2 Two-dimensionalism and the necessary a posteriori

As the name suggests, two-dimensional semantic theories posit two dimensions of meaning. There are many versions of two-dimensional semantics, which differ concerning the scope of the theory, the nature of the two dimensions and how they are related. Some of these versions, for instance, are only applicable to a narrow range of linguistic expressions and/or do not treat the first dimension as a full-fledged semantic value. None of these accounts is suitable to (re-)establish the viability of conceptual analysis as a philosophical method. But there is one version of a two-dimensional theory of meaning which seems ideal for this purpose: According to what David Chalmers calls 'epistemic two-dimensionalism', every linguistic expression is connected with a semantic value which is a priori accessible to a speaker. In this book, I will therefore focus on this version of two-dimensional semantics (which I will henceforth simply call 'two-dimensionalism'), which is advocated by Chalmers (cf. Chalmers 2002b, 2004, 2006) and Frank Jackson (cf. Jackson 1998a, 1998b, 2004).[14] The following presentation of two-dimensionalism will draw heavily on their work.

2.1 Two-dimensionalism

2.1.1 Primary and secondary intensions

According to two-dimensionalism, every expression which is a candidate for having an extension is associated with two intensions, which correspond to two different ways of considering a possible world: One can consider it as actual, for example by asking questions like 'What if we find out / What if it turns out that the world is like this?'. Or one can consider it

[14] For a comprehensive discussion of other two-dimensional accounts cf. Chalmers 2006.

as counterfactual, by asking questions like ‘What if the world had been like this?’. Accordingly, the primary intension (or, in Jackson’s terminology, the A-intension) of an expression is a function from worlds considered as actual to extensions. Its secondary intension (or C-intension) is a function from worlds considered as counterfactual to extensions.

This second way of considering a possible world corresponds to a common understanding of various Twin Earth scenarios. As applied to Putnam’s original thought experiment, one asks: Given that the liquid in our rivers and lakes is H_2O, is a scenario where the rivers and lakes contain XYZ one in which they contain water? That is, the character of the actual world is taken to be fixed, and we are then invited to regard the hypothetical scenario in question as counterfactual. If we consider a world this way then no matter what kind of scenario we conceive of, we will always find that water is (nothing else than) H_2O. This suggests that secondary intensions are connected with metaphysical necessity, in the following way:

(2D1) A sentence S has a necessary secondary intension iff S is metaphysically necessary.

The secondary intension of an expression is thus simply its post-Kripkean intension: The secondary intension of ‘water’ picks out H_2O in all worlds, which implies that the secondary intension of ‘water = H_2O’ is true with respect to all worlds. But since it is not a priori that water is H_2O, secondary intensions need not be accessible a priori.

Primary intensions work quite differently. Let us suspend our empirical knowledge about the molecular structure of water for a moment and imagine a situation where we are just about to make a chemical analysis of various samples of water. At least from this point of view, it could turn out that water is XYZ. This indicates that if we consider a Twin Earth scenario not as counterfactual, but rather as actual we get a different result. The scenario corresponds to the epistemic possibility that water is XYZ. Here, ‘epistemic possibility’ has to be understood in a broad sense – it is compatible with everything we can know a priori. Primary intensions are thus tied to apriority:

(2D2) A sentence S has a necessary primary intension iff S is a priori.

However, it may still be a bit unclear what it means to say that the world could turn out to be a certain way and whether this notion is really connected with apriority. For example, when the schoolboy is asked to calculate 2^7, then from his point of view, it could turn out to be 64, or 256, or whatever. The same is true more generally for complex mathematical statements: Before we have determined their truth-value, they could turn out either way. What is thus required is an idealized notion of epistemic possibility, where something could turn out to be the case if it is not ruled out even by ideal rational reflection. According to Chalmers, a sentence S is true with respect to a world considered as actual, and thus epistemically possible, if and only if there is a D such that D epistemically necessitates S. Here, D stands for a canonical description of the world in question in a specific kind of vocabulary – I will say more about the characteristics of such a description later on. The notion of epistemic necessitation again involves an idealization. It can roughly be taken to mean that given D, ideal rational reasoning would lead a thinker to conclude S (or alternatively, that given D, a thinker *should* conclude S).[15] In this case, in Chalmers' terms the world which corresponds to D verifies S.

The two-dimensional account of meaning can be represented in a matrix. Here is the matrix for 'water is H_2O':

[15] Chalmers discusses this issue in much more detail, for instance in Chalmers 2006, sections 3.3 and 3.9.

'water is H_2O'	w_1	w_2	w_3
w_1 (WS: H_2O)	T	T	T
w_2 (WS: XYZ)	F	F	F
w_3 (WS: ABC)	F	F	F

(Figure 1)

The worlds on the left are worlds considered as actual; the worlds on the top are worlds considered as counterfactual. 'WS' stands for 'watery stuff' (this is again Chalmers' term, cf. e.g. Chalmers 1996, 57), which can be taken as an abbreviation of a description like 'the drinkable colorless liquid in rivers and lakes which sometimes falls from the sky in drop shape'. One can see that in w_1, this liquid's molecular structure is H_2O. w_1 can thus be taken to be the actual world. w_2 is like the Twin Earth world: Here, the molecular structure of the liquid which satisfies the description is XYZ; etc. Accordingly, the top row represents the secondary intension of 'water is H_2O', or maybe rather the set of extensions of the expression with respect to different worlds considered as counterfactual. 'Water is H_2O' thus has a necessary secondary intension, which is in line with Putnam's insight that water is necessarily H_2O. Obviously, to determine the secondary intension of the expression one has to have empirical knowledge. To stay in the picture, one has to know in which row one is located, i.e. one must know which of those worlds on the left really is the actual world.

As I said before, two-dimensionalists claim that there is another semantic value – another intension – which is accessible a priori. The basic idea is that we have (implicit) *conditional* a priori knowledge of the following kind: We know that if w_1 is the actual world, i.e. if the watery stuff in our

world is H_2O, then water is H_2O. We can also say that if w_2 – the Twin Earth world – is actual, then water is XYZ; etc. This putative a priori knowledge is mirrored in the diagonal from the top left to the bottom right of the matrix, which thus represents the primary intension. One can see that the primary intension of 'water is H_2O' is contingent, i.e. its extension varies from world to world. This reflects the fact that it is not a priori that water is H_2O – it is epistemically possible that water is XYZ, or ABC, or whatever.

From (2D1) and (2D2), one can straightforwardly infer that all necessary a posteriori truths have this two-dimensional structure. Thus:

(2D3) A sentence S is necessary a posteriori iff S has a contingent primary intension and a necessary secondary intension.

Likewise, one can infer the following for contingent a priori truths:

(2D4) A sentence S is contingent a priori iff S has a necessary primary intension and a contingent secondary intension.

2.1.2 Metaphysical plenitude and two-fold world dependence

Two-dimensionalism thus posits an a priori accessible dimension of meaning even for a term like 'water' which is among the prime examples invoked by semantic externalists and whose involvement gives rise to necessary a posteriori truths. Although the post-Kripkean (secondary) intensions of 'water' and of 'H_2O' pick out the same substance in every possible world, there is still an intuitive sense in which they do not have the same meaning. This intuitive difference in meaning is straightforwardly captured by two-dimensionalism, exploiting the fact that 'water' and 'H_2O' do not pick out the same substance with *epistemic* necessity. Thereby, two-dimensionalism establishes a semantic value in the tradition of Frege's sense.

However, even if that much is granted, if one wants to draw any conclusions concerning the viability of conceptual analysis one has to face an obvious objection. So far, I have always talked of the scenarios in the

first dimension, i.e. those which constitute the primary intension, as worlds. But calling them 'worlds', so the objection goes, is already misleading. We saw for instance that the primary intension of 'water' assigns to some of those so-called worlds XYZ as extension, or ABC, or … The lesson we should have learned from Putnam and Kripke is precisely that there are no worlds where water is XYZ.

One can of course construe the worlds in the first dimension as purely epistemic possibilities, bearing no deeper connection with metaphysical possibility. In fact, in more recent writings Chalmers does propose such a version of two-dimensionalism (cf. Chalmers 2006, section 3.4.2; Chalmers 2011). For some purposes, such an account may be useful – for example to model the cognitive value of a linguistic expression for the subject, i.e. from her subjective point of view. But it is questionable whether it can serve as a basis for doing conceptual analysis, as usually understood. I take it that the aim of conceptual analysis is to gain insight into what is *really*, i.e. metaphysically, possible or necessary. If for instance conceptual analysis can only reveal that it is epistemically possible that water is not XYZ (which is just another way of saying that it is not a priori that water is not XYZ), then this hardly teaches us anything about water – after all, two-dimensionalists usually agree that water could not have really been XYZ.

To illustrate this problem: Kornblith, as a reliabilist who is also one of the most vigorous critics of conceptual analysis (cf. Kornblith 1998, 2002, 2007a) can agree with an epistemic internalist that it is not a priori that knowledge is reliably produced true belief. It may for instance be epistemically possible that a reliably produced true belief is not knowledge or even that some version of an internalist account of justification is true. But still, Kornblith will insist that knowledge is necessarily reliably produced true belief. He says that as an epistemologist, he is not interested in our concepts of justification and knowledge, but in justification and knowledge themselves. Therefore, the mere epistemic possibility of his account being wrong does not concern him.

