that are not restricted to complex sentences: see, for example, Yablo (2014a), Lewis (1988), or Fine (2016b).)

However, our results show that jointly these two principles are inconsistent: For take the two propositional sentences p and $p \vee (p \wedge q)$—which are an instantiation of axiom (A12). Then (P1) tells us that they are absolutely synonymous (since p and $p \vee (p \wedge q)$ cannot be distinguished by any scenario[123]). But the contraposition of (P2) tells us that they are not absolutely synonymous (since they don't share the same atoms).

$\varphi \wedge \neg\varphi \equiv (\varphi \wedge \neg\varphi) \wedge (\psi \vee \neg\psi)$.

[122] That is, a scenario in which one sentence is true while the other isn't, or in which one sentence is false while the other isn't.

[123] Assuming that our notion of a scenario captures the intuitive notion of a scenario used in (P1)—or, in other words, is an adequate rational reconstruction of the intuitive notion. We argue that this is the case in the paragraph after the next.

A theory about synonymy or content identity shouldn't leave this inconsistency uncommented. Our answer will be that there is a plurality of notions of synonymy or content identity and that there are notions satisfying one principle and "almost" the other—that is, they avoid the inconsistency while giving us as much as we can hope for. But to get there we first have to better understand the two principles. We already have a good understanding of (P1) – which we summarize now –, and then we develop a good understanding of (P2) in the next two subsections, allowing us to present the announced resolution in section 5.3.3.

We understand (P1) fairly well: A scenario on our reconstruction is a conceptualization or representation of a (possible) part the world—incompleteness and inconsistencies being allowed. When asked to spell out what it means for a scenario to make logically complex sentences true, we proceeded via the four-valued logic of first-degree entailment—and this arguably was the most natural choice. Thus, we arrived at an independently and intrinsically motivated reconstruction of the intuitive notion of a scenario (and when such a scenario distinguishes sentences). Hence content identity in our scenario semantics precisely satisfies (P1). Strict synonymy is how fine one can get if one takes scenarios as a starting point for semantics. We take this argument to show that the paradox doesn't indicate that our notion of strict synonymy is flawed, but that the paradox rather indicates that there must be a plurality of notions of (absolute) synonymy. But we'll discuss this in more detail in sections 5.3.3 and 5.4 below. For now, let's better understand (P2).

5.3.1 Characterizing Fine's analytic containment (AC)

In this section, we want to answer the question: What are the notions of synonymy that satisfy (P2)? The answer will be: AC provides one such notion but it is not the first above strict synonymy. The first is a notion that we will call super strict synonymy.

As a start, we know one notion of synonymy that does satisfy (P2). (Apart from the trivial syntactic identity relation.)

Analytic synonymy Two propositional sentences φ and ψ are analytically synonymous ($\varphi \approx_{\mathsf{AC}} \psi$) iff $\vdash_{\mathsf{AC}} \varphi \equiv \psi$.

(A straightforward induction on AC-proofs shows that, indeed, if $\vdash_{\mathsf{AC}} \varphi \equiv \psi$, then $\mathsf{At}(\varphi) = \mathsf{At}(\psi)$.) This notion was introduced as "analytic containment" by Angell (1977, 1989) and – as already mentioned – it recently was extensively elaborated by Fine (2016a)—which was one reason to base our system $\mathcal{S}$ on AC.

So analytic synonymy is one notion of synonymy in the class of notions satisfying (P2). We'll now discuss what other notions there are, and which of them are good representatives for the class.

For this discussion it will be helpful to introduce some terminology. A standard disjunctive form was defined in definition 5.2.3 and – following Fine (2016a, 214f.) – it is said to be *maximal* if whenever it contains a disjunct φ_i and literal l (appearing as a conjunct of some disjunct), then it contains the disjunct $\varphi_i \wedge l$ (modulo the order of the literals in the disjunct). Fine (2016a, theorem 18 & 22, p. 216f.) shows that every $\mathcal{L}_p$-sentence φ is AC-provably equivalent to a unique standard maximal disjunctive form φ_{mNF}. We write $\mathsf{L}(\varphi_{\mathrm{mNF}})$ for the set of literals occurring in φ_{mNF}.

Definition 5.3.1 (Overlap, literal overlap, FDE-preservation)**.** A binary relation $\approx$ between $\mathcal{L}_{\mathrm{p}}$-sentences is said to satisfy *overlap*, *literal overlap*, and *FDE-preservation* if, respectively

$$\forall \varphi, \psi \in \mathcal{L}_{\mathrm{p}} : \ \varphi \approx \psi \Rightarrow \mathrm{At}(\varphi) = \mathrm{At}(\psi) \qquad \text{(Overlap)}$$
$$\forall \varphi, \psi \in \mathcal{L}_{\mathrm{p}} : \ \varphi \approx \psi \Rightarrow \mathrm{L}(\varphi_{\mathrm{mNF}}) = \mathrm{L}(\psi_{\mathrm{mNF}}) \qquad \text{(Literal overlap)}$$
$$\forall \varphi, \psi \in \mathcal{L}_{\mathrm{p}} : \ \varphi \approx \psi \Rightarrow \big(\varphi \Leftrightarrow_{\mathrm{FDE}} \psi\big) \qquad \text{(FDE-preservation)}$$

where $\varphi \Leftrightarrow_{\mathrm{FDE}} \psi$ means that φ and ψ are equivalent in the logic of first degree entailment.

Thus, a notion that satisfies overlap is in the class of (P2)-synonymies. So one could pick as a representative the most coarse-grained notion of synonymy satisfying overlap: that is, the notion of "synonymy" where two sentences are synonymous if they have the same atoms. However, this is not a serious notion of synonymy since, for example, the contradiction $\mathrm{p} \wedge \neg \mathrm{p}$ is synonymous to the tautology $\mathrm{p} \vee \neg \mathrm{p}$. Demanding FDE-preservation is one way to exclude such notions. And – in light of (P1) – one should arguably demand that any serious notion of absolute synonymy between logically complex sentences should satisfy FDE-preservation—for example, strict synonymy satisfies FDE-preservation.

Moreover, note that literal overlap implies overlap, but the other direction doesn't hold (e.g. $\mathrm{p} \vee (\mathrm{p} \wedge \mathrm{q})$ and $\mathrm{p} \vee (\mathrm{p} \wedge \neg \mathrm{q})$ satisfy overlap but not literal overlap).

Now we can characterize AC to locate it in the class of (P2)-synonymies.

Theorem 5.3.2 (Characterization of AC)**.** *We have for all $\mathcal{L}_{\mathrm{p}}$-sentences φ and ψ that*

$$\varphi \approx_{\mathrm{AC}} \psi \textit{ iff } \begin{cases} \mathrm{L}(\varphi_{\mathrm{mNF}}) = \mathrm{L}(\psi_{\mathrm{mNF}}) & \textit{, and} \\ \varphi \Leftrightarrow_{\mathrm{FDE}} \psi & \textit{.} \end{cases} \tag{5.3}$$

In other words, $\approx_{\mathrm{AC}}$ is the maximally coarse-grained binary relation on $\mathcal{L}_{\mathrm{p}}$ satisfying literal overlap and FDE-preservation.

Proof of 5.3.2. We move the lengthy proof to the appendix (section B.2). □

The fact that we needed the stronger literal overlap instead of overlap simpliciter should make us wonder whether there is a more coarse-grained notion of synonymy above strict synonymy that satisfies overlap but is not yet analytic synonymy. And indeed, there is. To see this, we first define the following.

Definition 5.3.3 (Super strict synonymy)**.** Let $\mathcal{S}^*$ be the proof system $\mathcal{S}$ of definition 5.2.1 where we replace the axiom (A12) by

$$(\mathrm{A12}^*) \quad \varphi \vee (\varphi \wedge \psi) \equiv \varphi \vee (\varphi \wedge \neg\psi).$$

And we define the following notion of synonymy.

Super strict synonymy Two propositional sentences φ and ψ are *super strictly synonymous* ($\varphi \approx_{\mathtt{sstr}} \psi$) iff $\vdash_{\mathcal{S}^*} \varphi \equiv \psi$.

Clearly, super strict synonymy also satisfies overlap (that is, it is in the class of (P2)-synonymyies). We can now characterize super strict synonymy to locate it in the class of (P2)-synonymies.

Theorem 5.3.4 (Characterization of super strict synonymy)**.** *We have for all $\mathcal{L}_p$-sentences φ and ψ that*

$$\varphi \approx_{sstr} \psi \textit{ iff } \begin{cases} At(\varphi) = At(\psi) & \textit{, and} \\ \varphi \Leftrightarrow_{FDE} \psi & . \end{cases} \tag{5.4}$$

In other words, $\approx_{sstr}$ is the maximally coarse-grained binary relation on $\mathcal{L}_p$ satisfying overlap and FDE-preservation.

