On What Matters in Defining Concepts

Denis Kazankov



Thesis submitted for BPhil in Philosophy
2019-2021 cohort

Abstract

Defining concepts has become an integral part of philosophical methodology. However, it isn't clear what the primary purpose of this activity is. There are two competing models answering this question: according to the Conservative Model, the primary purpose of defining concepts is to accurately report their actual meaning. According to the Revisionary Model, the primary purpose of defining concepts is to make a normative judgement about what they should mean, and, if necessary, revise their actual meaning accordingly. The objective of this thesis is to argue that the Conservative Model doesn't give us a satisfactory picture of what is at stake in defining concepts and that we should therefore rely more upon the Revisionary Model when working with concepts in philosophy.[12]

[1] I dedicate this thesis to the memory of my dog, Kiarka, and my grandmother, Olinka. You will both always be in my heart.

[2] I am indebted to my supervisor Prof Ofra Magidor for her invaluable insights and comments to the previous drafts of this thesis.

Table of Contents

Chapter 1 - Introduction: Two Models of Defining Concepts……5

Chapter 2 - Conservative Model Implemented: PK-Method……13

2.1. An Overview of the PK-Method……13

2.2. Clarification 1: PK-Method vs. Canberra Plan……15

2.3. Clarification 2: Essential Properties……...18

2.4. Clarification 3: The Implicit Metaphysical Burden of the PK-Method.....20

2.5. Clarification 4: Predicative Concepts vs Referential Concepts.....24

Chapter 3 - The Implementation Problem for the PK-Method.....…….....26

3.1. PK-Method as a Three-Step Procedure……..26

3.2. Problem 1: The Identification of a Paradigmatic Instance……..27

3.3. Problem 2: Competing Individuating Criteria……..34

Chapter 4 - Reasons for Defining Concepts……41

4.1. The Wrong Kind of Reason Objection……...41

4.2. The Central Function of Concepts……..45

4.3. The Constitutive Purpose of Defining Concepts……....54

4.4. Some Objections……...67

Chapter 5 - Normative Encroachment in Meta-Semantics……78

5.1. Temporal Externalism and Two Phenomena……78

5.2. Phenomenon 1: Determinacy....……..79

5.3. Phenomenon 2: Deference……...86

5.4. Towards an Alternative Explanation of Determinacy and Deference…….88

Chapter 6 - Conclusion……..99

Bibliography……101

No doubt, in accepting our words (up to a certain point) we sometimes succeed in identifying with them. Then we seem to express the truth. But this happens only the moment we step aside, the moment we break with our past and future in order to be the past and future of the word, the moment we become the silence of our five senses, ...

— Edmond Jabès, *The Book of Questions*

Chapter 1 - Introduction: Two Models of Defining Concepts

Defining concepts has become an integral part of philosophical methodology. However, it isn't clear what the primary purpose of this activity is and, relatedly, what kind of considerations philosophers should be responsive to when defining concepts. The main objective of this thesis is to argue against the standard view, according to which the primary purpose of defining a concept is to simply track its actual meaning irrespective of wider normative considerations about what it *should* mean.

Let me first set up the debate more carefully. Consider the following two definitions:

(1) 'Water' means H_2O.

(2) 'Knowledge' means JTB.

Both (1) and (2) are nominal definitions, i.e., the definitions whose definienda are concepts themselves (the concepts 'water' and 'knowledge', respectively). In this thesis, I interpret concepts simply as lexical representational devices: for X to be a concept in a language is for it to be a word with a certain representational function in that language.[3] For simplicity, I will henceforth call nominal definitions 'conceptual definitions'.

[3] This interpretation shouldn't be confused with other 'thicker' accounts of concepts as mental representations (Fodor, 1975; Margolis and Laurence, 2007) or semantic contents of words (Peacocke 1992; Zalta 2001).

Conceptual definitions are commonly contrasted with real definitions, i.e., the definitions whose definienda aren't concepts themselves but rather entities that concepts express, such as properties, relations or kinds.[4][5] For example, the real definitions corresponding to (1) and (2) are standardly taken to be (1*) and (2*), respectively.

(1*) Water is H_2O.

(2*) Knowledge is JTB.