These considerations indicate that while the existence of a semantic value which is (a priori) accessible to a speaker is plausibly a necessary condition

for the viability of conceptual analysis, it is not a sufficient one. If the possibilities which make up the first dimension are construed as merely epistemic possibilities, then two-dimensionalism can hardly help to vindicate conceptual analysis as a valuable philosophical method. Thus, if one's aim is to defend conceptual analysis, one seems committed to holding that the scenarios involved in an expression's primary intension correspond to genuine possible worlds. Chalmers postulates such a correspondence in the following principle:

Metaphysical plenitude: For all S, if S is epistemically possible, there is a centered metaphysically possible world that verifies S. (Chalmers 2006, 82)

At first glance, this thesis stands in direct conflict with the conceded fact that sentences like 'water is XYZ' express epistemic possibilities, but not metaphysical possibilities. However, things are not that simple, as is witnessed by the fact that Kripke himself may quite plausibly be taken to endorse something at least roughly like this principle. I will discuss this issue in more detail in 2.2; for now I will just outline the basic idea. Contrary to what his considerations were taken to imply by many, Kripke argues in *Naming and Necessity* that whenever we conceive of something, what we conceive corresponds to some metaphysical possibility. It is just that sometimes, we have not conceived quite what we think we have (cf. Kripke 1980, 142ff.). In cases where rigid designators are involved, we are prone to fall prey to what has been called a 'proposition confusion' (cf. Stoljar 2006). That is to say, what we have conceived is not really, say, that water is XYZ, but something else. So what is this something else? Let us go back to the two-dimensional matrix of 'water is H_2O': There, w_2 was taken to be a world where the drinkable colorless liquid in rivers and lakes which sometimes falls from the sky in drop shape is XYZ. This surely represents a metaphysical possibility. Putnam's Twin Earth scenario itself describes precisely such a world, and it has rarely been suspected to be impossible.[16] Similarly, Kripke argues that when it seems to us that

[16] Except for the fact that no Twin-Earthling would be an exact duplicate of a person from Earth, since inhabitants of Twin Earth would largely consist of XYZ.

Hesperus could not have been Phosphorus, we really conceive of a world in which the brightest object visible in the morning sky is not identical to the brightest object visible in the evening sky – which could have surely been the case. The upshot is that the principle of metaphysical plenitude is compatible with Putnam's and Kripke's discoveries. The epistemic possibility that water is not XYZ corresponds to a genuine metaphysical possibility. It is just that this possibility is misdescribed by 'water is XYZ'.
Metaphysical plenitude raises another kind of problem in connection with indexicality. Which possible world verifies the epistemic possibility that I am famous, or that it is cold here, or that it is now 10 p.m.? It is well-known that even a complete objective characterization of a possible world is insufficient to settle all questions involving indexical expressions (cf. Castañeda 1967; Lewis 1979; Perry 1979). Therefore, the possible worlds in the first dimension should be construed as centered worlds. The notion of a centered world dates back to W.V.O. Quine (cf. Quine 1969). A centered world is simply a world with a marked individual at a time. 'I am famous' is thus verified by a centered world if the individual at the center of the world is famous; 'It is cold here' is verified if the individual at the center is located at a cold place; etc.

On the current reading of two-dimensionalism, the worlds involved in both the secondary and primary intension should be understood as metaphysically possible worlds. Both intensions thus range over the same space of possibilities, the only difference being that the worlds in the first dimension have a marked center. Accordingly, the difference between primary and secondary intensions is just the one mentioned at the beginning of this chapter: In the primary intensions, the worlds are considered as actual; in the secondary intensions, they are considered as counterfactual. This may still appear a bit puzzling: At least in some cases, the two intensions assign different extensions to the same possible world. But how is this possible, i.e. why should the extension of an expression with respect to a world be dependent so to speak on the perspective from which one considers the world? One way to explain this is by pointing out a

peculiar feature of names and natural kind terms, i.e. those terms which give rise to necessary a posteriori truths. I already mentioned that determining the extension of such terms with respect to a (counterfactual) world requires empirical knowledge. This shows that such a term's extension is dependent not only on the character of the world to be evaluated, but also on the actual world. In the case of names, this is explained by the fact that they are rigid designators: In every world, a name picks out the individual which it picks out in the actual world. Something similar is true for natural kind terms: In every world, they pick out the kind which they pick out in the actual world, regardless of the superficial properties of that kind in the world considered.

It may be worthwhile to have a closer look at the semantics of these terms in order to figure out what exactly it is about the actual world that determines their extension across all possibilities.[17] In the case of names, Kripke argued that their reference depends on causal chains. More precisely, he said that all that matters is the *actual* causal chain which connects *our* use of the name with its bearer. Whenever we use a name, we thus refer to the individual at the beginning of that chain. Natural kind terms seem to function quite similarly. They refer to the kind with which we are actually acquainted across all worlds. For this reason, the term 'water' does not refer to XYZ, even if we consider Twin Earth not just as a counterfactual world, but rather – like in Putnam's original scenario – as a remote planet in our galaxy. If this is correct, then names and natural kind terms behave similarly to indexicals: their extension with respect to a world is dependent on the actual referent which in turn depends on its relation to us, or more precisely to the speaker. In fact, Putnam himself argued in *The Meaning of 'Meaning'* that 'water' is an indexical (cf. Putnam 1975, 233f.; cf. also Haas-Spohn 1997). This suggests that the centering of the worlds in the first dimension is not only required to evaluate indexical expressions, but also for names and natural kinds. Imagine a world like the one described in Putnam's Twin Earth thought experiment. There are two

[17] I will try to spell out the primary intensions of names and natural kind terms in more detail in chapter 3.

‘watery stuffs’ on two different planets. In order to determine the extension of the term ‘water’, one needs more than an objective characterization of the world: One additionally has to know which of these substances we are acquainted with. Another thing to note in this context is that it is surely no coincidence that two-dimensional accounts have often been used to describe the semantics of indexical (or otherwise context-sensitive) expressions (cf. e.g. Kaplan 1989): In such accounts, the two-dimensional matrices are supposed to capture how content, i.e. the horizontal, varies with the context of use.
Two-dimensional semantics can straightforwardly capture the twofold world-dependence of specific kinds of terms. The two-dimensional structure reflects the fact that, as was just shown, the extension of an expression is dependent both on the world to be evaluated and on the actual world. In fact, the notion of a counterfactual world already contains this idea: A world can only be counterfactual relative to another – the actual – world. In the two-dimensional framework it is possible to hypothetically consider any world as actual. This can be useful since, although of course only one world is actual, without the relevant empirical information we do not know which of all the possible worlds we inhabit.
After what has just been said one can see why considering the same world in different ways can lead one to assign different extensions to it: It is because some terms make an implicit reference to the actual world, or to be more precise to our relation to the referent.

2.1.3 Scrutability and canonical descriptions

In my discussion of the two-dimensional matrix of ‘water is H_2O’, I mentioned that according to two-dimensionalism speakers have a specific kind of conditional knowledge: Although we do not know a priori that water is H_2O, we do know a priori that if the drinkable colorless liquid in rivers and lakes which sometimes falls from the sky in drop shape is H_2O, then water is H_2O. According to Chalmers and Jackson, this conditional

ability to identify the extension of 'water' given specific empirical information can be generalized:

(CJ) If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension. (Chalmers & Jackson 2001, 323)

In the background of this thesis is a very important idea which Chalmers calls the 'scrutability of truth and reference' (which may thus just be called the 'scrutability of extension') (cf. e.g. Chalmers 2002a). The idea is roughly that once we are given complete information about the character of the actual world, we are in a position to determine the referents of all of our expressions and the truth-values of all sentences.[18] Without any further specifications, this thesis does not sound very bold. If I am told for example that 'water' refers to H_2O, or that 'water is XYZ' is false, it is easy to determine the extension of 'water' or of 'water is XYZ'. To give the thesis its bite, one therefore has to say more about what 'sufficient' or 'complete' information amounts to. I already mentioned in passing that possible worlds can be characterized by the canonical description D. There are two basic constraints on the vocabulary used in D. Firstly, in order to avoid triviality it must be a limited vocabulary. And secondly, it must not contain terms like 'water' or 'Hesperus'. The reason for this qualification is that a sentence like 'the oceans are filled with water' may metaphysically determine the nature of the substance in the oceans; but it is epistemically compatible with there being H_2O, or XYZ, or … Consequently, the terms used in D must be such that their secondary application conditions are a priori accessible; let me call such terms 'epistemically transparent'. A term can only be epistemically transparent if its primary and its secondary

[18] Mainly due to potential problems regarding indeterminacies of reference, Chalmers is more confident in endorsing the scrutability of truth, though (cf. e.g. Chalmers 2002a, 174f.).

intension coincide;[19] let me call those terms whose two intensions are equivalent 'semantically neutral'.[20]

The scrutability thesis thus states that given a complete description of the actual world in a limited and semantically neutral vocabulary – let me call this description $D_@$ –, one is in a position to know all truths. Before I look at a proposal as to what such a vocabulary could look like, it still has to be clarified what it means to be in a position to know all truths. Once again, the notion involves an idealization: No actual subject would be able to grasp a complete description of our world, let alone draw all the required inferences from it. One can try to spell out what it means to be in a position to know something with the help of the notion of epistemic necessitation, which was introduced above: A given body of evidence E puts a subject in a position to know a true sentence S if E epistemically necessitates S, i.e. if given E and given ideal rational reflection, a subject would conclude S. The scrutability thesis can now be phrased as saying that for all S, if S is true, $D_@$ epistemically necessitates S.