Proof of 5.3.2. We move the lengthy proof to the appendix (section B.3). □

So we now have an answer to the question: What do we get if we make strict synonymy more fine-grained so that it satisfies overlap but change it as little as possible? The answer is: We precisely get super strict synonymy (since strict synonymy already satisfies FDE-preservation). Analytic synonymy is even more fine-grained but, too, satisfies overlap.

Also note that super strict synonymy lies strictly in between strict and analytic synonymy (for $\mathcal{L}_p$-sentences)

$$\approx_{str} \supsetneq \approx_{sstr} \supsetneq \approx_{AC},$$

since the relations are equivalent, respectively, to the properly increasingly stronger conditions $\Leftrightarrow_{FDE}$, $\Leftrightarrow_{FDE}$ & overlap, and $\Leftrightarrow_{FDE}$ & literal overlap. To have an explicit example of the latter, $p \vee (p \wedge q)$ and $p \vee (p \wedge \neg q)$ are $\mathcal{S}^*$-provably equivalent but they are not AC-equivalent (since they are two non-identical maximal normal forms).

5.3.2 Move on up: From strict synonymy to analytic synonymy

Via the inconsistency of (P1) and (P2) mentioned in the beginning of this section, we know that for analytic synonymy – which satisfies (P2) – we cannot have (P1). That is, the fact that two sentences cannot be distinguished by any scenario – no matter how incomplete or inconsistent – doesn't guarantee that the two sentences are analytically synonymous. In this section, we want to show that we can formulate a weaker version of (P1) that *is* satisfied by analytic synonymy. This necessarily involves moving one level up: from scenarios making sentences true or false, to sets of scenarios making sentences true or false.

The key result of Fine (2016a, theorem 21, p. 216) is that AC is sound and complete with respect to his truthmaker semantics—a kind of semantics that can be traced back to Van Fraassen (1969). In short, the semantics is as follows. A state model M is a triple $(S, \sqsubseteq, |\cdot|)$ where $\sqsubseteq$ is a complete partial order[124] on S and $|\cdot|$ maps atomic sentences p to pairs $(|p|^+, |p|^-)$ of non-empty subsets of S. (Elements of $|p|^+$ are called verifiers of p and elements of $|p|^-$ are called falsifiers of p.) As in Fine (2016a, p. 205), we recursively define when a $\mathcal{L}_p$-sentence φ is verified/falsified by a state $s \in S$ (in signs: $s \Vdash \varphi$ / $s \dashv\Vdash \varphi$).

- $s \Vdash p$:iff $s \in |p|^+$, and $s \dashv\Vdash \varphi$:iff $s \in |p|^-$
- $s \Vdash \neg\varphi$:iff $s \dashv\Vdash \varphi$, and $s \dashv\Vdash \neg\varphi$:iff $s \Vdash \varphi$

[124] So $\sqsubseteq$ is reflexive, transitive, anti-symmetric, and every subset of S has a least upper bound.

- $s \Vdash \varphi \wedge \psi$:iff $\exists u, t \in S : u \Vdash \varphi$ & $t \Vdash \psi$ & $s = u \sqcup t$ [125], and
  $s \dashv \varphi \wedge \psi$:iff $s \dashv \varphi$ or $s \dashv \psi$
- $s \Vdash \varphi \vee \psi$:iff $s \Vdash \varphi$ or $s \Vdash \psi$, and
  $s \dashv \varphi \vee \psi$:iff $\exists u, t \in S : u \dashv \varphi$ & $t \dashv \psi$ & $s = u \sqcup t$

Following Fine (2016a, p. 208), the exact content of φ is $|\varphi| := \{s \in S \mid s \Vdash \varphi\}$, and the (replete) content of φ – denoted $[\varphi]$ – is the convex closure of the complete closure of $|\varphi|$.[126] For reasons of space, we cannot go into the philosophical difference between these two notions of content—but see Fine (2016a, sec. 4–5).

Analogous to the canonical model of Fine (2016a, 215f.), we define the *canonical scenario model* as $C := (\mathcal{P}(\Sigma), \subseteq, |\cdot|_C)$ where $|p|_C := \big\{(V, F) \mid V = \{\{s_p\}\}, F = \{\{s_{\neg p}\}\}\big\}$ where s_p is the program $\{\top \to p\}$ and $s_{\neg p}$ is the program $\{\bot \to p\}$. (It's easily verified that C indeed is a state model.)

It's straightforwardly checked that $[p \vee (p \wedge q)]_C = \big\{\{s_p\}, \{s_p, s_q\}\big\}$ and $[p]_C = \big\{\{s_p\}\big\}$. So the set of scenarios $\{s_p, s_q\} \in \mathcal{P}(\Sigma)$ makes $p \vee (p \wedge q)$ true but not p. Thus, if we move from scenarios making sentences true to sets of scenarios making sentences true, then we get a semantics that can distinguish between the strictly synonymous sentences p and $p \vee (p \wedge q)$. Indeed, in general we have the following.

Theorem 5.3.5 (Semantics with sets of scenarios)**.** *The following are equivalent*

(i) $\varphi \equiv \psi$ *is valid in truthmaker semantics (i.e. for every state model* M *we have* $[\varphi]_M = [\psi]_M$*).*

(ii) $\vdash_{AC} \varphi \equiv \psi$

(iii) φ *and* ψ *literally overlap and* $\varphi \Leftrightarrow_{FDE} \psi$

(iv) $[\varphi]_C = [\psi]_C$ *(where* C *is the canonical scenario model based on sets of scenarios).*

Proof of 5.3.5. We move the proof to the appendix (section B.4). □

In other words, if we do semantics based on scenarios, then no matter how incomplete or inconsistent scenarios we allow for, we never get to a level of granularity where we can distinguish all analytically synonymous sentences. If we wish to achieve such a level of granularity, we have to move one level up: we have to do semantics based on sets of scenarios. Here "have to" means the following: It is sufficient to move one level up to get a semantics that individuates content according to analytic synonymy. And any other semantics that individuates content according to analytic synonymy is equivalent – in terms of content individuation – to the sets-of-scenarios semantics. In other words, moving to sets of scenarios is not just an *ad hoc* move to get a notion of content that individuates according to analytic synonymy, it rather is an "extensionally" necessary move.

This yields an intuitive two-level picture of content: On the first granularity level, content can be modeled by first-order objects like scenarios or possible worlds. And

[125] Where $u \sqcup t$ denotes the least upper bound of $\{u, t\}$ which exists by the completeness of the partial order.

[126] A set $T \subseteq S$ is convex if for all $s, t, u \in S$, if $s, t \in T$ and $s \sqsubseteq u \sqsubseteq t$, then $u \in T$. And a set $T \subseteq S$ is complete if for every $T_0 \subseteq T$ the least upper bound of T_0 is in T. The complete (resp. convex) closure of a set $T \subseteq S$ is the smallest set $T^* \subseteq S$ such that $T \subseteq T^*$ and T^* is complete (resp. convex).

on the second granularity level, content is modeled by second-order objects: like sets of scenarios.

Finally, we see that the situation is this. Truthmaker (or sets-of-scenarios) semantics satisfies (P2), that is, its notion of content identity satisfies overlap. And, as already seen, it hence cannot satisfy (P1). However, it satisfies a weak version of (P1), namely

(P1') If there is no *set of* scenarios that (exactly) verifies one sentence but not the other, then the sentences are absolutely synonymous.

To recite the above example: While there is no single scenario making $p \vee (p \wedge q)$ true but not p, there is a set of scenarios – namely $\{s_p, s_q\}$ – that verifies $p \vee (p \wedge q)$ but doesn't (exactly) verify p. (For reasons of space, we won't make this more precise.)

Of course, we could go on and find a similar semantic characterization for super strict synonymy, but for reasons of space and since analytic synonymy is the much more famous notion, we don't.

5.3.3 Resolving the paradox

Given the last two subsections, here is a way to look at the paradoxical nature of synonymy mentioned at the beginning of the section.

It arguably is surprising to realize that there cannot be one strong notion of synonymy – the one alluded to by the intuitive term "absolute synonymy" – that satisfies both principles. But we've now seen that there are notions that come very close to satisfying both. Strict synonymy is one example. It exactly satisfies (P1) but it doesn't satisfy (P2). However, the instances of sentences violating (P2) can be traced back to exactly one axiom: $\varphi \equiv \varphi \vee (\psi \wedge \varphi)$. Another example is super strict synonymy. It exactly satisfies (P2), but it doesn't satisfy (P1). However, it satisfies a weaker version of (P1): If neither scenarios nor sets of scenarios can distinguish two sentences, then they are absolutely synonymous. And analytic synonymy also satisfies (P2) – though not exactly – but satisfies exactly the weaker version of (P1).

Taking a step back, this paradox seems to point us towards a pluralistic conception of synonymy. We will now take this up in the next section.

5.4 Pluralism about synonymy

In this last section, we take stock of all the notions of synonymy that we've seen. And we show how a pluralistic conception of synonymy can reconcile the many opposing features of synonymy that we've seen by now. This provides two arguments for pluralism about synonymy: the sheer number of independently well-motivated notions of synonymy, and the fact that only a plurality of notions of synonymy can reconcile the opposing features.