Despite their form, (1*) and (2*) aren't equivalent to ordinary identity statements. Instead, they seek to describe *the essential properties* that together make water/knowledge what they are. To phrase it more generally:

(3) The real definition of the form 'G is F' is true iff its definiens F encapsulates all the essential properties of its definiendum G that together determine the identity of G.

This account of real definitions as specifying the essential properties of their definienda has been dominant in the history of philosophy ever since Aristotle formulated it when writing that 'a definition is an account (*logos*) that signifies an essence' (Topica, 102a3) before reiterating that same point in several other places.[6] I will discuss it in greater detail in §2.3..

[4] See Cargile (1991, p.21) and Gupta (2015, Section 1.1.).

[5] The distinction between properties and kinds as the semantic values of concepts is somewhat blurry. While properties are simply any entities instantiated by objects, 'kinds' are commonly associated with clusters of properties that are shared by objects based on some systematic similarities between them. See Boyd (1991) and Khalidi (2013) for a comprehensive discussion. Since the distinction is tangential to my arguments in this thesis, I will mostly talk about concepts expressing properties, except for natural kind concepts.

[6] See, for example, (Metaphysics, 1029b-1032b) or (Posterior Analytics, 73a34–5). Other historically influential expositions of this account of real definitions can be found in the works of Spinoza (Improvement of the Understanding, 95), Locke (Essay, III.iii.15-20, pp.414-420) or Moore (1942, pp.664-665).

The philosophical inquiry about conceptual definitions seems tightly connected to their corresponding real definitions. For example, when philosophers discuss concepts such as 'knowledge' or 'justice', what they ultimately investigate is what the property that these concepts pick out (or should pick out) actually is, i.e., what is knowledge, or what is justice.[7] Hence, we may hold that, even though conceptual inquiry is a part of philosophical methodology, philosophy is ultimately concerned with *worldly phenomena* that become salient to us by being conceptualised. In that case, discussions that philosophers pursue about conceptual definitions cannot be characterised as *merely* verbal disputes about concepts independent of what these concepts pick out (or should pick out) in the real world.[8]

The target problem of this thesis now enters the stage. Imagine that a linguistic community is formulating a definition of a concept in their language that will guide their understanding and use of it. The concept has already been in use for some time, and thus has a semantic value, such as a property it expresses. At the same time, based on their best normative theorising involving various considerations, the community reaches the judgement that the concept would be more valuable for them if it expressed a different property than the one it in fact expresses. What conceptual definition should the community implement in their language? Is it the definition that tracks the real definition of the property actually expressed by the concept? Or is it the definition that stipulates what property the concept should express? This dilemma boils down to the question as to what is the primary purpose of defining concepts already existing in our language. There seem to be two dominant models of defining concepts, each of which provides a different answer:

[7] See Robinson (1963, pp.150-193) for numerous examples from the history of philosophy supporting this claim.

[8] Accordingly, I agree with Williamson (2007, p.21) that philosophical inquiry isn't distinctively conceptual.

Conservative Model (CM): The primary purpose of defining a concept is to track its actual semantic value, i.e., the property that it *in fact* expresses. Whether this purpose is met is the only relevant consideration that cancels all other reasons for preferring one definition of the concept over the others.[9] The task of defining agents is, therefore, to first identify what specific property the defined concept expresses, find out what the real definition of this property is, and then to articulate the concept's definition accordingly. Therefore, the only definitions of concepts that defining agents should endorse are those that match the real definitions of what they actually express.

Revisionary Model (RM): The primary purpose of defining a concept is to make a normative judgement about what property the concept should express and, if necessary, revise its meaning accordingly. For each concept, there are several competing candidates for its definitions and different reasons that count for and against each of them. Whether a candidate definition of a concept tracks its actual semantic value is only one of many reasons that defining agents must weigh against one another when deciding which candidate definition to endorse; there is nothing that would make this reason systematically privileged over the others.

The difference between CM and RM is sometimes presented as the difference between *descriptive* and *stipulative* definitions of concepts. Descriptive definitions simply report what concepts actually mean. In contrast, stipulative definitions stipulate what concepts should mean and establish their meaning accordingly.[10] However, I prefer avoiding these labels here because

[9] Whenever I discuss reasons for defining a concept, I mean *normative reasons* concerning why an agent should or shouldn't define a concept in a certain way, as opposed to motivating reasons that merely explain agents' decisions. For the distinction between the two types of reasons see Dancy (2002) or McNaughton and Rawling (2018).