In *Conceptual Analysis and Reductive Explanation* (2001), Chalmers and Jackson make a quite ambitious claim as to what the vocabulary used in $D_@$ might look like. They claim that a conjunction of the following kinds of information is sufficient to know all truths: microphysical information, information about phenomenal states, indexical information, and a concluding clause, which basically says that the world contains nothing beyond what has been stated in the description. Chalmers and Jackson call this conjunction PQTI (for **P**hysics, **Q**ualia, **T**hat's all, **I**ndexicals). It should be noted, however, that neither two-dimensionalism in general nor the scrutability thesis in particular are committed to the thesis that PQTI epistemically necessitates all truths. One could add a lot more (explicit)

---

[19] As I will argue in chapter 5, there is no such entailment in the opposite direction – there can thus be semantically neutral expressions which are not epistemically transparent.

[20] Here I deviate slightly from Chalmers' usage of the term, whose 'semantic neutrality' is rather supposed to capture what I call epistemic transparency (cf. his discussion in Chalmers 2006, 86ff.).

information to $D_{@}$ and still have, depending on one's purposes, a reasonably narrow (nontrivial) 'scrutability base'. There is also no compelling reason to assume that there is some privileged vocabulary which has to be used in $D_{@}$ – there may thus be several kinds of descriptions which satisfy the scrutability thesis.

Chalmers and Jackson hold that the scrutability thesis is not only true with respect to the actual world, but also with respect to other worlds which we can hypothetically consider as actual. Given the scrutability thesis, this seems to be a reasonable assumption: If the entailment from D to some S is a priori anyway, then why should it matter whether D describes the actual world or some other possible world?[21] This idea again goes nicely with the thesis that mastering a term or understanding a complex expression, i.e. grasping its primary intension, is connected with a specific kind of conditional a priori knowledge, as mentioned above: We are not only able to judge that since the watery stuff on Earth is H_2O, water is H_2O. We can also say that if the watery stuff on Earth had turned out to be XYZ, water would be XYZ, etc. That is to say that we have a priori knowledge of a kind that enables us to determine the extension of an expression with respect to different worlds considered as actual, if provided with sufficient nontrivial information about that world.

One minor complication is that if the relevant world is sufficiently remote from ours, then we need a different kind of vocabulary. PQTI is insufficient if a world contains (additional) nonphysical or nonphenomenal ingredients. Another thing to note is that it is most probably not a good idea to hold that there is a sensible scrutability statement with respect to all possible worlds. Why should there for instance not be worlds where only fundamental properties are instantiated – or even worlds with just one fundamental property? In this case, any description of the world would either have to explicitly describe this fundamental property, thus rendering the scrutability thesis trivial, or rather pointlessly somehow encode the information.

---

[21] Laura Schroeter nevertheless rejects this seemingly innocuous step. Her reasons for this will be discussed in chapter 4.

I just gave an outline of the two-dimensionalist framework, its key theses and its motivation. Before I turn to a more thorough discussion of the Kripkean and the two-dimensionalist accounts of the necessary a posteriori in 2.2, let me address two issues related to the two-dimensionalist framework which still deserve clarification: Firstly, the relation between two-dimensionalism and Jackson's descriptivism, and secondly, the notion of apriority implicit in the two-dimensionalist account.

2.1.4 Two-dimensionalism and Jackson's descriptivism

Jackson explicitly states that he adheres to a two-dimensional account of meaning (cf. e.g. Jackson 2004). However, elsewhere he says that he is committed to descriptivism (cf. e.g. Jackson 1998b). The combination of these views may seem surprising. In the following, I will try to spell out how they go together.

It is important to note that Jackson does not hold that the reference of an expression is determined by an associated description, where a description is understood as a linguistic expression. He understands descriptivism as the thesis that reference is determined by associated *properties*. But what exactly does this claim amount to? For a start, what counts as a genuine property is a hotly debated metaphysical question. But I think it is obvious that one should understand 'property' in a more liberal sense here, since there is no reason to expect that we associate words only with particularly 'natural' or fundamental properties. Jackson's take on this issue thus seems reasonable: On his view, what is required for a term to be associated with a property is just for the term to be associated with some kind of pattern, however grue-like (cf. Jackson 2007, 140; Jackson 1998b, 202). Given these considerations, let me propose to understand a property here simply as a set of extensions across possible worlds.[22] Taken this way, to say that an expression is associated with a property is equivalent to holding that it

[22] This is of course not to propose a general reductive account of properties.

has an intension, i.e. a function from possible worlds to extensions. Unlike in traditional descriptivist accounts, this thesis does not imply that the associated properties have to be expressible by a linguistic description. In fact, Jackson does argue that it should be possible to put the associated properties into words,[23] but I think this view should not be considered as an essential part of his descriptivism.

A descriptivist obviously wants to claim more than just that every linguistic expression has an intension. What is required is that the relevant property is associated with the expression by competent speakers. But it still has to be clarified what it means for a speaker to associate an expression with a property. According to Jackson, a speaker need not be able to articulate the associated properties. In most cases, her knowledge of these properties is only implicit. Jackson likens this to a card player's knowledge of the card game's rules or a competent speaker's knowledge of the rules of grammar (cf. Jackson 1998b, 211f.; Jackson 2004, 272f.; Jackson, Mason & Stich 2009). The player need not be able to spell out all the rules of the game, just like the average speaker would be at loss to give a detailed description of the grammar of her language. Nevertheless, they do follow the relevant rules. Take our average speaker: She is capable of forming grammatically correct sentences and of distinguishing grammatical expressions from ungrammatical ones. If we transfer this to the case of associated properties, then implicit knowledge would amount to being able to tell whether the expression applies to a particular situation or not, i.e. to determine its extension with respect to that situation.[24]

If one generalizes this idea and abstracts away from a subject's cognitive limitations, the resulting thesis is already quite close to Chalmers and Jackson's account of concept possession quoted above, according to which

[23] Cf. e.g. Jackson 2005. Cf. also my discussion in 7.2.

[24] Jackson holds that the competent speaker cannot only classify a given sentence as grammatical or ungrammatical, but that she can also give a reason for this; and apparently, he thinks that implicit knowledge of the associated properties is connected with a similar ability (cf. Jackson 1998b; Jackson 2004, 272). I will leave the question aside what exactly this ability amounts to.

possession of a concept enables a speaker to determine the concept's extension with respect to a hypothetical scenario. But there is still one more step to go: Chalmers and Jackson's thesis is clearly about worlds considered as actual, they speak about the speaker being provided hypothetical information about the actual world. For the descriptivist claim that speakers have implicit knowledge of the associated properties to have any chance of being correct, one would have to make the same restriction. This is because as we have seen, to determine the extension of an expression with respect to a world considered as counterfactual, one often requires empirical information which need not be accessible to the speaker. One could define two kinds of (associated) properties here: A primary (or primarily associated) property is a set of extensions across worlds considered as actual; a secondary (or secondarily associated) property is a set of extensions across worlds considered as counterfactual. Given metaphysical plenitude any primary property will be identical with some secondary property. Thus, these definitions should not be taken to mark a genuine metaphysical distinction, but rather as being purely epistemically and semantically motivated.

We can now formulate two theses concerning knowledge of associated properties, which should be considered as true by definition:

(KAP1) If a speaker has knowledge of the primary property associated with an expression, then she can determine its primary extension with respect to every possible world (or alternatively, she can determine its extension with respect to every world considered as actual), given a canonical description of the world and ideal rational reflection.

(KAP2) If a speaker has knowledge of the secondary property associated with an expression, then she can determine its secondary extension with respect to every possible world (she can determine its extension with respect to every possible world considered as counterfactual), given a canonical description of the world and ideal rational reflection.

Since speakers are often unable to determine a term's extension with respect to a world considered as counterfactual and thus do not know the secondary property associated with an expression, the descriptivist's thesis should be that a competent speaker knows the associated primary properties of the terms she uses. In fact, in *Why We Need A-Intensions*, Jackson himself invokes the notion of associated properties in connection with A-intensions, i.e. his version of primary intensions (cf. Jackson 2004, 264). What does this imply for Jackson's thesis that reference is determined by speaker associations? First of all, as a thesis about reference with respect to worlds considered as actual, it is already implied by two-dimensionalism. The same is true for reference in our world, since in the two-dimensionalist framework, the primary and the secondary extension of an expression with respect to the actual world are always identical. The thesis that speaker associations determine reference with respect to worlds considered as counterfactual seems clearly false, however, given the existence of necessary a posteriori truths. But one could hold a slightly weaker thesis, namely that speaker associations determine the reference of an expression with respect to a world considered as counterfactual in a context, or given a specified actual world with a marked center. Once again, this thesis is straightforwardly implied by two-dimensionalism.

To sum up: On its most plausible reading, Jackson's claim that competent speakers know the property associated with a given linguistic expression is equivalent to the thesis that they grasp its primary intension. His account thus offers a natural way to understand mastery of a term or possession of a concept, which involves the ability to determine the term's or concept's extension with respect to a given hypothetical scenario. On this interpretation, Jackson's variety of descriptivism should not be taken to exceed two-dimensionalism. Rather, it provides another way of framing some familiar two-dimensionalist theses about meaning, reference, and concept-possession.