The sheer number argument. The relation between the notions of synonymy is visualized in figure 5.4 (where $\approx_{cr}$ is cognitive-role synonymy and $\approx_{rt}$ is reasoned-to synonymy). Of course, by using impossible worlds semantics we could go even more fine-grained than analytic synonymy—all the way to syntactic identity. But for reasons of space, we stop here. This sheer number of independently well-motivated notions of synonymy is one argument for a plurality of notions of synonymy.

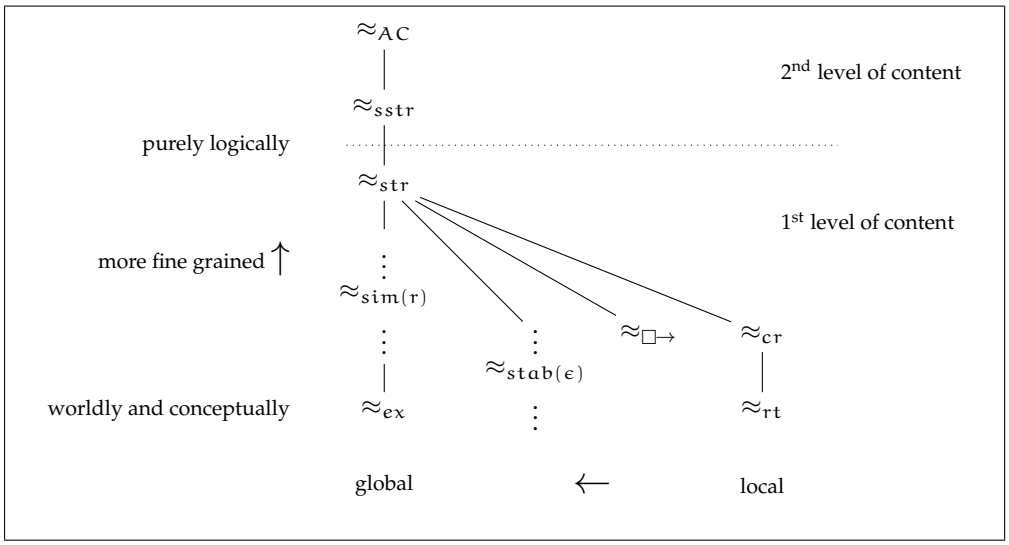


Figure 5.2: The inclusion ordering of the notions of synonymy

The reconciling opposing features argument. The other argument for pluralism about synonymy is that it seems to be the only way to account for the many opposing features of synonymy. We've already mentioned several of them in the introduction (section 1.1): contextual stability vs. flexibility, objective vs. subjective, externalism vs. internalism, respecting logical equivalence vs. not, equivalence vs. similarity, and extensional vs. intensional. Moreover, in chapter 2, we've seen that substitution salva veritate synonymy is inconsistent with cognitive synonymy. And finally, in the preceding section 5.3 we've seen that the two intuitive principles (P1) and (P2) about "absolute" synonymy are inconsistent.

We'll now sketch how we can account for these opposing features—for reasons of space we only exemplarily pick some of them.

Ad stability vs. flexibility. We've observed that synonymy between sentences has a certain contextual stability, but that we can also (basically) always cook up a context where the meaning of two non-identical (atomic) sentences comes apart. Similarly, there is, on the one hand, overwhelming evidence that the meaning of expressions varies massively across contexts, and on the other hand there has to be some "contextually stable" meaning of expressions—already because otherwise communication would not be possible (as words would basically always mean something different on every occasion).[127] In the scenario framework, this is accounted for as follows (the details are provided in chapter 3, so we stay rather informal here). We take two sentences p and q to be synonymous if we have the defeasible rules "If p, then q" and "If q, then p" in our knowledge base (resp. if the two units representing the sentences are linked in the neural implementation of our knowledge base). Now, in a (cooked up) context that makes p and q come apart in meaning, the context adds additional information to our knowledge base that triggers an abnormality that blocks the link between p and q. (Neurally speaking, if this context is perceived, it increases activity in the units feeding into abnormality unit between the units representing p and q.)

[127] Herman Cappelen and collaborators have precisely articulated this phenomenon and worked a lot on it: E.g. Cappelen (2008), Cappelen and Lepore (2008) or Cappelen and Lepore (2005). But their account for this phenomenon is different from the one we'll present now.

However, in most contexts where p and q are used, no such abnormality is added, so the synonymy of p and q has a certain contextual stability. In short, stability is provided by rules and flexibility is provided by abnormalities that can inhibit the rules. (For more on how logic programming describes reasoning to an interpretation of a sentence see Besold et al. (2017).) Formally, this is can be reconstructed by the notions of counterfactual or stable synonymy.

Ad objective vs. subjective. A subjective and descriptive notion of synonymy is obtained if we consider scenarios whose program represents the knowledge base of particular agents. An objective and normative notion of synonymy is obtained if we consider scenarios whose program represents the generic knowledge base (see section 3.1.3).

Ad extensional vs. intensional. We've seen that modal intensions can be captured in the scenario framework by regarding possible worlds as limits of appropriate sequences of scenarios. Representational intensionality (that is, a form of hyperintensionality) is available since the scenarios can be interpreted as describing how an agent represents or conceptualizes the world.

Ad scenario respecting vs. subject matter preserving. As seen in section 5.3.3, if we adopt a pluralistic conception of synonymy, then we can resolve the inconsistency between (P1) and (P2) as well as we could hope for: There are many notions of synonymy all on a par with each other. In different contexts, we use different notions. On this picture, it is fine (and non-surprising) that some notions – especially the more coarse-grained ones – don't satisfy one (or any) of the jointly inconsistent principles. Nonetheless, one might be convinced that there are well-motivated weaker forms of synonymy but still think that there should be one strong notion of synonymy that satisfies both principles. On this view, it is surprising to realize that there cannot be such a notion. But with strict synonymy and super strict (resp. analytic) synonymy, we've provided two equally justified notions of synonymy where each exactly satisfies one principle and nearly – that is, as much as we could hope for – the other principle.

Ad cognitive synonymy vs. substitution salva veritate. In chapter 3, we reconstructed the notion of cognitive synonymy (for atomic sentences) as cognitive-role identity. This yielded a compromise between being cognitively adequate and still having some substitution properties for belief. But of course, this is not the only notion of synonymy that finds a such a compromise. Others could be obtained – not by starting from cognitive synonymy but – by weakening the notion of co-hyperintensionality (or substitution salva veritate) until one thus arrives at a notion a notion of synonymy that is cognitively adequate. For example, one could drop the minimality condition of co-hyperintensionality to also allow for intransitive relations (such that their transitive closure is the original co-hyperintensionality relation).[128] Or instead of formulating the substitution condition by (transitive) valid inference one could formulate it as a "correct" (possibly intransitive) conditional—which is essentially what we did via the defeasible rules. Also, one could replace "(always) valid" by "almost always valid"—which is captured by similarity or exception-tolerant synonymy (equivalence on all scenarios except for on a negligible set of scenarios).

In sum, pluralism about synonymy seems to be the only way to account for the many opposing features of synonymy. Like many other philosophical concepts,

[128] Cf. cognitive-role synonymy vs. reasoned-to synonymy.

synonymy is not a single concept but rather a cluster of related concepts. Each member of the cluster has some properties that others don't have, but together they form a plurality of notions of synonymy.

APPENDIX

A More on knowledge bases

We continue the discussion of the agent-based and generic knowledge base from section 3.1.3. We consider two worries (and replies) concerning (i) the notion of a set of relevant agents and (ii) a generic knowledge base.

Ad (i). Isn't it a problem that we don't know exactly what agents should be in A? Of course, it's very difficult to spell out exactly what the relevant possible agents are: If we're too permissive and include, say, an utterly ignorant possible agent, no law whatsoever would hold since no cognitive synonymy would ever be detected. If we're too restrictive and include, say, only logically omniscient possible agents, identity of cognitive role would coincide with logical equivalence—and we precisely set out to avoid this since this is not how actual humans work. So the relevant possible agents are those agents whose intellectual capabilities lie somewhere strictly between utter ignorance and logical omniscience (also cf. Jago 2014, p. 11).

So it's hard to spell out exactly the notion of a relevant possible agent, and it might even be context-sensitive; but fortunately, all we need here is that the set A of relevant possible agents exists—we don't need to know exactly what its members are. (And if this set were not to exist, there would be no truth to the matter of what the cognitively speaking correct individuation of content is.)

Moreover, we will treat the set A as a variable: The following will work for any particular choice of A, and A can also be context-sensitive—in short, A can be chosen depending on the intended interpretation.

Ad (ii). How exactly can the generic knowledge base be seen as the common ground of the agent-based knowledge bases? Given an agent $a \in A$ we can consider her knowledge base KB_a. Now, KB is the result of merging all the individual knowledge bases KB_a into one that combines what is common to all the individual ones. Of course, the crucial question is how this merging should work exactly.