[10] See Horwich (2005, pp.138-146), Lycan (1991, pp.121-129), Shields (forthcoming) or Hale (2021, pp.2002-2004) for some discussion of stipulative definitions.

I consider them misleading for two reasons. Firstly, even CM allows that, when defining a new concept with no prior meaning, we stipulate it, and so it doesn't dismiss stipulative definitions altogether. Secondly, the label 'stipulative definition' creates the wrong impression that, according to RM, defining a concept consists simply of making an arbitrary command such as 'Let "X" mean Y', which may be completely insensitive to the prior meaning of 'X'. This might be the case if we introduce a completely new concept to our language, but my focus here is on *the already existing concepts with a prior meaning*. RM states that, when defining these concepts, we should be responsive to all kinds of considerations, including the concept's actual meaning. Moreover, the model allows that there might be cases when this consideration overrides all others, and thus the most rational thing to do is to retain the concept's actual meaning as it is. It only claims that this doesn't always have to be the case, and thus revising a concept's meaning might be likewise sometimes the most rational thing to do. Hence, as RM envisages it, the act of defining a concept is neither arbitrary nor insensitive to its prior meaning.

CM and RM may be considered to be complementary in some respects, since grasping what a concept actually means might help us in making an informed normative judgement over what it should mean. Still, at the end of the day, philosophers must decide which model to prioritise whenever they define a concept whose actual meaning isn't the best possible one: they must decide whether to revise its meaning or to just report it. The best illustration of this dilemma is provided by debates about controversial concepts such as 'woman'. Suppose, for example, that we know that the actual meaning of 'woman' doesn't anyhow reflect the various forms of injustice that accompany the classification of people as women and that revising its meaning to the form that reflects it will raise the awareness of its users about the injustice, which will

help to eliminate it. Taking these considerations into account, should we revise 'woman' or not?[11] Clearly, we cannot have it both ways.

CM is a standard model of the philosophical engagement with concepts that has been pursued by analytic philosophers for a long time. If CM is true, it imposes a substantive constraint upon how concepts should be defined in philosophy: each time philosophers disagree over which of the candidate definitions of a concept to accept, the two central questions that they need to address are what property the concept expresses and what its real definition is. All other theoretical and practical virtues that various definitions may exhibit are of secondary importance.

RM has recently also received a lot of attention due to the increased interest of philosophers in conceptual engineering and meta-linguistic negotiation.[12] Some of these philosophers think that assessing, negotiating and improving concepts in response to our normative judgements is a promising activity which should supplement or even replace CM in the methodology of philosophy. The core argument that underlies their view can be summarised as follows: since the actual semantics of concepts has evolved as a result of various random factors, it's very unlikely for it to be the best possible semantics that they can have. However, it's important to make sure that our language contains concepts with best the possible semantics, and therefore we must revise concepts' meanings, just as RM claims.[13]

[11] For some revisionist answers, see Haslanger (2000) or Jenkins (2016). For some conservative answers, see Saul (2006) or Mikkola (2009). Of course, debates about controversial concepts are more nuanced than that: for example, in the parallel debate about 'race', some argue that we should eliminate controversial concepts (Appiah, 1996), or substitute them with alternative concepts (Blum, 2002, pp.147-178; Glasgow, 2009, pp.133-154).

[12] See especially Burgess and Plunkett (2013a, 2013b), Cappelen (2018), various contributions to Cappelen et al. (2020), Plunkett (2015) or Plunkett and Sundell (2019). Also, RM is sometimes thought to trace back to Carnap's discussion of conceptual explication (1950, pp.1-18). See e.g., Cappelen (2017, pp.756-757), Brun (2020, pp.930-934) or Dutilth Novaes (2020).

[13] See Cappelen (2018, pp.39-48, 2020) or Eklund (2015, pp.376-378).