2.1.5 Two notions of apriority

According to two-dimensionalism, a sentence is a priori if and only if it has a necessary primary intension, i.e. if and only if it is true with respect to every epistemic possibility. On the face of it, connecting apriority with epistemic necessity seems plausible: If a sentence is epistemically necessary, then it is in principle possible to determine its truth by going through all epistemic possibilities in one's mind. If, however, a sentence is not epistemically necessary, then its truth is dependent on contingent features of the world. It seems reasonable to think that it can thus only be known empirically. However, that latter claim is at least debatable. Let me illustrate this by means of a hypothetical case:
Suppose that a piece of chocolate which I put into my safe just a few hours ago is gone. The safe shows no signs of a violent break-in. I know that my room-mate, who likes chocolate a lot, is the only person who was in the room at that time. She is also the only person aside from me who knows the combination of the safe's lock. When I ask her, she tells me outright that she opened the safe and ate the chocolate. I thereupon come to believe that my room-mate took my chocolate. It is very plausible that (absent defeaters) this belief amounts to knowledge. Now take all the evidence I have in that hypothetical case, all my knowledge relevant to the judgment and, if necessary, also other relevant empirical background facts to build the antecedent of a conditional: 'If there is no chocolate in safe A at time T and there was chocolate in the safe two hours before T and the safe was subsequently locked and … and the laws of nature are such-and such …'. The consequent of the conditional is my judgment about the fate of the chocolate, i.e., that it was taken by my room-mate. If a subject is presented with that conditional, thinks carefully about it and concludes that it is correct, then her justification for the resulting belief is at least as strong as mine for my belief that my room-mate took the chocolate. But that subject's justification, unlike mine, is not based on empirical evidence: All the information relevant for justifying the conclusion is already in the antecedent of the conditional. Accordingly, it is natural to say that the

subject knows the conditional a priori.[25] At the same time, it is plausible that the conditional is not epistemically necessary. For example, the prevailing laws of nature may make it extremely unlikely that the chocolate suddenly dematerialized, but they do not exclude this possibility. Consequently, the conditional in question is knowable a priori, even though it is not epistemically necessary.

This result seems to raise a problem for two-dimensionalism which postulates an intimate connection between apriority and epistemic necessity, as seen above. An obvious way to solve that problem is to require stronger justification conditions for a priori knowledge. This would mean that the conditional I mentioned above would not be knowable a priori. Chalmers does indeed construe apriority as requiring 'conclusive' justification (cf. e.g. Chalmers 2006, 98). I do not like this idea, however. It is hard to see a reason to require stricter justification conditions for a priori knowledge (cf. also Casullo 2003, 33ff.; BonJour 1998, 110ff.). Moreover, since such an understanding makes it unnecessarily hard to acquire a priori knowledge, it could be detrimental to the project of defending the epistemic fruitfulness of conceptual analysis.

Let me therefore propose to distinguish two notions of apriority: The first one is equivalent to epistemic necessity, i.e. to having a necessary primary intension. One might call this notion strong apriority.[26] On the second notion, which one might thus call 'weak apriority' or 'standard apriority', a sentence (thought/proposition) is a priori if it can be known on the basis of justification which is independent of sense experience, where the relevant justification conditions are just as strong as those we impose on empirical knowledge. It is plausible, barring skepticism about for instance complex mathematical or logical truths, that if a sentence is strongly a priori, it is also standardly a priori, though the reverse does not generally hold. Both of these notions of apriority can be useful, depending on one's purposes. In

[25] We will encounter a similar move in an argument for a priori scrutability in 4.1.

[26] This notion is related, but not equivalent, to Hartry Field's 'strong aprioricity' which is tied to indefeasibility by empirical evidence (cf. Field 1996).

the following, I will briefly discuss which roles they could play within the two-dimensionalist framework.

Above I explained that according to two-dimensionalism, the grasp of an expression's primary intension bestows a subject with conditional a priori knowledge, i.e. the ability to determine the extension of the expression with respect to every world considered as actual a priori. Now it might be that the grasp of a term often, or even usually, enables a subject to determine the expression's extension conclusively, because the description of the scenario necessitates the extension. But I do not want to consider this as a requirement. I therefore hold that for the two-dimensionalist claim to be correct, it suffices that we are able to determine the extension of an expression with respect to a scenario in a standardly a priori way. Notice that this entails that conclusive justifiability and strong a priority can come apart, since it is possible that a sentence is epistemically necessary without being conclusively justifiable. For even if the truth of a sentence with respect to every scenario can be determined in a standardly a priori way, the justification we can have for at least some of these judgments may still be inconclusive. Let me therefore introduce the notion of super-strong apriority: A sentence is super-strongly a priori if and only if it is knowable with conclusive a priori justification.[27] It is plausible that just like strong apriority entails standard apriority, super-strong apriority entails strong apriority.

I just argued that two-dimensionalism should only require that the extension of an expression with respect to worlds considered as actual is determinable in a standardly a priori way. However, for other purposes, the notion of strong apriority is plausibly more relevant. It seems reasonable, for example, that for a feature F to be an a priori associated property of 'A' in Jackson's sense, F must be correlated with A with epistemic necessity, i.e. 'A → F' must be strongly a priori.

[27] This is the notion which may be equivalent to Field's notion of strong aprioricity (cf. Field 1996), cf. the preceding footnote.

For the majority of this book, the distinction between standard and strong apriority will not be important. I will mention it whenever it does become relevant.

So far, we have seen that two-dimensionalism promises to reconcile Kripke's and Putnam's insights with a broadly Fregean picture of meaning. But in itself, the existence of the two-dimensional framework does not suffice to establish such a picture, much less does it show that conceptual analysis is or can be a valuable philosophical method. I will address the questions of what exactly the two-dimensional framework shows and of what is still to be shown in 2.3. However, I presented it as a key motivation for adopting this framework that it can account for the Kripkean a posteriori necessities. So I think this is one thing which two-dimensionalism really needs to accomplish in order for the whole project to offer any prospects of success: It must be able to explain these a posteriori necessities in a way which leaves room for an a priori dimension of meaning and thereby establishes a connection between epistemic and metaphysical modality. I will therefore now turn to the question of whether a two-dimensionalist account manages to accomplish this task. It may seem curious that I will take Kripke's theory of modal error as the starting point for addressing this question. However, this approach draws its justification from the fact that the two-dimensionalist account can not only be understood as a reaction to Kripke's arguments, but also as a derivative of his modal epistemology. My general aims will be these: Firstly, I will carve out the similarities as well as the differences between the Kripkean and the two-dimensionalist explanation of a posteriori necessities. Secondly, I will point out what Kripke's account is and what it is not committed to, as compared with two-dimensionalism. And thirdly, at those points where Kripke's account faces objections, I will examine how two-dimensionalism can deal with them.

2.2 Modal illusions according to Kripke and according to two-dimensionalism

Due to the connection between meaning and modality, any account which tries to defend the tenability of conceptual analysis as an a priori method has to establish some kind of a priori access to what is possible and what is necessary. I outlined above how an intimate relation between modality and the a priori can be posited within a two-dimensional framework. In this respect, two-dimensionalism seems to be very much at odds with Putnam's and Kripke's views. Indeed, the lesson which Putnam drew from the existence of necessary a posteriori truths was to reject any kind of a priori access to modality. Kripke has often been assumed to take a very similar stance, not least by Putnam himself. But actually, as was briefly sketched above, many aspects of the two-dimensionalist explanation of the necessary a posteriori can already be found in Kripke's *Identity and Necessity* (1971) and in *Naming and Necessity* (1980).[28] In the following, I will outline Kripke's account of so-called 'modal illusions' in more detail and highlight its similarities with two-dimensionalism. I will then discuss some objections which have been raised and some which could be raised against Kripke's view. Insofar as these objections prove critical, I will examine how two-dimensionalism can fare against them.

Take a sentence like 'Hesperus = Phosphorus'. Since proper names are rigid designators, the sentence expresses a necessary truth. But still, it appears contingent. It at least seems conceivable that Hesperus is not identical with Phosphorus. Likewise, it seems as though it could have turned out that Hesperus is not Phosphorus. After all, for a long period of time in human history people believed that Hesperus and Phosphorus are not the same celestial body. Unlike many other semantic externalists, Kripke does not rest content with merely calling attention to the existence of such illusions of contingency. He tries to give an explanation for them. The first thing he points out is that sentences like 'Hesperus ≠ Phosphorus',

[28] In a later paper, Putnam admitted that he had earlier taken Kripke's position to be closer to his own than it actually is (cf. Putnam 1990).

though being metaphysically impossible, are nevertheless epistemically possible. This mere fact may account for the appearance of possibility. It is also an important part of the motivation for endorsing two-dimensionalism – after all, primary intensions are built precisely on the notion of epistemic possibility. However, saying that it is epistemically possible that Hesperus is not Phosphorus can just be taken as another way of stating that it is not (strongly) a priori that Hesperus is Phosphorus. As was pointed out above, this is of not much help to the proponent of conceptual analysis unless there is some kind of connection between epistemic and metaphysical modality.