One obvious first choice would be to define KB as the intersection of the individual knowledge bases: $KB := \bigcap_{a \in A} KB_a$. This is the minimal choice (in the set-inclusion ordering of possible definitions of KB). For the time being, we consider it an open question whether there are more permissive ways of defining KB (e.g. by weighing the importance of rules by their probability of occurring among all the KB_a's).[129] But again, for our purposes, it only matters that there exist such a generic knowledge base—how it looks like exactly is not important for now.

[129] For other ways, we can look at database theory or at the theory of group knowledge.

Note that the rules (or "laws") in KB are grounded in (or supervene on) the clauses of the individual knowledge bases—in the sense that KB is built out of the individual knowledge bases.

B Outsourced proofs

B.1 Proof of the replacement rule

Lemma 5.2.2 (Replacement)**.** The following rules are admissible, that is, if their premiss is S-derivable, then the conclusion is S-derivable. (When $\chi[\varphi]$ is a formula containing occurrences of φ, then $\chi[\psi]$ is the result of replacing the occurrences of φ by ψ).

$$\text{(PR)}\ \frac{\varphi \equiv \psi}{\chi[\varphi] \equiv \chi[\psi]}\ \text{where the occurrences of } \varphi \text{ in } \chi[\varphi] \text{ are not in the scope of } \neg.$$

$$\text{(NR)}\ \frac{\varphi \equiv \psi}{\neg\varphi \equiv \neg\psi}$$

$$\text{(FR)}\ \frac{\varphi \equiv \psi}{\chi[\varphi] \equiv \chi[\psi]}$$

Proof. The proof of the admissibility of (PR) works in our setting exactly as in the setting of (AC), and for this see Fine (2016a, lemma 1, p. 202).

For (NR), the proof is by induction on the proof of $\varphi \equiv \psi$ (from which $\neg\varphi \equiv \neg\psi$ is then derived). All the cases corresponding to the axioms and rules that are also in the system (AC) are dealt with in Fine (2016a, lemma 2, p. 203f.). So we only have to consider the new axiom (A12). So suppose $\varphi \equiv \psi$ is $\varphi_0 \vee (\varphi_0 \wedge \psi_0) \equiv \varphi_0$. We have to show that $\neg\varphi \equiv \neg\psi$ is derivable. Indeed, we have

$$\begin{aligned}
\neg\varphi \;=\; \neg\big(\varphi_0 \vee (\varphi_0 \wedge \psi_0)\big) &\equiv \neg\varphi_0 \wedge \neg(\varphi_0 \wedge \psi_0) && \text{by DeMorgan}\\
&\equiv \neg\varphi_0 \wedge (\neg\varphi_0 \vee \neg\psi_0) && \text{by DeMorgan}\\
&\equiv (\neg\varphi_0 \wedge \neg\varphi_0) \vee (\neg\varphi_0 \wedge \neg\psi_0) && \text{by Distributivity}\\
&\equiv (\neg\varphi_0) \vee (\neg\varphi_0 \wedge \neg\psi_0) && \text{by (A5) and (R4)}\\
&\equiv \neg\varphi_0 \;=\; \neg\psi && \text{by (A12).}
\end{aligned}$$

Finally, (FR) follows from (PR) and (NR), which is noted in Fine (2016a, corollary, p. 204). □

B.2 Proof of the theorem characterizing AC

Theorem 5.3.2 (Characterization of AC)**.** We have for all $\mathcal{L}_p$-sentences φ and ψ that

$$\varphi \approx_{AC} \psi \text{ iff } \begin{cases} L(\varphi_{mNF}) = L(\psi_{mNF}) & \text{, and} \\ \varphi \Leftrightarrow_{FDE} \psi & \text{.} \end{cases} \tag{5.5}$$

In other words, $\approx_{AC}$ is the maximally coarse-grained binary relation on $\mathcal{L}_p$ satisfying literal overlap and FDE-preservation.

Proof. Let's start with the right-to-left direction. We show $\approx_{AC}$ implies $\Leftrightarrow_{FDE}$ by contraposition: If $\varphi \not\Leftrightarrow_{FDE} \psi$, then $\varphi \not\approx_{str} \psi$, so $\varphi \not\approx_{AC} \psi$ (since $\mathcal{S}$ is an extension of AC). And $\approx_{AC}$ implies literal overlap because of the following. If $AC \vdash \varphi \equiv \psi$, then $AC \vdash \varphi_{mNF} \equiv \varphi \equiv \psi \equiv \psi_{mNF}$. So φ_{mNF} and ψ_{mNF} are two sentences in standard maximal normal form that are AC-equivalent to φ. Since a sentence's standard maximal normal form is unique, $\varphi_{mNF} = \psi_{mNF}$. Hence, in particular, $L(\varphi_{mNF}) = L(\psi_{mNF})$.

Let's concentrate on the left-to-right direction. We claim that for all $\mathcal{L}_p$-sentences φ and ψ in standard maximal normal form we have

$$\text{If } L(\varphi_{mNF}) = L(\psi_{mNF}) \text{ and } \varphi \Leftrightarrow_{FDE} \psi, \text{ then } \varphi = \psi. \tag{5.6}$$

Before proving this, let's see that it finishes the proof. Assume $L(\varphi_{mNF}) = L(\psi_{mNF})$ and $\varphi \Leftrightarrow_{FDE} \psi$. We have $\vdash_{AC} \varphi_{mNF} \equiv \varphi$ and $\vdash_{AC} \psi_{mNF} \equiv \psi$. Moreover, we've seen in the right-to-left direction that AC-equivalence entails FDE-preservation, so we have

$$\varphi_{mNF} \Leftrightarrow_{FDE} \varphi \Leftrightarrow_{FDE} \psi \Leftrightarrow_{FDE} \psi_{mNF}.$$

Since $(\varphi_{mNF})_{mNF} = \varphi_{mNF}$ we have

$$L\big((\varphi_{mNF})_{mNF}\big) = L\big(\varphi_{mNF}\big) = L\big(\psi_{mNF}\big) = L\big((\psi_{mNF})_{mNF}\big).$$

Hence, applying (5.6) to φ_{mNF} and ψ_{mNF} we get that $\varphi_{mNF} = \psi_{mNF}$. And since $\vdash_{AC} \chi \equiv \chi$ we get that AC indeed proves

$$\varphi \equiv \varphi_{mNF} \equiv \psi_{mNF} \equiv \psi.$$

So it remains to show (5.6). We use an argument similar to the proof of lemma 5.2.5. Let $\varphi = \varphi_1 \vee \ldots \vee \varphi_n$ and $\psi = \psi_1 \vee \ldots \vee \psi_m$ be the sentences in standard maximal normal form. By standardness, it suffices to show that $C(\varphi) = C(\psi)$, that is,

$$\big\{L(\varphi_1), \ldots, L(\varphi_n)\big\} = \big\{L(\psi_1), \ldots, L(\psi_m)\big\}.$$

Assume for contradiction that not. Then, without loss of generality, there is a φ_i ($i \leqslant n$) such that $L(\varphi_i) \notin C(\psi)$ (the other case is analogous). Without loss of generality, we may assume that $L(\varphi_i)$ is $\subseteq$-minimal with this property (otherwise take the minimal one). That is,

$$\forall j \leqslant n : L(\varphi_j) \subsetneq L(\varphi_i) \Rightarrow L(\varphi_j) \in C(\psi). \tag{5.7}$$

We consider two cases. Case 1: For all ψ_k ($k \leqslant m$) we have $L(\psi_k) \not\subseteq L(\varphi_i)$. Then, for each $r \leqslant m$, there is an $l_r \in L(\psi_r)$ with $l_r \notin L(\varphi_i)$. Let $M := \{l_1, \ldots, l_m\}$. Note that $M \cap L(\varphi_i) = \emptyset$. For a set of literals L write L^+ for the positive (i.e. atomic) literals in L and L^- for the negative literals (i.e. negated atoms) in L. Consider the four-valued

valuation $v : S \to \{0, 1, u, \bot\}$ defined by

$$v(p) := \begin{cases} 1 & \text{, if } p \in L(\varphi_i)^+ \text{ and } \neg p \notin L(\varphi_i)^- \\ 0 & \text{, if } p \notin L(\varphi_i)^+ \text{ and } \neg p \in L(\varphi_i)^- \\ \bot & \text{, if } p \in L(\varphi_i)^+ \text{ and } \neg p \in L(\varphi_i)^- \\ u & \text{, if } p \notin L(\varphi_i)^+ \text{ and } \neg p \notin L(\varphi_i)^-. \end{cases}$$

Then $v(\varphi_i) \in \{1, \bot\}$ since each literal in φ_i is either 1 or $\bot$ under v. Hence $v(\varphi) \in \{1, \bot\}$ since φ is a disjunction with disjunct φ_i.