As a proponent of RM myself, I suspect, though, that this argument alone wouldn't convince most of the advocates of CM to adopt the revisionist approach to concepts. After all, they might retort that, even if a concept's meaning isn't the best possible one it might have, it might still be worth preserving as long as it isn't seriously defective in the sense of being internally incoherent or nonsensical. For example, suppose that the actual meaning of some purely theoretical concept from epistemology or metaphysics turns out to have bad practical effects which override all its theoretical virtues and thus isn't the best possible meaning that the concept might have. Still, they might legitimately argue that, insofar as the elimination of these effects by revising the concept will lead to the loss of its theoretical virtues, we shouldn't adopt RM.

Consequently, I believe that if we want to base the practice of defining concepts in philosophy more in line with RM than CM, some additional story about the problems with CM that will steer the debate toward a new course needs to be put forward. The objective of this thesis is to do so by arguing that, even under its strongest interpretation, CM doesn't give us a satisfactory picture of what is at stake in defining concepts.

The thesis is structured as follows: in Chapter 2, I explain what, to my knowledge, is the most compelling account of how the process of defining concepts according to CM looks like in practice: Putnam and Kripke's definitional method. This lays groundwork for my first argument against CM in Chapter 3. In that chapter, I show that it's problematic to implement CM to those concepts that are uninstantiated in the actual world or those expressing properties that aren't attributed to objects based on clear general criteria. In Chapter 4, I argue that CM cannot be defended on the grounds that defining a concept according to RM is to define it for the wrong kind of reason. The gist of my argument is that, as defining agents, we are committed

to associating such a property with the defined concept expressing which makes it an all-things-considered valuable representational tool. Finally, in Chapter 5, I reinforce my argument by responding to the objection that we have no independent reason to think that the semantics of concepts may be sensitive to normative facts about their all-things-considered value qua representational tools. I argue that allowing that such facts may encroach upon the semantics of concepts enables us to better explain some meta-semantic phenomena that are commonly thought to be best explained by temporal externalism about meaning.

Chapter 2 - Conservative Model Implemented: PK-Method

2.1. An Overview of the PK-Method

There is perhaps no theory that better shows how defining concepts is supposed to work according to CM than the one introduced by Putnam (1973, 1975) and Kripke (1980). Even though Putnam and Kripke originally introduced their theory with particular respect to the definitions of natural kind concepts such as 'water', there have also been some attempts to apply it to non-natural kind concepts.[14] In this chapter, I will summarise what the theory says and clarify its main tenets.

According to Putnam (1975 pp.153-173) and Kripke (1980, pp.106-134), the definition of 'water' (1) is formulated somewhat as follows: a community of speakers use 'water' to talk about the colourless and transparent liquid around them. This description gives them some basic grasp of what 'water' applies to in their language, but it doesn't yet specify its meaning. The community knows, however, that 'water' was introduced to their language by the statement (4) for the purpose of referring to a particular substance. They also know that 'water' still continues to refer to this substance and that the demonstrative expression 'these samples' in (4) picks out some paradigmatic instances to which they are strongly disposed to apply the concept.

[14] See Mallon (2017), Löhr (2021) or Haslanger (2003, 2006, 2012). Haslanger is an especially interesting case because while in her earlier work (2000) she clearly advocates for RM, arguing that we should ameliorate the definitions 'woman' or 'race', in these works, she relies more upon CM, arguing that we should debunk what 'woman' and 'race' *in fact* mean.

(4) □ ∀*x* (*x* is water ↔ *x* is a sample of dthat [the same substance that *these* samples are samples of]).[15]

Both Putnam (1975, pp.141-142) and Kripke (1980, pp.54-60) describe statements like (4) as 'ostensive definitions', but what they mean isn't that (4) literally defines water. Kripke (1980, p.60) would argue that the function of (4) is merely to make it the case that 'water' means H_2O but doesn't directly enter its semantics. Similarly, Putnam (1975, p.150) claims that (4) merely points out a standard for being water but doesn't analytically specify what water is. Nevertheless, ostensive statements like (4) play an important role in defining concepts, as defining agents can use them to identify what property their defined concept expresses even before they have enough information about its real definition.