Kripke's next step is to ask what it could mean to say that although it could have turned out that Hesperus is not Phosphorus, it could not have been the case. If it turns out or had turned out that Hesperus is not Phosphorus, would it not be / have been the case that Hesperus is not Phosphorus? Kripke accepts this line of reasoning and states that for every (metaphysical) impossibility I, it cannot turn out that I (cf. Kripke 1980, 141). He argues that when we say that it could have turned out that Hesperus is not Phosphorus, we are just speaking loosely. Something similar can be said about conceivability: When it seems conceivable that Hesperus is not Phosphorus, this is mere seeming – we do not really conceive of a situation where Hesperus is not identical with Phosphorus.

Kripke's way of putting things is disputable. Chalmers, for example, uses 'could turn out' purely epistemically throughout his writings. The same is true for Kripke's use of 'conceiving': Most writers have taken conceiving (and related notions) to be tied to apriority by definition (cf. e.g. Yablo 1993; Tidman 1994; Menzies 1998). However, these are just terminological issues. The important point is that according to Kripke, when a necessary falsehood appears to be conceivable, what we have really conceived is something else. And when we say, for some necessary falsehood I, that it could turn out that I, what could have really been the case is some I*. So the natural question to ask is: What is this I*, i.e. what is the genuine possibility which we really conceive?

2.2.1 Kripke's two models of modal error

Kripke offers two different models for explaining (or explaining away) these modal illusions. Yablo calls them the 'epistemic counterpart model' and the 'reference fixer model' (cf. Yablo 2006). Kripke illustrates the first of these models by means of his example of the wooden table (cf. Kripke 1980, 142): We may intuitively believe that this table could, contrary to appearances, turn out to be made of ice. But since facts about the origin and the constitution of a material object are necessary, this is impossible. That is, if the table is made of wood, it is essentially made of wood and it could thus not turn out / have turned out to be made of something else. However, it could have really been the case that there was

> *a table* looking and feeling just like this one and placed in this very position in the room, which was in fact made of ice. In other words, I (or some conscious being) could have been *qualitatively in the same epistemic situation* that in fact obtains, I could have the same sensory evidence that I in fact have, about *a table* which was made of ice. (Kripke 1980, 142)

That is to say, although there is a possible state of affairs involved, it is not about this particular table but rather about an 'epistemic counterpart' of it, that is a table which looks and feels just like it. It is certainly possible for a table made of ice to be an epistemic counterpart of a given wooden table. Moreover, there is at least some initial plausibility to the view that what is really conceived is just some table – after all, we cannot even distinguish that particular table from an epistemic counterpart. Thus, Kripke holds that when we seem to conceive of an impossible situation where this very table is made of ice, what we conceive is something possible: Namely that some table which looks like this wooden one is made of ice (cf. Kripke 1971, 160f. – though he talks about a lectern there).

Kripke introduces his second model by drawing on the already familiar example of the identity statement 'Hesperus = Phosphorus'. In his critique of descriptivism, he noted that although an associated description cannot give the meaning of a proper name, it may still be used to fix its reference. Take for example 'the brightest object in the evening sky'. The reason why this definite description cannot give the meaning of 'Hesperus' is that the

expressions are not modally equivalent: As was pointed out above, names refer rigidly, unlike the associated descriptions. But this does not mean that the description cannot fix the reference in the actual world, given that Hesperus really is the brightest object in the evening sky. Now if we replace the rigid designators in the metaphysically impossible sentence 'Hesperus ≠ Phosphorus' by corresponding reference-fixing descriptions, we get a sentence like 'the brightest object in the evening sky ≠ the brightest object in the morning sky' which is still false, but contingently so. This observation provides the basis for the reference fixer model. For any identity statement involving two rigid designators '$R_1 = R_2$', the reference of R_1 and R_2 can also be fixed by the descriptions D_1 and D_2, yielding the contingent '$D_1 = D_2$' (cf. Kripke 1980, 143f.). It is our confusing the former with the latter sentence which produces the illusion that 'Hesperus = Phosphorus' and the like are contingent. It should be noted that to create a contingent statement, it suffices to replace one of the rigid designators by a reference-fixing description – provided that the property by which the reference is fixed is not an essential property of the referent.

One can see some obvious similarities between Kripke's models and the two-dimensionalist treatment of necessary a posteriori truth. Both views concede that there are epistemically possible sentences which are metaphysically impossible, but argue that in each such case, there is a genuine metaphysical possibility nearby. Kripke's remarks thus suggest that he is committed to a principle quite similar to the thesis of metaphysical plenitude – the thesis that for every epistemic possibility there is a corresponding (centered) metaphysically possible world. And just like in the two-dimensionalist account, Kripke's explanation can be taken to imply that the seeming impossibilities are simply misdescribed: In the case of the wooden table, we mistake a situation where an epistemic counterpart of the table is made of ice for one where the table itself is. In the case of Hesperus and Phosphorus, the diagnosed misdescription is evident. We confuse the metaphysical possibility 'the brightest object in the evening sky ≠ the brightest object in the morning sky' with the impossibility 'Hesperus ≠ Phosphorus'.

2.2.2 Doubts about the accounts of modal error

2.2.2.1 Doubts about the epistemic counterpart model

Kripke's explanation of the modal illusions has been disputed. Stephen Yablo, for instance, thinks that although Kripke's models can account for some of the cases which he discusses, they cannot account for all of them (cf. Yablo 2006). The biggest part of Yablo's criticism is directed against the epistemic counterpart model. He begins by pointing out that being an epistemic counterpart of the wooden table can mean either of two things. It can mean that there is a counterfactual situation in which a table looks and feels to some individual just like the wooden table actually looks and feels to me. Or it can mean that there is a counterfactual table which looks and feels the same *to us* or *to me* as the wooden table. The first kind of scenario is certainly possible. If the individual in the counterfactual scenario has a neural (or whatever kind of) architecture which is sufficiently different from ours, any kind of table or non-table can look and feel to her like the wooden table does to us. But as Yablo remarks, this hardly explains the modal illusion. It is not very plausible that what we really imagine when we think we imagine this (wooden) table to be made of ice is, say, an 'ordinary' table made of ice which due to a counterfactual observer's extraordinary neural architecture looks to her like the wooden table looks to us. The proper reading of the epistemic counterpart model should thus be that there has to be a counterfactual table which would be an epistemic counterpart of the wooden table for us, i.e. including our actual neural architecture. Still, the case seems unproblematic on this reading, too. It is plausibly possible that a cleverly prepared table made of ice looks and feels just like a wooden one. However, Yablo tries to argue that the epistemic counterpart model fails with respect to another one of Kripke's examples, or at least a specific version of it.

An important set of a posteriori necessities which Kripke discusses are theoretical identifications, like 'heat = mean molecular kinetic energy'. Here, the epistemic counterpart model should be able to explain the intuition that heat could have turned out to be something else and point up

the possibility which is really conceived. Yablo grants that there could have possibly been an epistemic counterpart of heat, i.e. some phenomenon other than mean molecular kinetic energy which causes the same sensations in us. But let us consider a slightly more specific scenario. It seems likewise that heat (here understood as the opposite of cold) could have turned out to be low mean molecular kinetic energy, instead of high mean molecular kinetic energy (here, 'high' and 'low' are highly relative, of course). But the seeming possibility of 'heat = low mean molecular kinetic energy' cannot be accounted for by the epistemic counterpart model, according to Yablo. Kripke does note that it is contingent that mean molecular kinetic energy is felt as heat: Had our neural architecture been different, mean molecular kinetic energy could have caused very different sensations in us (cf. Kripke 1980, 133). This seems irrelevant, however. After what has been shown above, the required kind of epistemic counterpart of mean molecular kinetic energy would have to be felt as heat for us, with the neural architecture we have. But Yablo argues that it is impossible for mean molecular kinetic energy to be felt differently than it is felt, if we keep our actual neural architecture fixed. It is thus also impossible for low mean molecular kinetic energy to be felt as heat. If so, Kripke's model fails: There is no genuinely possible state of affairs to be found which is really conceived. It seems like the epistemic possibility of heat turning out to be low mean molecular kinetic energy corresponds to no metaphysical possibility.

If Yablo's argument succeeds, this is bad news not only for Kripke's view, but possibly also for the thesis of metaphysical plenitude, which is supposed to play a key role in the defense of conceptual analysis. In the following, I will therefore discuss in some detail how Yablo's challenge can be met. I will first identify two principled ways in which his 'fool's cold' case could be rejected and point out how Yablo would have to respond to these objections. Against this background, it will be possible to show straightforwardly that Yablo's critique does not threaten the two-dimensionalist account of modal epistemology. After that, I will argue that Yablo's critique does not provide a refutation of Kripke's model, either.

For a start, one could wonder whether 'fool's heat' – low mean molecular kinetic energy which is felt as heat – is not possible after all. The conceivability of fool's heat could even be used as a premise in a kind of 'inverted spectra' argument against physicalism. The acceptance of such a premise would hardly be a problem for either Kripke or Chalmers, both of whom in fact put forth quite similar arguments against physicalism. Yablo could respond by insisting that not only the actual physical basis of our neural architecture has to be held fixed in the counterfactual scenario, but also, say, the possibly partly non-physical laws which govern our brain processes. Thus understood, fool's heat is surely impossible. It is not clear that this is the most plausible way to interpret the fool's heat scenario. But I grant that it can be understood this way. Moreover, on this version fool's heat still seems conceivable. So Yablo's argument still poses a challenge.