Moreover, we claim that $v(\psi) \in \{0, u\}$. It suffices to show that for each $k \leqslant m$, $v(\psi_k) \in \{0, u\}$. Indeed, pick a $k \leqslant m$. Then $l_k \in M$. If $l_k = q$ for an atomic sentence q, then $q \notin L(\varphi_i)$ (since $M \cap L(\varphi_i) = \emptyset$), and in particular, $q \notin L(\varphi_i)^+$. Hence, if $\neg q \in L(\varphi_i)^-$, then $v(q) = 0$, and if $\neg q \notin L(\varphi_i)^-$, then $v(q) = u$. Hence $v(q) = v(l_k) \in \{0, u\}$. Dually, if $l_k = \neg q$ for an atomic q, then $v(\neg q) = v(l_k) \in \{0, u\}$. Since $v(l_k) \in \{0, u\}$, also $v(\psi_k) \in \{0, u\}$ since ψ_k is a conjunction with conjunct l_k.

Hence, $v(\varphi) \in \{1, \bot\}$ while $v(\psi) \in \{0, u\}$ in contradiction to φ and ψ being FDE-equivalent.

Case 2: There is a ψ_k ($k \leqslant m$) such that $L(\psi_k) \subseteq L(\varphi_i)$. We again can take $L(\psi_k)$ to be minimal, that is,

$$\forall j \leqslant m : L(\psi_j) \subseteq L(\psi_k) \text{ and } L(\psi_j) \subseteq L(\varphi_i) \Rightarrow L(\psi_j) = L(\psi_k). \tag{5.8}$$

Again, if there is no φ_r such that $L(\varphi_r) \subseteq L(\psi_k)$, we can construct – as in case 1 – a four-valued valuation v such that $v(\psi) \in \{1, \bot\}$ and $v(\varphi) \in \{0, u\}$, in contradiction to φ and ψ being FDE-equivalent.

So assume that there is a φ_r such that $L(\varphi_r) \subseteq L(\psi_k)$. Since $L(\varphi_i) \neq L(\psi_k)$ (otherwise φ_i would be in ψ), we have $L(\varphi_r) \subseteq L(\psi_k) \subsetneq L(\varphi_i)$, so by (5.7) we have $L(\varphi_r) \in C(\psi)$. So by (5.8), $L(\varphi_r) = L(\psi_k) \in C(\psi)$. We write $A := L(\varphi_i) \setminus L(\varphi_r)$, so $L(\varphi_i) = L(\varphi_r) \cup A$. Since φ and ψ have the same literals, the literals in A also occur in ψ. So, since $\varphi_r = \psi_k$ is a disjunct of ψ and ψ is in maximal normal form, $L(\varphi_r) \cup A$ is a disjunct of ψ, in contradiction to $L(\varphi_i) \notin C(\psi)$. □

B.3 Proof of the theorem characterizing super strict synonymy

Theorem 5.3.4 (Characterization of super strict synonymy)**.** We have for all $\mathcal{L}_p$-sentences φ and ψ that

$$\varphi \approx_{sstr} \psi \text{ iff } \begin{cases} At(\varphi) = At(\psi) & \text{, and} \\ \varphi \Leftrightarrow_{FDE} \psi & . \end{cases}$$

In other words, $\approx_{sstr}$ is the maximally coarse-grained binary relation on $\mathcal{L}_p$ satisfying overlap and FDE-preservation.

Proof. The right-to-left direction is immediate by induction on $\mathcal{S}^*$-proofs: For $\mathcal{S}^*$-axioms $\varphi \equiv \psi$ we have that $At(\varphi) = At(\psi)$ and $\varphi \Leftrightarrow_{FDE} \psi$ (because φ and ψ are strictly synonymous), and these two properties are preserved by the $\mathcal{S}^*$-rules.

So let's concentrate on the left-to-right direction. We first introduce the notion of a standard *maximal positive* disjunctive form. A standard disjunctive form φ was

defined in definition 5.2.3 and it is said to be maximal positive if:

(i) For every disjunct φ_i of φ, there is an $A \subseteq At(\varphi)$ and a minimal disjunct φ_0 of φ (i.e., there is no disjunct φ_0' of φ such that $L(\varphi_0') \subsetneq L(\varphi)$) such that $L(\varphi_i) = L(\varphi_0) \cup A$, and

(ii) if φ_i is a disjunct of φ and $p \in At(\varphi)$, then $\varphi_i \wedge p$ is a disjunct of φ (modulo the order of the literals).

We show that every $\mathcal{L}_p$-sentence φ is $\mathcal{S}^*$-provably equivalent to a standard maximal positive disjunctive form φ_{mpNF}. Fine (2016a, theorem 18, p. 215) shows that φ is AC-provably (and hence $\mathcal{S}^*$-provably) equivalent to its maximal disjunctive normal form φ_{mNF}. Similar to lemma 5.2.2, we show that the replacement rules also hold for $\mathcal{S}^*$ (essentially because in $\mathcal{S}^*$ we can prove that $\neg\big(\varphi \vee (\varphi \wedge \psi)\big) \equiv \neg\big(\varphi \vee (\varphi \wedge \neg\psi)\big)$). Now, let $\varphi_1, \ldots, \varphi_r$ be the minimal disjuncts of φ_{mNF}. Then every disjunct φ' of φ_{mNF} is of the form $\varphi' = \varphi_i \wedge L$ (modulo ordering) for an $i \leqslant r$ and a (possibly empty) set of literals occurring in φ. By using replacement, axiom $(A12^*)$, and idempotence several times, we can $\mathcal{S}^*$-provably replace each $\varphi_i \vee \varphi'$ by $\varphi_i \vee (\varphi_i \wedge At(L))$ and thus end up with a formula φ^* that still is $\mathcal{S}^*$-provably equivalent to φ.[130] Clearly, φ^* satisfies (i), and it also satisfies (ii): Let φ' be a disjunct of φ^* and $p \in At(\varphi^*)$. Then $\varphi' = \varphi_i \wedge At(L)$ for an $i \leqslant r$ and a set L of literals occurring in φ_{mNF}, and p occurs in a literal l_p of φ_{mNF} (since $At(\varphi^*) = At(\varphi_{mNF})$. Then, by the maximality of φ_{mNF}, $\varphi_i \wedge (L \cup \{l_p\})$ is (modulo order) a disjunct of φ_{mNF}. By our replacement process, $\varphi_i \wedge At(L \cup \{l_p\}) = \varphi_i \wedge (At(L) \cup \{p\}) = \varphi' \wedge p$ is a disjunct of φ^*. Consequently, φ^* is a maximal positive normal form of φ.

Now we claim that for all $\mathcal{L}_p$-sentences φ and ψ in standard maximal positive normal form we have

$$\text{If } At(\varphi) = At(\psi) \text{ and } \varphi \Leftrightarrow_{FDE} \psi, \text{ then } \varphi = \psi. \tag{5.9}$$

Before proving this, let's see that it finishes the proof. Assume $At(\varphi) = At(\psi)$ and $\varphi \Leftrightarrow_{FDE} \psi$. We have $\vdash_{\mathcal{S}^*} \varphi_{mpNF} \equiv \varphi$ and $\vdash_{\mathcal{S}^*} \psi_{mpNF} \equiv \psi$. Moreover, we've seen in the right-to-left direction that $\mathcal{S}^*$-equivalence entails overlap and FDE-preservation, so we have

$$At(\varphi_{mpNF}) = At(\varphi) = At(\psi) = At(\psi_{mpNF}), \text{ and}$$
$$\varphi_{mpNF} \Leftrightarrow_{FDE} \varphi \Leftrightarrow_{FDE} \psi \Leftrightarrow_{FDE} \psi_{mpNF}.$$

[130] To be a bit more precise: Say $\varphi' = \varphi_i \wedge \bar{p} \wedge \neg q_1 \wedge \ldots \wedge \neg q_m$. Then, by maximality, $\varphi_i \wedge \bar{p} \wedge \neg q_1 \wedge \ldots \wedge \neg q_{m-1}$ is a disjunct of φ_{mNF}, too. By axiom $(A12^*)$, $\mathcal{S}^*$ proves that

$$(\varphi_i \wedge \bar{p} \wedge \neg q_1 \wedge \ldots \wedge \neg q_{m-1}) \vee \Big((\varphi_i \wedge \bar{p} \wedge \neg q_1 \wedge \ldots \wedge \neg q_{m-1}) \wedge \neg q_m\Big)$$
$$\equiv (\varphi_i \wedge \bar{p} \wedge \neg q_1 \wedge \ldots \wedge \neg q_{m-1}) \vee \Big((\varphi_i \wedge \bar{p} \wedge \neg q_1 \wedge \ldots \wedge \neg q_{m-1}) \wedge q_m\Big),$$

so we can replace the formula to the left of $\equiv$ which is a subformula of φ_{mNF} by the formula to the right and obtain an $\mathcal{S}^*$-equivalent formula φ_1.