The act of defining 'water' is then performed via the following method: first, the speakers identify paradigmatic instances to which they apply 'water' as the ones demonstratively denoted by 'these samples' in (4). Accordingly, they identify the kind that 'water' expresses as the one whose instances bear the *same substance* relation to these paradigmatic instances. Finally, once they discover that the paradigmatic instances are H_2O, they endorse the real definition of water, (1*), along with the definition of 'water', (1). Let's call this method the 'PK-method'. In what follows, I will do justice to the PK-method by making four clarifications regarding it.

[15] 'Dthat' functions here as an indexical operator which converts a singular term such as a description into a rigid designator. Semantically, it works as an incomplete demonstrative term which must be completed by a singular term 'α' relative to a context. Once 'α' completes it in a context, 'dthat[α]' rigidly designates the referent of 'α' in that context (Kaplan, 1989, pp.521-522).

2.2. Clarification 1: PK-Method vs. Canberra Plan

Each step of the PK-method seems to correspond to a different stage that we must go through when defining a concept. We start by identifying one or more objects to which we are disposed to apply the concept as its paradigmatic samples. This puts us in position to identify the property that the concept expresses as the same property as the one instantiated by these objects. Afterwards, we can start exploring why the given objects instantiate this property. If our inquiry is successful, it should reveal the property's real definition.

Indeed, our practice of defining concepts often seems to look like that. For example, when defining 'person', we first identify one or more entities that we are disposed to call 'person'. Our initial understanding of personhood becomes 'the same kind of entities as these entities are'. Then, we start exploring what makes those entities qualify as persons. If we are successful, we arrive at the real definition of personhood and formulate the definition of 'person' in accordance with it.

Since Putnam and Kripke propose this method specifically for natural kind concepts, in their original example defining agents arrive at the real definition of water through an empirical discovery about its paradigmatic sample. Yet, different epistemic activities might be involved in the process of defining other concepts. For example, in the case of 'person', in addition to gathering new empirical information about the entities to which we are disposed to apply the concept, we might also engage in analysing the information that we already possess. We might also consult our intuitions on possible cases (Jackson, 1998, pp.31-37) by imagining possible entities that are similar to the actual entities that we call 'person' in some respects while differing in others, testing our intuitions about whether the concept still applies to them.

To avoid misunderstanding, even though the PK-method might resemble the so-called 'Canberra Plan' method of analysing concepts[16] in its reliance upon our intuitions on possible cases, it's better to set these methods apart in the present discussion. My focus here is solely on the PK-method, as it seems to me to be a less controversial way to implement CM than the Canberra Plan. To explain, the Canberra Plan consists of two steps (Nolan, 2009, pp.268-270). In the first step, defining agents collect folk assumptions about the property that the defined concept expresses. What guides them are their observations about how competent speakers are disposed to apply the concept in various actual and possible cases. These assumptions then reveal the role that the property expressed by the concept plays in their theory of the world. In the second step, the agents look at their best scientific theory of the world to find what (if anything) realises the role of the given property specified in the first step. This two-step method has been perhaps the most widely discussed implementation of CM, but it involves two objectionable assumptions that the PK-method avoids.

Firstly, the proponents of the Canberra Plan commonly endorse two-dimensional semantics for concepts, arguing that folk assumptions associated with a concept amount to one dimension of its semantics known as 'A-intension'.[17] This dimension is supposed to remain stable independently of which world is actual and should be thus knowable a priori by competent users of the concept. For example, the A-intension of 'water' would be the set of folk assumptions based on which competent speakers apply the concept, such as 'a transparent colourless liquid substance that quenches thirst, falls from the sky, fills the rain and lakes, etc'. However, the claim that folk assumptions are a priori knowable and unaffected by variations

[16] The Canberra Plan is associated mainly with the works of Jackson (1998, 1994, 2004 (with Pettit)) and Lewis (1970, 1972). In addition, see also Braddon-Mitchell and Nola (2009).
[17] See Jackson (1998, pp.46-52), Chalmers and Jackson (2001, pp.322-323) or Braddon-Mitchell (2004).

across possible worlds is highly controversial. As Nimtz (2004, pp.138-139) points out, for most descriptions featuring in the above folk assumptions that guide our application of 'water', we can conceive of an object which fails to satisfy them, and yet we are still disposed to call it 'water'. Also, our folk assumptions are often fallible and thus subject to empirical refutation. For instance, when Aristotle used the concept 'fire', he thought that he was talking about a substance rather than a chemical process and would presumably correct himself in the light of empirical evidence disconfirming this assumption (Schroeter, 2004, pp.436-442).