Another way to attack the argument is to deny that fool's heat is conceivable in the relevant sense. Recall the student in class who is supposed to calculate 2^7. Before she switches on her pocket-calculator, in one sense the correct answer could turn out to be 256 – unless she has thought sufficiently hard about it. But given the idealized notion of epistemic possibility, '$2^7 = 256$' is not really epistemically possible, i.e. it cannot really turn out that way. Chalmers explicitly argues that only the idealized sense of epistemic possibility is of relevance here. This thought is also present in Kripke's writings: Kripke says that the modal illusion that Hesperus could have turned out to be distinct from Phosphorus goes deeper than the (putative) illusion that the four color theorem could turn out to be false (cf. Kripke 1980, 103).[29] So maybe 'heat = low mean molecular kinetic energy' only seems to be epistemically possible and thus is not conceivable given ideal rational reasoning.

There is a second way in which non-ideal epistemic conditions could be responsible for the appearance that fool's heat is possible: We simply do not know enough about our neural architecture to see that it is not

[29] By the time Kripke *Naming and Necessity* was published, the theorem had not yet been proven, so he did not know if it would turn out to be true.

compatible with low mean molecular kinetic energy being felt as heat. We certainly do not know anything about non-physical laws of nature which are involved in the emergence of phenomenal states, if there are such laws. One could thus even argue that we do not really understand the proposition we are supposed to conceive. Let me therefore try to spell out its exact content. The relevant hypothetical situation is one where our actual neural architecture is to remain fixed and in which low mean molecular kinetic energy is felt as heat. Since we do not know what our actual neural architecture is like, we cannot specify this aspect of the situation descriptively. The content of the proposition thus has to be just this: 'Our neural architecture is as it actually is and low mean molecular kinetic energy is felt as heat.' Presumably, this is unproblematic on Yablo's view. He remarks that modal judgments involving explicit reference to the actual world are quite common. One example he gives is 'There could have been less ivory-billed woodpeckers than there actually are' (cf. Yablo 2006, 332). Of course, this judgment is based on the belief that the ivory-billed woodpecker has not yet gone extinct. If Yablo is wrong about this, then so is his modal belief.

The considerations just made will help to answer the question of whether the alleged conceivability of fool's heat is capable of undermining the two-dimensionalist explanation of modal error: The first thing to note is that on Yablo's account, modal judgments can explicitly rely on empirical information. This is most obvious in his wood-pecker example. Accordingly, conceivability thus construed cannot be an a priori source of knowledge. Secondly, and relatedly, recall how hypothetical scenarios are evaluated in two-dimensionalism. The scenarios are given via a complete canonical description in a vocabulary which must not include terms whose evaluation is implicitly dependent on characteristics of the actual world. On this account, any description of a scenario containing the term 'actual' is obviously not suitable. This suggests that the modal intuition that fool's heat is possible, if it is spelled out the way it is by Yablo is irrelevant for the two-dimensionalist account. In a canonical description, the term

'actual' would have to be replaced by a complete specification of the subject's neural architecture.

One might object that this response is too simple. Given that 'actually'-involving modal judgments are quite common, as Yablo argues, is it not just ad hoc to exclude this kind of judgment? After all, they seem to be a source of modal error and if Kripke's account and two-dimensionalism do not deal with this type of modal error, so much the worse for them. However, I think this objection is ill-founded. Consider for instance the illusion that water could have failed to be H_2O. As was elaborated above, two-dimensionalism can be taken to explain this illusion, or respectively the fact that 'water is H_2O' is both metaphysically necessary and epistemically contingent, by pointing out that the term 'water' makes an implicit reference to the actual world. The solution to this problem has two parts. The first part involves the thesis that in each of these cases, there is a genuine possibility which corresponds to the impossibility. This possibility is revealed by focusing on the primary intension of the relevant expression. The second part of the solution is to prevent any possible misdescription of the conceived possibility by avoiding the use of terms whose secondary intensions differ from their primary intensions. Thus, the implicit dependence of certain terms on the character of the actual world is identified as the source of modal error, and the exclusion of those terms in a canonical description is an important part of the strategy to avoid such error. Accordingly, if one encounters a modal illusion and then tries to replace the (supposedly) misleading description of the conceived situation, it would not make much sense to replace it by one which makes *explicit* reference to the actual world.

Of course, this still leaves open the question what is really conceived when we deem fool's heat to be possible. It is certainly not what might be suggested by what I said about the requirements of a canonical description. That is, it is not what we get by replacing 'actual' in 'our neural architecture is as it actually is and low mean molecular kinetic energy is felt as heat' by a complete specification of that architecture. For firstly, such a situation would presumably no longer be conceivable given ideal

rational reflection. Secondly, it is not what we do or even something we can imagine, given our cognitive limitations and our empirical ignorance. What we actually imagine, I think, is something very unspecific. The details of course depend on how sketchy the conceiver's knowledge of the human brain is. In my case, it would probably be something like 'It is possible that some person has a brain weighting a bit more than one kilogram, consisting inter alia of a couple of billion neurons and even more glial cells, and experiences low mean molecular kinetic energy as heat.[30] This modal judgment is quite obviously correct, and I actually think this is sufficient to explain away the seeming possibility of fool's heat.

I have just shown that the modal illusion that low mean molecular kinetic energy could have been felt as heat does not threaten the thesis of metaphysical plenitude and the two-dimensionalist explanation of the necessary a posteriori in general. It is still worth checking whether it nevertheless threatens Kripke's epistemic counterpart model. There are at least two ways for Kripke to reject Yablo's conclusion. One is to deny that fool's heat is conceivable in the relevant sense. That is to say, once the relevant scenario is spelled out in sufficient detail, any appearance of possibility would be just due to the subject's cognitive limitations. The modal illusion that fool's heat is possible would then be no different in kind from the illusion that the four color theorem could have turned out to be false. Another way to deny the conclusion is to reject Yablo's interpretation of the fool's heat scenario. It is at least not completely clear that what is really required for there to be an epistemic counterpart of heat is that there is a phenomenon which is felt as heat for us as we actually are, up to every detail in our neural architecture. This need not imply that the possibility of a creature with a completely alien perceptual system experiencing low mean molecular kinetic energy as heat is sufficient to establish the possibility of fool's heat. One could also take an intermediary position and say for instance that what is required is that low mean molecular kinetic

[30] This description is still not very accurate since I do not know exactly what neurons, glial cells or molecular kinetic energy are.

energy is possibly felt as heat given all we currently know about our neural architecture.[31] It is not clear which of these lines Kripke should take, and of course it's even less clear which of them he would take. But in any case it is safe to say that the example of fool's heat does not provide an outright refutation of the epistemic counterpart model. However, I think there is another lesson to be learned here, which is not directly related to this particular example. If an epistemic counterpart is understood the way Yablo understands it, then in addition to the relevant object or property, there would also have to be a counterpart of the conceiver present in the scenario. Taken as a general requirement, this seems problematic for two reasons: Firstly, it makes the epistemic counterpart model's value as a heuristic device problematic. It is too demanding to require a subject to imagine a duplicate of herself, or to judge how she would experience some object or property given her neural architecture. And secondly, it has to be possible to talk about hypothetical scenarios where no observer is present. It seems sensible to ask for example if the wooden table would still look the same if all life in the universe had been extinguished. So the fools' cold scenario might be interpreted as a special case such that a counterpart of the subject has to be present in the scenario conceived. But it seems wrong to take the epistemic counterpart model to posit not only that the genuinely possible (and genuinely conceived) scenario has to contain a counterpart of what is conceived, but also a counterpart of the conceiver. In general, an epistemic counterpart of an object or property should thus be understood in a weaker sense: An epistemic counterpart of for example a table is an object which looks, feels, smells, … like the original table, without further qualification.

[31] The latter proposal would be in line with an understanding of a qualitatively identical situation as one in which all the evidence we have, sensory or not, is the same.

2.2.2.2 Doubts about the reference fixer model

Now let me turn to Kripke's reference fixer model. In many cases, this strategy for explaining away modal illusions seems to work just fine. Just replace the rigid designators flanking the identity sign in a sentence like 'Hesperus = Phosphorus' by contingent reference-fixing descriptions and you get a contingent sentence. The same seems to apply to 'heat = mean molecular kinetic energy'. If one replaces 'heat' by something like 'the phenomenon which causes heat sensations', the resulting sentence is plausibly contingent. However, if we understand the latter case the way Yablo proposes, this immediately raises some problems. As was just discussed, Yablo holds that many of our modal judgments involve reference to the actual world. I do not want to deal with the question of how specific examples should be interpreted. I will grant that there are modal judgments which refer to features of the actual world and just deal with the general question of whether and how such cases can be accounted for by the reference fixer model.

So, to stick with the current example, what if the question we are concerned with is not just whether molecular kinetic energy could have failed to cause heat sensations, but whether it could have failed to cause such sensations in us as we actually are? One possible solution proposed by Yablo himself amounts to 'rigidifying' the relevant description, in our case yielding 'the actual cause of heat sensations'. But in his view, this would not work either (cf. Yablo 2006, 338). He thinks that the reference fixing description would have to be a 'piece of language'. However, a token of 'the actual cause of heat sensation' uttered in some counterfactual scenario does not refer to the actual world, but to the counterfactual scenario itself, so this does not help. And even if there was a description which referred to the actual world no matter which world it was uttered in, then such a description would not be understandable in a counterfactual scenario.