We continue this process with $\varphi_i \wedge (\bar{p} \wedge q_m) \wedge \neg q_1 \wedge \ldots \wedge \neg q_{m-1}$ by using the disjunct $\varphi_i \wedge (\bar{p} \wedge q_m) \wedge \neg q_1 \wedge \ldots \wedge \neg q_{m-2}$ that was in the original φ_{mNF} and still is in φ_1. So we can $\mathcal{S}^*$-provably replace $\neg q_{m-1}$ by q_{m-1} and obtain φ_2. We continue until we replaced all the $\neg q_j$'s by q_j's.

And if this replacement process applied to another $\varphi'' = \varphi_i \wedge \bar{p} \wedge \neg\neg r$ also requires a disjunct $\varphi_i \wedge \bar{p} \wedge \neg q_1 \wedge \ldots \wedge \neg q_k$, then we first add a copy of this disjunct to the current φ_j (which we $\mathcal{S}^*$-provably can do by idempotence) and then use one of them to replace $\neg q_k$ with q_k.

To not be overly tedious, we omit a fully detailed proof of this fact.

Hence we have by (5.6) that $\varphi_{\mathrm{mpNF}} = \psi_{\mathrm{mpNF}}$. Since $\mathcal{S}^* \vdash \chi \equiv \chi$, $\mathcal{S}^*$ indeed proves

$$\varphi \equiv \varphi_{\mathrm{mpNF}} \equiv \psi_{\mathrm{mpNF}} \equiv \psi.$$

So it remains to show (5.9). We use an argument similar to the proof of lemma 5.2.5. Let $\varphi = \varphi_1 \vee \ldots \vee \varphi_n$ and $\psi = \psi_1 \vee \ldots \vee \psi_m$ be the sentences in standard maximal positive normal form. By standardness, it suffices to show that $C(\varphi) = C(\psi)$, that is,

$$\{L(\varphi_1), \ldots, L(\varphi_n)\} = \{L(\psi_1), \ldots, L(\psi_m)\}.$$

Assume for contradiction that not. Then, without loss of generality, there is a φ_i ($i \leqslant n$) such that $L(\varphi_i) \notin C(\psi)$ (the other case is analogous). Without loss of generality, we may assume that $L(\varphi_i)$ is $\subseteq$-minimal with this property (otherwise take the minimal one). That is,

$$\forall j \leqslant n : L(\varphi_j) \subsetneq L(\varphi_i) \Rightarrow L(\varphi_j) \in C(\psi). \tag{5.10}$$

We consider two cases. Case 1: For all ψ_k ($k \leqslant m$) we have $L(\psi_k) \not\subseteq L(\varphi_i)$. Then, as in the proof in section B.2, we get a contradiction to $\varphi \Leftrightarrow_{\mathrm{FDE}} \psi$.

Case 2: There is a ψ_k ($k \leqslant m$) such that $L(\psi_k) \subseteq L(\varphi_i)$. We again can take $L(\psi_k)$ to be minimal, that is,

$$\forall j \leqslant m : L(\psi_j) \subseteq L(\psi_k) \text{ and } L(\psi_j) \subseteq L(\varphi_i) \Rightarrow L(\psi_j) = L(\psi_k). \tag{5.11}$$

Again, if there is no φ_r such that $L(\varphi_r) \subseteq L(\psi_k)$, we get – as in the proof in section B.2 – a contradiction to $\varphi \Leftrightarrow_{\mathrm{FDE}} \psi$.

So assume that there is a φ_r such that $L(\varphi_r) \subseteq L(\psi_k)$. Without loss of generality, φ_r is minimal in φ. Since $L(\varphi_i) \neq L(\psi_k)$ (otherwise φ_i would be in ψ), we have $L(\varphi_r) \subseteq L(\psi_k) \subsetneq L(\varphi_i)$, so by (5.10) we have $L(\varphi_r) \in C(\psi)$. So by (5.11), $L(\varphi_r) = L(\psi_k) \in C(\psi)$.

Now, write $A := \big(L(\varphi_i) \setminus L(\varphi_r)\big)$. We consider two cases. Case (a): A doesn't contain a negative literal. Then $A \subseteq At(\varphi)$. Hence, since ψ is maximally positive and $\varphi_r = \psi_k$ and $At(\varphi) = At(\psi)$, we have by clause (ii) that $\varphi_r \wedge A \in C(\psi)$. So $\varphi_i \in C(\psi)$, contradiction.

Case (b): A contains a negative literal. Then, since φ is maximally positive we have, by clause (i), that $\varphi_i = \varphi_0 \wedge A'$ for a minimal disjunct φ_0 of φ and an $A' \subseteq At(\varphi)$. We claim that $L(\varphi_0) \in C(\psi)$. If not, then we can reason as above that there must be a $\psi_{k'}$ and a minimal $\varphi_{r'}$ such that $L(\psi_{k'}) = L(\varphi_{r'}) \subseteq L(\varphi_0)$, so by minimality of φ_0, $L(\varphi_0) = L(\varphi_{r'}) \in C(\psi)$. Since $\varphi_i = \varphi_0 \wedge A'$ with φ_0 being a disjunct of ψ and $A' \subseteq At(\varphi) = At(\psi)$, we have by clause (ii) that $\varphi_i \in C(\psi)$, contradiction.[131] □

B.4 Proof of the "semantics with sets of scenarios" theorem

Theorem 5.3.5 (Semantics with sets of scenarios)**.** The following are equivalent

(i) $\varphi \equiv \psi$ is valid in truthmaker semantics (i.e. for every state model M we have $[\varphi]_M = [\psi]_M$).

[131] If I had more time, I would have simplified this proof.

(ii) $\vdash_{AC} \varphi \equiv \psi$

(iii) φ and ψ literally overlap and $\varphi \Leftrightarrow_{FDE} \psi$

(iv) $[\varphi]_C = [\psi]_C$ (where C is the canonical scenario model based on sets of scenarios).

Proof. $(i) \Leftrightarrow (ii)$. See Fine (2016a, theorem 21, p. 216). $(ii) \Leftrightarrow (iii)$. See our theorem 5.3.2. $(i) \Rightarrow (iv)$. Immediate. So we're done if we show $(iv) \Rightarrow (ii)$. Indeed, assume $[\varphi]_C = [\psi]_C$. Since φ is AC-provably equivalent to its standard maximal disjunctive form φ_{mNF}, and ψ to ψ_{mNF}, we have (since we've already shown $(ii) \Rightarrow (i) \Rightarrow (iv)$) that $[\varphi_{mNF}]_C = [\psi_{mNF}]_C$. This implies $\varphi_{mNF} = \psi_{mNF}$ by lemma 20 of Fine (2016a, p. 216) adapted to the canonical scenario model. Hence AC proves $\varphi \equiv \varphi_{mNF} \equiv \psi_{mNF} \equiv \psi$ □

BIBLIOGRAPHY

Anderson, A. R. and N. D. Belnap (1975). *Entailment: The Logic of Relevance and Neccessity, Vol. I*. Princeton: Princeton University Press.

Anderson, A. R., N. D. Belnap, and J. M. Dunn (1992). *Entailment, Vol. II*. Princeton: Princeton University Press.

Angell, R. (1977). "Three Systems of First Degree Entailment." In: *Journal of Symbolic Logic* 47, p. 147.

— (1989). "Deducibility, Entailment and Analytic Containment." In: *Directions in Relevant Logic*. Ed. by J. Norma and R. Sylvan. Dordrecht: Kluwer. Chap. 8, pp. 119–144.

Bader, R. M. (2013). "Towards a Hyperintensional Theory of Intrinsicality." In: *Journal of Philosophy* 110.10, pp. 525–563.

Balkenius, C. and P. Gärdenfors (1991). "Nonmonotonic Inferences in Neural Networks." In: *Principles of Knowledge Representation and Reasoning. Proceedings of the Second International Conference (KR91)*. Ed. by J. Allen, R. Fikes, and E. Sandewall. San Mateo: Morgan Kaufmann Publishers.

Barwise, J. (1981). "Scenes and Other Situations." In: *The Journal of Philosophy* 78, pp. 369–397.

Barwise, J. and J. Perry (1983). *Situations and Attitudes*. Bradford Books. Cambridge, Mass.: MIT Press.

Beall, J. and G. Restall (2005). *Logical Pluralism*. Oxford: Oxford University Press.

Berto, F. (2010). "Impossible Worlds and Propositions: Against the Parity Thesis." In: *The Philosophical Quarterly* 60.240, pp. 471–486.

— (2013). "Impossible Worlds." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2013. https://plato.stanford.edu/archives/win2013/entries/impossible-worlds/. Metaphysics Research Lab, Stanford University.

— (2017). "Impossible Worlds and the Logic of Imagination." In: *Erkenntnis*, pp. 1–21.

Besold, T. R., A. d'Avila Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kühnberger, L. C. Lamb, D. Lowd, P. Machado Vieira Lima, L. de Penning, G. Pinkas, H. Poon, and G. Zaverucha (under review). "Neural-Symbolic Learning and Reasoning: A Survey and Interpretation."