Secondly, the agenda of the Canberra Plan is a naturalist one: to show how higher-order metaphysical properties described by our folk assumptions can be located in the world *as described by natural science* (Jackson, 1998, pp.2-4). Yet, arguably, only a few of those philosophers who have engaged in defining concepts according to CM believe that such a naturalistic account of metaphysical properties is in fact achievable. Moreover, many concepts seem to be such that even if we discovered the underlying natural realisers of their expressed properties, we wouldn't regard this discovery as being relevant to their semantics. Cohnitz and Haukioja (2020, pp.136-137) nicely illustrate this point on the following example: suppose we discovered that all unmarried men have a certain neural structure N. It seems that we would still be reluctant to conclude that 'bachelor' in fact expresses the property of having N. After all, if we came across a few married men with N, we wouldn't be disposed to call them 'bachelors'.

In contrast, the PK-method as outlined above isn't embroiled in these issues. Firstly, assumptions that competent speakers associate with a concept don't have to enter its semantics. Some of these assumptions might be only what Putnam (1975, pp.150-151) calls 'operational definition', i.e., the descriptions used to identify paradigmatic instances of the property

expressed by the concept but, just like the above-discussed ostensive statements, they aren't its real definition. As such, they don't have to be satisfied by every object to which we are disposed to apply the concept but may still serve their purpose in defining it. Others of these assumptions may prove to track the concept's meaning. However, in the course of the PK-method, defining agents may come to know that this is so by gathering new empirical information as in the case of 'water', by analysing already known information, or by reflecting upon their intuitions about the application of the concept to possible cases. Hence, the picture of defining concepts that drops out of the PK-method is much less constrained than the Canberra Plan because it isn't exclusively committed to a priori conceptual analysis. Additionally, nothing in the PK-method implies that a concept can be successfully defined only if we discover the natural realiser of its expressed property, and thus the method is also spared from the second problem. This makes the PK-method a significantly more acceptable and generalisable implementation of CM than the Canberra Plan.

2.3. Clarification 2: Essential Properties

If the PK-method is successfully implemented, the speakers defining 'water' should arrive at the real definition (1*). As specified by (3), (1*) says that H_2O contains all the essential properties of the kind *water* that jointly determine its identity. But what does it take to be an essential property of something? For a long time, it was standardly thought that the essential properties of an entity X are simply those necessarily instantiated by anything that is X. However, this account of essence lost traction after Fine (1994a) presented a compelling argument against it.[18] His argument consists of various examples of necessary but intuitively

[18] See also Fine (1994b, pp.56-58).

unessential properties, showing that our pre-theoretic understanding of what properties are essential to an object cannot be accurately captured simply in terms of metaphysical necessity. For example, the property of being water and such that 2+2=4 is necessarily instantiated by any sample of water but isn't essential to water. In that case, only those necessary properties of an object that meet an additional qualification are essential to it.

There has been a long-standing debate over what (if anything) is an appropriate qualification of essential properties, or over whether essence is a primitive concept explanatorily prior to necessity as Fine himself argues (1994a, p.9). Unfortunately, I cannot afford to fully cover this debate here.[19] At the same time, I believe that there is one intuitive characterisation of essence that we can rely upon in our discussion. In his later work, Fine (2015, pp.306-308) argues that real definitions of the form 'G is F' should be interpreted as equivalent to (5).[20]

(5) Being F is essential to being G, and being F fully grounds being G.

The grounding relation in (5) is a metaphysical explanatory relation: the fact *a* fully grounds the fact *b* iff *a* corresponds to those aspects of the world which together make it the case that b.[21]

I take it that (5) just rephrases my initial explication of real definitions, (3): the definiens F of the real definition of G is the property which is both essential to G and such that whatever

[19] Two notable proposals have been that essential properties are metaphysically sparse and joint-carving (Wildman, 2013; Cowling, 2013) or somehow intrinsically related to their instances (Denby, 2014, Livingston-Banks, 2017). Both qualifications, though, seem to be too narrow by not including some properties that seem pre-theoretically essential. For example, many have argued that some social and human kinds are essentially mind-dependent (Searle, 1995, pp.37-42; Sveinsdóttir, 2013). But being mind-dependent is neither joint-carving, nor intrinsic to the instances of these kinds.