I do not think this objection is decisive. The main thing it teaches us is that reference-fixing descriptions should just be understood as (linguistic) types associated with expressions by speakers. One should thus not require that a

token of them has to be present in a counterfactual scenario. After all, rigid designators can also be evaluated with respect to possible worlds which do not contain any token of the expression – as in 'Aristotle could have never existed'. There is no reason why Kripke should be committed to the existence of a token of the relevant description in the counterfactual scenario.
However, it is easy to see that invoking rigidified reference-fixing descriptions does not help anyway, since the expression which is supposed to replace the original one is not contingent, either. Thus, for example 'the phenomenon which actually causes heat sensations ≠ mean molecular kinetic energy' is necessarily false and does not help to explain away the modal illusion. It is thus crucial to take care that the reference-fixing description which replaces the rigid designator is really contingent. Accordingly, the solution to the problems raised by Yablo's examples does not lie in a modification of the reference fixers. Rather, one should proceed like in the case of the epistemic counterpart model discussed above: Any explicit reference to features of the actual world has to be obliterated and replaced by a description which explicitly specifies these characteristics. There is no apparent reason why this solution should not be available here as well.

The reference fixer model can only be successful if the statement which replaces the original one plausibly represents a situation which is really imagined by the subject. I.e., it does not suffice that for a given rigid designator there exists a reference-fixing description which turns a false possibility statement like 'heat could have failed to be mean molecular kinetic energy' into a true one like 'the phenomenon which causes heat sensations could have failed to be mean molecular kinetic energy'. The reference-fixing description also has to be accessible to the subject and it has to be strongly associated with the term in question. With this in mind, the reference fixer model seems even closer in spirit to two-dimensionalism than the epistemic counterpart model. Where two-dimensionalism holds that our evaluation of an expression with respect to a possible world is

guided by our grasp of a primary intension, Kripke's model seems to imply that we evaluate these worlds by relying on reference-fixing descriptions which we associate with the relevant expressions. The extent of these similarities is quite surprising, given the general picture of meaning Kripke is otherwise committed to. Precisely this serves as the starting point of the critique of George Bealer (cf. Bealer 2006) and Christian Nimtz (cf. Nimtz 2007). They argue that the reference fixer model is not compatible with Kripke's anti-descriptivism. And indeed, this model seems to commit Kripke to holding that speakers have access to descriptions which fix the reference of the terms they use.[32] But this is at odds with Kripke's own so-called arguments from Ignorance and Error: Kripke himself pointed out that speakers frequently do not know anything which could determine the reference of a given term. Nevertheless, they do refer when they use it. Take Alvin who uses the name 'Gell-Mann'. The only thing he believes about Gell-Mann may be that he is a famous physicist. He therefore does not know anything to distinguish Gell-Mann from any other famous physicist like, say, Richard Feynman. Or take Batu who does have more specific beliefs about Gell-Mann. She believes that Gell-Mann is the famous physicist who developed the theory of quantum electrodynamics. But as it happens, she is wrong: It was Feynman who developed quantum electrodynamics. Nothing that Batu associates with the name can thus determine its reference, either. For her belief is still a false belief about Gell-Mann, not a true one about Feynman. Thus, when she utters the name 'Gell-Mann', she refers to Gell-Mann, just like Alvin. If this is correct, then it seems that speakers need to know nothing which determines the reference of the terms they use. But then there is no basis for the reference fixer model. In cases of Ignorance or Error, there is no suitable reference-fixing description to be found which can replace the relevant rigid designator, in order to explain away the modal illusion.

[32] This does not seem to be Bealer's line of argument, however. In fact, I think that his critique of Kripke at this point relies on a failure to distinguish between associated descriptions which give the meaning and those which merely fix the reference of an expression.

There are two possible conclusions which could be drawn from this. One could either hold that since Kripke himself is committed to saying that speakers have access to reference-fixing descriptions, maybe there is some room for a moderate version of descriptivism after all. This is in effect Nimtz' proposal. Or one could conclude that Kripke's reference fixer model is flawed, since it is based on an idea which has been refuted by Kripke himself.

In fact, I think both of these conclusions are premature. Kripke does not claim that the reference fixer model can be applied to all cases of modal error. So he could just say that the model applies in those cases where speakers do have access to an appropriate reference- fixing description. But things are not quite as simple for Kripke who is committed to a strong thesis concerning the scope of his models. For in his argument against materialism, he argues that since the seeming possibility of pain without C-fiber activity cannot be explained away by one of his models, there could have been pain without C-fiber activity (cf. Kripke 1980, 144ff.). Thus, for his argument to work, it must be possible to explain away all kinds of seeming modal error. We saw above that it is not altogether clear whether the epistemic counterpart model can be applied to all kinds of cases. Now if Kripke's second model is only of limited range as well, the ambition to explain away all kinds of modal illusions may not be satisfiable. Fortunately, as Alma Barner points out, the range of the reference fixer model can be greatly extended with a small modification (cf. Barner ms.). One should not require that a rigid designator is replaced by a description which suffices to determine the term's reference. It suffices if it is replaced by some description which represents what the speaker actually does associate with the term. Take for example the modal illusion that Richard Feynman could have been Murray Gell-Mann's brother. In the case of Alvin, we get 'Some famous physicist could have been the brother of some (other) famous physicist'; in the case of Batu it could be 'The author of 'Ivanhoe' could have been the brother of the inventor of quantum electrodynamics'. Both of these sentences clearly express possibilities. But do they really explain the modal illusion? It might be argued that especially

in the first case, the allegedly imagined possibility is much too unspecific and the truth of the modal judgment is thus trivial. I actually think this rather speaks in favor of the current proposal, though. If what a speaker associates with a term is very unspecific, it would be odd if the resulting scenario was much more detailed. For the model to be psychologically adequate, the replacing description has to reflect the subject's state of knowledge, or ignorance.[33]

It is plausible that the modified 'reference fixer model', which now has to be considered misnamed is applicable to all kinds of cases of modal error.[34] This requires only that a speaker associates something with a given term. These associations do not even have to be semantic. They must only be sufficiently strongly connected with the term to make it plausible that they represent what the subject has in mind when she conceives of a corresponding situation. Thus understood, the model is compatible with Kripke's other theoretical commitments. At the same time, this clearly distinguishes his view from two-dimensionalism which posits that what a speaker associates with a term enables her to determine the term's reference with respect to every possible world considered as actual. Since two-dimensionalism cannot appeal to Kripke's model of modal error in this respect, it has to deal with the arguments from Ignorance and Error. I will discuss the question of whether the mastery of a term, or the possession of a concept, really amounts to grasping its primary intension in detail in chapter 3.

---

[33] It has been objected that in some cases, Kripke's model is inadequate because it does not provide a *de re* possibility, for instance by Janine Jones (cf. Jones 2004). A complaint to that effect could be raised here as well – it is not clear which philosophers the judgment of Alvin is about. However, I do not see why his judgment has to be about particular philosophers. I would again consider the fact that the description is just as unspecific as the speaker's ideas of Gell-Mann and Feynman as a virtue.

[34] Whether it can be used to explain away all kinds of modal error is a separate issue, though (cf. also 2.3).

2.3 Summary and outlook: What has been shown and what is yet to be shown

I think it has become clear that two-dimensionalism offers some very useful tools for a defense of conceptual analysis. But of course, the existence of the two-dimensional framework as such does not show that conceptual analysis can play a substantial role in philosophical inquiry. In the remaining part of this chapter, I will sum up what has been shown so far, and examine what is yet to be shown. This will also set the agenda for the following chapters. Some of the theses connected with two-dimensionalism go beyond what is required to offer conceptual analysis a theoretical foundation. I will thus also identify those parts of the two-dimensionalist account which seem dispensable with respect to my aims in this book.

A very important virtue of two-dimensionalism is that it offers a way to account for the Kripkean a posteriori necessities which still leaves room for an a priori dimension of meaning in the tradition of Frege. In the foregoing section, it transpired that two-dimensionalism does indeed offer a convincing explanation for the necessary a posteriori and the accompanying modal illusions, provided that there really are primary intensions associated with the relevant terms. However, the existence of primary intensions has not yet been established. Moreover, we saw that although Kripke's own theory of modal illusions is compatible with the existence of such a semantic value, it by no means presupposes it. The existence of primary intensions is obviously crucial for a defense of conceptual analysis. In the following two chapters, I will therefore try to dismantle a number of arguments against primary intensions and provide some positive reasons for positing them.

There is another question related to the modal illusions which is still open. Even if two-dimensionalism is able to explain the typical Kripkean a posteriori necessities, this does not imply that it can explain all of them.

That is to say, maybe metaphysical plenitude is not true after all – there could still be epistemic possibilities to which no metaphysical possibility corresponds. The question of whether there are such so-called 'strong necessities' has been most extensively discussed in the context of the debate on physicalism in the philosophy of mind. I am not planning to discuss these questions here in any detail – I will address the issue briefly in chapter 6, though. I think that the main motivation for denying that we have a priori access to modalities stems from Kripke's and Putnam's examples. Since examples of this kind are so common, they also pose a considerable threat to the prospects of conceptual analysis. Thus, if it is conceded that two-dimensionalism can explain away these typical cases, then any critique of conceptual analysis which invokes the necessary a posteriori no longer has much force. For even if there are local exceptions to metaphysical plenitude, this does not yet undermine the existence of a sufficiently reliable route from epistemic to metaphysical possibilities.