Besold, T. R., A. d'Avila Garcez, K. Stenning, L. van der Torre, and M. van Lambalgen (2017). "Reasoning in Non-probabilistic Uncertainty: Logic Programming and Neural-Symbolic Computing as Examples." In: *Minds and Machines* 27.1, pp. 37–77.

Bimbó, K. (2007). "Relevance Logics." In: *Philosophy of Logic*. Ed. by D. Jacquette. Amsterdam: Elsevier, pp. 723–790.

Bjerring, J. C. (2014). "On Counterpossibles." In: *Philosophical Studies* 168.2, pp. 327–353.

Bjerring, J. C. and W. Schwarz (2017). "Granularity problems." In: *The Philosophical Quarterly* 67.266, pp. 22–37.

Brogaard, B. and J. Salerno (2013). "Remarks on Counterpossibles." In: *Synthese* 190.4, pp. 639–660.

Cappelen, H. (2008). "Content Relativism and Semantic Blindness." In: *Relative Truth*. Oxford University Press, pp. 265–286.

Cappelen, H. and E. Lepore (2005). *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Blackwell Pub.

— (2008). "Shared Content." In: *The Oxford Handbook of Philosophy of Language*. Oxford: Oxford University Press. Chap. 40.

Carnap, R. (1947). *Meaning and Necessity. A Study in Semantics and Modal Logic*. Chicago: The University of Chicago Press.

Chalmers, D. J. (2002). "On Sense and Intension." In: *Philosophical Perspectives* 16, pp. 135–182.

— (2006a). "The Foundations of Two-Dimensional Semantics." In: *Two-Dimensional Semantics: Foundations and Applications*. Ed. by M. Garcia-Carpintero and J. Macia. New York: Oxford University Press, pp. 55–140.

— (2006b). "Two-Dimensional Semantics." In: *Oxford Handbook of the Philosophy of Language*. Ed. by E. Lepore and B. Smith. Oxford: Oxford University Press.

Chemla, E., P. Egré, and B. Spector (2016). "Characterizing Logical Consequence in Many-Valued Logics." Draft from http://semanticsarchive.net/Archive/GQzYTM4N/Chemla-Egre-Spector-LCrelations.pdf (last checked 5 October 2016).

Choe, Y. (2013). "Anti-Hebbian Learning." In: *Encyclopedia of Computational Neuroscience*. Ed. by D. Jaeger and R. Jung. New York: Springer New York, pp. 191–193.

Church, A. (1954). "Intensional Isomorphism and Identity of Belief." In: *Philosophical Studies* 5, pp. 65–73.

Clark, S. (2015). "Vector Space Models of Lexical Meaning." In: *The Handbook of Contemporary Semantic Theory*. Ed. by S. Lappin and C. Fox. 2nd ed. Sussex: Wiley-Blackwell, pp. 493–522.

Cobreros, P., P. Egré, D. Ripley, and R. van Rooij (2012). "Tolerance and Mixed Consequence in the S'valuationist Setting." In: *Studia Logica* 100.4, pp. 855–877.

Cresswell, M. J. (1975). "Hyperintensional Logic." In: *Studia Logica: An International Journal for Symbolic Logic* 34.1, pp. 25–38.

Dantsin, E., T. Eiter, G. Gottlob, and A. Voronkov (2001). "Complexity and expressive power of logic programming." In: *ACM Computing Surveys (CSUR)* 33.3, pp. 374–425.

D'Avila Garcez, A. S., L. C. Lamb, and D. M. Gabbay (2009). *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies. Berlin, Heidelberg: Springer.

Doets, K. (1994). *From Logic to Logic Programming*. Cambridge, MA: MIT Press.

Duží, M., B. Jespersen, and P. Materna (2010). *Procedural Semantics for Hyperintensional Logic: Foundations and Applications of Transparent Intensional Logic*. Dordrecht: Springer.

Eddon, M. (2011). "Intrinsicality and Hyperintensionality." In: *Philosophy and Phenomenological Research* 82.2, pp. 314–336.

Faroldi, F. L. G. (2016). "Co-Hyperintensionality." In: *Ratio*, forthcoming.

Ferguson, T. (2014). "A computational interpretation of conceptivism." In: *Journal of Applied Non-Classical Logics* 24.4, pp. 333–367.

Fine, K. (1994). "Essence and Modality." In: *Philosophical Perspectives* 8, pp. 1–16.

— (2012). "Guide to Ground." In: *Metaphysical Grounding*. Ed. by F. Correia and B. Schnieder. Cambridge, MA: Cambridge University Press, pp. 37–80.

— (2014). "Truth-Maker Semantics for Intuitionistic Logic." In: *Journal of Philosophical Logic* 43.2-3, pp. 549–577.

— (2016a). "Angellic Content." In: *Journal of Philosophical Logic* 45.2, pp. 199–226.

— (2016b). "Review of Steve Yablo's 'Aboutness'." Draft retrieved from http://www.academia.edu/15555407/Review_of_Steve_Yablos_Aboutness (last checked 27 May 2017).

— (2016c). "Truthmaker Semantics. Chapter for the Blackwell Philosophy of Language Handbook." Draft retrieved from https://www.academia.edu/10908756/survey_of_truthmaker_semantics (last checked 31 October 2016).

Føllesdal, D. (2004). *Referential opacity and modal logic*. Reprint of his 1961 PhD thesis. New York: Routledge.

— (2013). "Preface to the New Edition." In: Quine, W. V. *Word and Object*. 2nd ed. Cambridge, Massachusetts: The MIT Press.

Font, J. M. (1997). "Belnap's Four-Valued Logic and DeMorgan Lattices." In: *Logic Journal of the IGPL* 5.3, pp. 1–29.

Frege, G. (1891). *Funktion und Begriff*. Translated in Frege (1960). Jena: Hermann Pohle.

— (1892). "Über Sinn und Bedeutung." In: *Zeitschrift für Philosophie und philosophische Kritik* 100, pp. 25–50. Translated in: Frege (1948).

— (1948). "Sense and Reference." In: *The Philosophical Review* 57.3, pp. 209–230.

— (1960). "Function and Concept." In: *Translations from the Philosophical Writings Of Gottlob Frege*. Ed. by P. Geach and M. Black. 2nd ed. Oxford: Blackwell, pp. 21–41.

— (1979). "A brief Survey of my logical Doctrines." In: *Posthumous Writings*. Ed. by H. Hermes, F. Kambartel, and F. Kaulbach. Trans. by P. Long and R. White. Oxford: Blackwell, pp. 197–202.

— (2008[1891]). "Funktion und Begriff." In: *Funktion, Begriff, Bedeutung. Fünf logische Studien*. Göttingen: Vandenhoeck & Ruprecht, pp. 1–22.

French, R. (2017). "A Simple Sequent Calculus for Angell's Logic of Analytic Containment." In: *Studia Logica*, pp. 1–24.

Gioulatou, I. (2016). "Hyperintensionality." MA thesis. Amsterdam: Institute for Logic, Language and Computation.

Goodman, N. (1949). "On Likeness of Meaning." In: *Analysis* 10.1, pp. 1–7.

Hähnle, R. (1993). "A new translation from deduction into integer programming." In: *Artificial Intelligence and Symbolic Mathematical Computing: International Conference AISMC-1 Karlsruhe, Germany, August 3–6, 1992 Proceedings*. Ed. by J. Calmet and J. A. Campbell. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 262–275.

Harris, Z. S. (1951). *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

— (1954). "Distributional Structure." In: *Word* 10.2–3, pp. 146–162.

Hodges, W. (2001). "Formal features of compositionality." In: *Journal of Logic, Language, and Information* 10, pp. 7–28.

Hölldobler, S. and Y. Kalinke (1994). "Towards a Massively Parallel Computational Model for Logic Programming." In: *Proceedings of the ECAI94 Workshop on Combining Symbolic and Connectionist Processing, ECCAI*, pp. 68–77.

Humberstone, L. (2015). "Sentence Connectives in Formal Logic." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2015. http://plato.stanford.edu/archives/fall2015/entries/connectives-logic/. Metaphysics Research Lab, Stanford University.

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

Jespersen, B. (2010). "How hyper are hyperpropositions?" In: *Language and Linguistic Compass* 4.2, pp. 96–106.

Jespersen, B. and M. Duží (2015). "Introduction." In: *Synthese* 192.3, pp. 525–534.

Kencana Ramli, C. D. P. (2009). "Logic Programs and Three-Valued Consequence Operators." MA thesis. International Center for Computational Logic, TU Dresden.

King, J. C. (1995). "Structured Propositions and Complex Predicates." In: *Noûs* 29.4, pp. 516–535.

— (2016). "Structured Propositions." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2016. https://plato.stanford.edu/archives/win2016/entries/propositions-structured/. Metaphysics Research Lab, Stanford University.

Kleene, S. C. (1952). *Introduction to Metamathematics*. Amsterdam: North-Holland.

Kment, B. (2014). *Modality and Explanatory Reasoning*. Oxford: Oxford University Press.