[20] Two slightly different analyses of real definitions that also appeal to grounding have been recently independently developed by Correia (2017, pp.60-61) and Rosen (2015, pp.197-200).

[21] See Fine (2001, pp.15-16; 2012, pp.37-40).

is G is so *in virtue of* having it. In that case, it seems that F is a complex property encapsulating all simpler essential properties of G as its constituents. Accordingly, the other essential properties of G can be understood as the constituent properties of F. These properties *partially ground* G in the sense of non-redundantly contributing to the metaphysical explanation of G. To illustrate, let 'G' stand for 'water' and 'F' for 'H_2O'. The real definition (1) then says that being H_2O is both essential to and fully grounding of being water. Also, each constituent property of being H_2O, such as the property of consisting of molecules containing two hydrogen atoms or the property of being a chemical compound, is also essential to water because of partially grounding being water. Hence, the essential properties of an object seem to be those that somehow take part in its metaphysical grounding. While this characterisation of essential properties isn't completely transparent because grounding is a topic in its own right, it should suffice for our purposes.[22]

2.4. Clarification 3: The Implicit Metaphysical Burden of the PK-Method

Notice that the real definition (1*) can be derived from (4) and the premise about the chemical structure of a paradigmatic instance of water only insofar as the same substance relation consists in having the same chemical structure. In this respect, the PK-method, as originally presented, is slightly misleading since it makes it seem that the community is justified in accepting (1*) only based on (4) and their discovery that the paradigmatic sample is H_2O independently of any prior metaphysical assumptions. But, in fact, they can do so only with the hidden metaphysical belief about what being of the same substance as something consists in.

[22] See Audi (2012), Raven (2015) or Thompson (2016) for further discussion of the explanatory nature of grounding.

To show this point, Salmon (1981, p.166-167) unpacks the inference involved in the PK-method as follows:

(Premise 1) □ ∀*x* (*x* is water ↔ *x* is a sample of dthat [the same substance that these samples are samples of]).

(Premise 2) Being a sample of the same substance as something consists in having the same chemical structure with that thing.

(Premise 3) These samples have the chemical structure H_2O.

(C) □ ∀*x* (*x* is water ↔ *x* is H_2O).

Premise 2 is a hidden metaphysical belief here. I interpret this premise as specifying a general essentialist criterion for individuating objects as water. This is a slight digression from Salmon who doesn't analyse Premise 2 in terms of essence but merely as saying that, for any two worlds W_1 and W_2, a sample X in W_1 is a sample of the same substance as sample Y in W_2 iff X in W_1 has the same chemical structure as Y has in W_2 (Ibid., p.179). Yet, I believe that my interpretation is preferable. Salmon seems to have missed that, strictly speaking, C isn't equivalent to the real definition of water (1*) even though it follows from it. C only says that water and H_2O are co-intensional, i.e., equivalent across all possible worlds. Given Clarification 2 though, (1*), unlike C, says not only that water and H_2O are co-intensional but also that H_2O is the essence of water. The above argument can establish that this is so only under my interpretation of Premise 2.

Nonetheless, Salmon's main point seems correct: Premise 1 and Premise 3 cannot be used to establish either C or (1*) unless Premise 2 is presupposed. This is because unless the speakers are justified in believing that Premise 2 is true, they cannot rule out the possibility that being water is grounded in something different than its chemical structure. Moreover, Premise 2 certainly doesn't look like a claim that the speakers could infer from Premise 1 or Premise 3. Rather, it's an independent and non-trivial metaphysical claim that requires its own justification (Salmon, Ibid., pp.183-192). Consequently, the PK-method is a more epistemically demanding method than it might seem.

It might be objected that it's misleading to present Premise 2 as an independent premise, for the tight relation between the identity of substances and their chemical structures is built into the semantics of the concept 'substance' in Premise 1; and thus, whoever understands the meaning of Premise 1, knows that it semantically entails (6).[23]

> (6) $\Box \forall x$ (x is water $\leftrightarrow$ x is a sample of dthat [the chemical structure that these samples have]).