Another central component of two-dimensionalism is the scrutability of truth. The scrutability thesis could also play an important role in establishing the viability of conceptual analysis, which becomes apparent from the fact that, as was already pointed out in chapter 1, conceptual analysis is often done via thought experiments: From the description of a hypothetical scenario, we are supposed to judge whether the case described is a case of knowledge, or a good action, or a free choice, etc. I.e., we are supposed to determine the extension of a term with respect to a particular hypothetical scenario. The scrutability thesis can be taken to provide the theoretical foundation for our ability to judge such cases. Grasping a term's primary intension just means to be able to determine the term's extension when given hypothetical information about the world. Thus, two-dimensionalism provides a very straightforward rationale for the reliability of our judgments about hypothetical cases, basing it on our mastery of the relevant terms, or respectively on our possession of the relevant concepts. If successful, this would be an important achievement in itself, given the

ubiquity of the method of thought-experimentation in philosophical practice.

Against this background, it is quite remarkable that the most influential of Kripke's, Putnam's, and Tyler Burge's (cf. e.g. Burge 1979) arguments which are supposed to refute any Fregean account of meaning or content are themselves based on thought experiments – for example Kripke's 'Gödel and Schmidt', Putnam's 'Twin Earth', and Burge's 'arthritis' scenario. This might suggest another way to try and sustain the existence of primary intensions: One could argue that these arguments presuppose our ability to correctly evaluate hypothetical cases – and thus that they implicitly rely on the grasp of primary intensions. In fact, at a number of places Chalmers and Jackson seem to suggest precisely this (cf. e.g. Chalmers 2002b, 169; Jackson 1998b, 213). So do the arguments against semantic internalism really presuppose an internalist dimension of meaning? In my view, there is something to say in favor of this suspicion in the case of Putnam's 'Twin Earth'. Putnam gives a description of a hypothetical scenario with respect to which we are supposed to determine the extension of 'water'. One might think that our judgment that the liquid on Twin Earth is not water is based on our grasp of the primary intension of 'water' plus our empirical knowledge that the liquid in our lakes and seas is H_2O. However, it does not have to be understood this way. Firstly, the thought experiment does not really require us to determine the extension of water from a qualitative description of the scenario. We only have to infer that the liquid on Twin Earth is *not* water. To do this, it would suffice if for example we merely grasped some necessary condition for being water which XYZ fails to meet. And secondly, the inference in question – from the liquid on Earth being H_2O to the liquid on Twin Earth not being water – can still be taken to be based on empirical considerations, say, considerations from the history of science.

So what about Kripke's thought experiments? Take the afore-mentioned hypothetical case about Schmidt and Gödel: Kripke argues that it could turn out that it was not Gödel who discovered the incompleteness of arithmetic, but rather a man called 'Schmidt'. Gödel may have stolen it

from Schmidt and published it under his own name (cf. Kripke 1980, 83f.). For now, it need not concern us what the thought experiment is supposed to show.[35] The important questions here are how it works and what it presupposes. Chalmers argues in effect that Kripke's thought experiment conforms completely to the scrutability thesis (cf. Chalmers 2002b, 169): We are given a description of the hypothetical scenario and judge what the names 'Gödel' and 'Schmidt' refer to. Chalmers likens this to Gettier cases in the context of the analysis of knowledge. There, we are given a description of a scenario which we judge to be a case of justified true belief without knowledge. Thus, if Chalmers is correct, then far from providing a refutation of conceptual analysis, Kripke's thought experiment itself has to be considered as an instance of conceptual analysis. But again, things are not so simple. In fact, it would be odd if Kripke's thought experiment really worked as described by Chalmers. After all, Kripke does not tire to point out that possible worlds are not given to us qualitatively. In his view, we are not provided with a description of a world and then have to infer what is the case in that world. Rather, we just stipulate what is the case in a possible world (cf. Kripke 1980, 42ff.). On closer inspection, one can see that his 'Gödel/Schmidt' scenario is completely in line with this account: He does not provide a purely qualitative description of two persons from which we are supposed to judge who of them is Gödel and who is Schmidt. Rather, it is stipulated that it is Schmidt who discovered the incompleteness of arithmetic in the scenario, and Gödel who then stole and published it (cf. also Byrne & Pryor 2006, 50). So if these things are simply stipulated, then how do we know that things could really turn out to be as described in the scenario? I think the work in Kripke's argument is done by the following two intuitions: firstly, an intuition to the effect that discovering something is a contingent feature of a person; and secondly, the intuition that the names in question, or names in general, are used to refer to the same person with respect to every possible world. There is thus an important difference between Kripke's thought experiment and a typical Gettier scenario: In the latter case, it is not part of the description of the scenario that it is a case of

[35] I will return to this case in chapter 3.

non-knowledge. Therefore, it can be taken to rely on a conditional ability to determine the extension of our terms as expressed by the scrutability thesis, unlike Kripke's Gödel/Schmidt case. This does not mean that Kripke's account is in principle incompatible with the idea that we can infer the extension of a term with respect to a possible world from a qualitative description of that world. Even less has he shown that this cannot be done. But it is just false to say that he is committed to such a view.

Jackson often stresses that the 'method of cases' cannot refute two-dimensionalism even in principle (cf. e.g. Jackson 1998b, 213). For whatever one's verdict about a hypothetical scenario is, two-dimensionalism can take this judgment to be guided by our grasp of the relevant expressions' primary intensions. This may well be so.[36] But still, neither Putnam's nor Kripke's reliance on thought experiments has to commit them to (implicitly) assuming an internalist semantic value. This observation once more highlights that what is required is an independent defense of primary intensions, i.e. one which does not exclusively rely on the existence of the two-dimensional framework.

Since the scrutability thesis is quite ambitious, one may wonder if a proponent of conceptual analysis has to be committed to it. I pointed out before that there are many ways to construe the vocabulary used in the canonical description D, so the scrutability thesis need not be based on anything like PQTI. Moreover, the tenability of conceptual analysis surely does not depend on whether *every* truth is scrutable from the scrutability base. It has been argued for instance that there are true mathematical statements which are nevertheless not discoverable even by ideal rational reflection. In this case, they would not be epistemically necessitated by D, even though they are metaphysically entailed. But it is hardly plausible to derive an argument against conceptual analysis from the undecidability of certain mathematical statements. It is also questionable whether conceptual analysis has to assume that there is one basic vocabulary from which the

[36] It still has to be discussed how two-dimensionalism is able to handle the so-called epistemic arguments, which are also often based on hypothetical cases, though. This will be done in the following chapter.

other truths can be inferred. No actual cognizer is able to grasp a complete description of a world anyway, so for many practical purposes, this requirement seems irrelevant. In order to pursue conceptual analysis via the construction and evaluation of hypothetical cases, it is sufficient that we are able to determine the extension of an expression if given a description of a situation which does not explicitly use that term. Here, the described situation will typically only represent a tiny part of a possible world. And it is not even clear whether something like this can be done for each of our expressions, since there might be primitive expressions which defy any kind of analysis. Still, even though conceptual analysis need not be based on the scrutability thesis, this does not mean that an adherent of conceptual analysis should abandon the thesis. If correct, it may provide the theoretical foundation for conceptual analysis. And if it is combined with the idea that the grasp of a primary intension manifests itself in the respective ability to determine a term's extension with respect to a hypothetical scenario, then it can also help to give a reason for considering the intuitions elicited by thought experiments reliable. For these reasons, various versions of the scrutability thesis will play an important role throughout this thesis.

The scrutability thesis does not require that all terms are definable with the help of those in the scrutability base. This becomes apparent when one realizes that the point of many thought experiments is to undermine a proposed analysis of a specific term. Many have argued that it is impossible to give explicit analyses of most, or at least many philosophically interesting terms. These people often base their claim precisely on the fact that innumerable proposed analyses have been refuted by counterexamples. This arguably shows that it is possible to evaluate the invoked hypothetical scenarios even in the absence of any definition (cf. Chalmers & Jackson 2001, 320ff.). Given this, the fact that the primary intension, or the associated property, of a term does not have to be expressible by a description has to be considered a virtue of the theory. However, it may also be a reason to worry. If it turns out that it is impossible to extract any kind of analysis from a term's primary intension, then it becomes

questionable if primary intensions are of any use for conceptual analysis. The more general issue here is this: Even if it can be successfully argued that linguistic expressions are connected with primary intensions, it has not yet been shown that it is possible to gain any substantial philosophical insights via conceptual analysis. The second central aim of the following chapters – besides from the defense of the existence of primary intensions – will thus be to examine the practical epistemic value of conceptual analysis. There are two main issues to be addressed here: Firstly, how substantial are the a priori implications connected with a term – i.e., how much can be gained from the analysis of primary intensions? This question will mainly be discussed in chapter 5. And secondly, how is conceptual analysis to proceed – i.e., what could its method look like, and what are its aims? These questions will be the subject of chapter 6 and especially of chapter 7.