Krakauer, B. (2012). "Counterpossibles." Dissertations. 522. PhD thesis. University of Massachusetts – Amherst.

Kripke, S. (2008). "Frege's Theory of Sense and Reference: Some Exegetical Notes." In: *Theoria* 74.

Kripke, S. A. (1980). *Naming and necessity*. Revised and enlarged edition. Library of philosophy and logic. Oxford: Blackwell.

Künne, W. (2010). "Sense, Reference and Hybridity: Reflections on Kripke's Recent Reading of Frege." In: *Dialectica* 64.4, pp. 529–551.

Van Lambalgen, M. and F. Hamm (2005). *The Proper Treatment of Events*. Vol. 4. Explorations in Semantics. Oxford: Blackwell.

LaPorte, J. (2016). "Rigid Designators." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2016. https://plato.stanford.edu/archives/

spr2016/entries/rigid-designators/. Metaphysics Research Lab, Stanford University.

Lau, J. and M. Deutsch (2016). "Externalism About Mental Content." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2016. https://plato.stanford.edu/archives/win2016/entries/content-externalism/. Metaphysics Research Lab, Stanford University.

Leitgeb, H. (2001). "Nonmonotonic Reasoning by Inhibition Nets." In: *Artificial Intelligence* 128, pp. 161–201.

— (2003). "Nonmonotonic Reasoning by Inhibition Nets II." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 11.101, pp. 105–135.

— (2005). "Interpreted Dynamical Systems and Qualitative Laws: from Neural Networks to Evolutionary Systems." In: *Synthese* 146.1, pp. 189–202.

— (2008). "An Impossibility Result on Semantic Resemblance." In: *Dialectica* 62, 293–306.

Lenci, A. (2008). "Distributional semantics in linguistic and cognitive research." In: *Rivista di Linguistica* 20.1, pp. 1–31.

Levy, A. (2002). *Basic Set Theory*. Mineola, N.Y.: Dover Publications.

Lewis, D. K. (1970). "General Semantics." In: *Synthese* 22, 18–67.

— (1973). *Counterfactuals*. Page numbers refer to the reissued version of 2001. Oxford: Blackwell.

— (1982). "Logic for Equivocators." In: *Noûs* 16.3, pp. 431–441.

— (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

Lindström, S. (1991). "Critical study: Situations and Attitudes." English. In: *Noûs* 25.5, pp. 743–770.

Malink, M. and A. Vasudevan (2016). "The Logic of Leibniz's Generales Inquisitiones de Analysi Notionum et Veritatum." In: *The Review of Symbolic Logic* 9.4, 686–751.

Mates, B. (1952). "Synonymity." In: *Semantics and the Philosophy of Language*. Ed. by L. Linsky. Urbana: University of Illinois Press.

Mendelsohn, R. L. (2005). *The Philosophy of Gottlob Frege*. Cambridge: Cambridge University Press.

Meŏstrović, R. (2012). "Euclid's theorem on the infinitude of primes: a historical survey of its proofs (300 B.C.–2012) and another new proof." In: *ArXiv e-prints*. arXiv: 1202.3670 [math.HO].

Moschovakis, Y. N. (1994). "Sense and denotation as algorithm and value." In: *Lecture notes in logic*. Ed. by J. Väänänen and J. Oikkonen. Vol. 2. Berlin: Springer, pp. 210–49.

— (2006). "A Logical Calculus of Meaning and Synonymy." In: *Linguistics and Philosophy* 29.1, pp. 27–89.

Nolan, D. (1997). "Impossible Worlds: A Modest Approach." In: *Notre Dame Journal of Formal Logic* 38.4, pp. 535–572.

— (2014). "Hyperintensional metaphysics." In: *Philosophical Studies* 171.1, pp. 149–160.

Pacuit, E. (2017). "Neighborhood Semantics for Modal Logic." Draft from March 21, 2017. Retrieved from http://web.pacuit.org/files/neighborhoods/nbhd-v6-PENULTIMATE.pdf on May 20, 2017.

Penco, C. (2013). "Indexicals as Demonstratives: On the Debate between Kripke and Künne." In: *Grazer Philosophische Studien* 88, pp. 55–71.

Pietz, A. and U. Rivieccio (2013). "Nothing but the Truth." In: *Journal of Philosophical Logic* 42.1, pp. 125–135.

Priest, G. (2005). *Towards Non-Being: The Logic and Metaphysics of Intentionality*. Oxford: Oxford University Press.

— (2008). *An Introduction to Non-classical Logic. From If to Is*. 2nd ed. Cambridge: Cambridge University Press.

Putnam, H. (1973). "Meaning and Reference." In: *The Journal of Philosophy* 70.19, pp. 699–711.

— (1975a). *Mind, Language and Reality*. Philosophical Papers, Volume 2. New York: Cambridge University Press.

— (1975b). "The Meaning of 'Meaning'." In: *Minnesota Studies in the Philosophy of Science* 7. Reprinted in Putnam (1975a, pp. 215-271)., pp. 131–193.

Quine, W. V. O. (1941). "Whitehead and the Rise of Modern Logic." In: *The Philosophy of Alfred North Whitehead*. Ed. by P. A. Schlipp. Evanston and Chicago: Northwestern University Press, pp. 127–163.

— (1951). "Two Dogmas of Empiricism." In: *Philosophical Review* 60.1, pp. 20–43.

— (1960). *Word and Object*. Cambridge, Mass.: MIT Press.

— (1969). *Ontological Relativity and other Essays*. New York: Columbia University Press.

Rey, G. (2016). "The Analytic/Synthetic Distinction." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2016. https://plato.stanford.edu/

archives/win2016/entries/analytic-synthetic/. Metaphysics Research Lab, Stanford University.

Ripley, D. (2013). "Paradoxes and Failures of Cut." In: *Australasian Journal of Philosophy* 91.1, pp. 139–164.

Rojas, R. (1996). *Neural Networks: A Systematic Introduction*. Berlin: Springer.

Russell, B. (1903). *The principles of mathematics*. Cambridge: Cambridge University Press.

Sahlgren, M. (2008). "The Distributional Hypothesis." In: *Rivista di Linguistica* 20.1, pp. 33–53.

Schaffer, J. (2009). "On What Grounds What." In: *Metametaphysics*. Ed. by D. Chalmers, D. Manley, and R. Wasserman. Oxford: Clarendon Press, pp. 347–383.

Schellenberg, S. (2012). "Sameness of Fregean sense." In: *Synthese* 189.1, pp. 163–175.

Schnieder, B. (2011). "A Logic for 'because'." In: *The Review of Symbolic Logic* 4.3, pp. 445–465.

Schroeter, L. (2017). "Two-Dimensional Semantics." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2017. https://plato.stanford.edu/archives/sum2017/entries/two-dimensional-semantics/. Metaphysics Research Lab, Stanford University.

Soames, S. (1985). "Lost Inoccence." In: *Linguistics and Philosophy* 8.1, pp. 59–71.

— (1987). "Direct Reference, Propositional Attitudes, and Semantic Content." In: *Philosophical Topics* 15.1, pp. 47–87.

Speaks, J. (2017). "Theories of Meaning." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. https://plato.stanford.edu/archives/spr2017/entries/meaning/. Metaphysics Research Lab, Stanford University.

Stalnaker, R. (1968). "A Theory of Conditionals." In: *Studies in Logical Theory, American Philosophical Quarterly*. Monograph Series, 2. Oxford: Blackwell, pp. 98–112.

— (1984). *Inquiry*. Cambridge, MA: MIT Press.

— (2002). "Common Ground." In: *Linguistics and Philosophy* 25, pp. 701–721.

Stanojević, M. (2009). "Cognitive Synonymy: A General Overview." In: *Facta Universitatis – Linguistics and Literature* 7.2, pp. 193–200.

Stenning, K. and M. van Lambalgen (2008). *Human Reasoning and Cognitive Science*. A Bradford book. Cambridge, Massachusetts: MIT Press.

— (2016). "Logic programming, probability, and two-system accounts of reasoning: a rejoinder to Oaksford and Chater (2014)." In: *Thinking & Reasoning* 22.3, pp. 355–368.

Tversky, A. and D. Kahneman (1981). "The Framing of Decisions and the Psychology of Choice." In: *Science* 211.4481, pp. 453–458.

Van Fraassen, B. C. (1969). "Facts and Tautological Entailments." In: *The Journal of Philosophy* 66.15, pp. 477–487.

Williamson, T. (2013). *Modal Logic as Metaphysics*. Oxford: Oxford University Press.

— (2016). "Counterpossibles." In: *Topoi*, pp. 1–12.

Yablo, S. (2014a). *Aboutness*. Princeton University Press.

— (2014b). "Aboutness Theory." Online appendix to Yablo (2014a) retrieved from `http://www.mit.edu/~yablo/home/Papers_files/aboutnesstheory.pdf` (last checked 31 October 2016).

Zalta, E. N. (1988). *Intensional Logic and the Metaphysics of Intentionality*. Cambridge, MA: MIT Press.