I have two things to say in response to this objection. Firstly, even if it's true that Premise 2 is built into the semantics of 'substance', it's doubtful that this makes the PK-method epistemically easier. (6) would by no means be a trivial, or obvious semantic consequence of Premise 1, and so it seems that the speakers would still need a strong justification for why the given entailment holds. Given what substantive metaphysical implications this entailment has, it seems unsavoury to explain it away by a modal anti-realist argument to the effect that it's

[23] This objection has been raised in Mackie (2006, pp.187-192). Similar attempts to establish a semantic link between Premise 1 and Premise 2 can be also found in Law (2016, pp.287-290) and Haukioja (2015).

simply a matter of general convention that any object to which 'substance' applies is individuated by its chemical structure unless something more is said about what justifies us in complying with this convention rather than any other.[24] Hence, the objection doesn't relieve the PK-method of its epistemic burden.

Secondly, even if ordinary speakers generally think that natural kind concepts apply to objects in virtue of their microstructure,[25] the thesis that *whoever* understands what 'substance' means necessarily accepts Premise 2 seems like just another of the problematic understanding-assent links which have been forcefully attacked by Williamson (2007, pp.73-133). It seems completely adequate to say that speakers who have never thought about the existence of chemical structure can still manifest a decent understanding of 'substance' simply by applying the concept roughly to the same objects as they would if they accepted Premise 2.

Furthermore, the historical evidence suggests that there can be competent users of chemical kind concepts who are aware of the existence of chemical structure and yet aren't immediately compelled to accept Premise 2. LaPorte (2004, pp.100-102), for instance, points to the historical examples of divergent decisions on how to define chemical kind concepts that the scientific community took when faced with the discovery that there existed a mineral which had the same chemical structure but a different colour from an already known mineral. In the case of ruby, the community decided to restrict the concept 'ruby' to the mineral variety of the composition Al_2O_3 with a red colour. By contrast, in the case of topaz, they allowed 'topaz' to refer also to blue minerals with the composition $Al_2SiO_4(F, OH)_2$ even though the concept had been associated with a yellow colour. Yet, it would be ridiculous to conclude that scientists are

[24] See Sidelle (1989) for the most prominent anti-realist arguments of this kind.
[25] See Jylkkä et al. (2009) for some supporting experimental evidence.

less competent users of 'ruby' than of 'topaz' only because they consider the colour of the kind expressed by the former concept to be a more important aspect of it than its chemical structure.

2.5. Clarification 4: Predicative Concepts vs Referential Concepts

Putnam and Kripke argue that 'water' is a referential concept that directly picks out a natural kind. Yet, concepts are also sometimes treated as predicates expressing properties or kinds attributed to the objects to which they apply. These properties are commonly theoretically represented as *intensions*, i.e., functions from world-times to the classes of objects instantiating a given property in those world-times.[26] According to the former interpretation, real definitions such as (1*) entail identity statements such as (7) instead of equivalence statements such as (8).

(7) Water = H_2O[27]

(8) $\Box \forall x$ (water(x)$\leftrightarrow H_2O(x)$).

(7) carries heavier ontological commitments than (8), as it doesn't just state a *de dicto* quantified biconditional bound by the necessity operator but is a *de re* statement which includes direct reference to kinds that commits us to the existence of the kind water and the co-intensional kind H_2O. Yet, setting aside this ontological difference, does it make any substantive difference to the plausibility of the PK-method whether we treat (1*) as entailing

[26] An important caveat here is that, strictly speaking, intensions are too coarse-grained theoretical proxies for properties that cannot capture some metaphysical distinctions between properties in terms of grounding or essentiality (Schaffer, 2009, p.364; Nolan, 2014, p.157). For example, being H_2O grounds being water but doesn't ground being H_2O even though water and H_2O are co-intensional. Also, as written above, while the property of being H_2O and the property of being water and such that 2+2 =4 are co-intensional, only the former is essential to water. Thus, it might be better to individuate them instead by hyperintensional semantics such as the one developed in Tichý (1988) or in Berto and Jago (2015, pp.162-184).

[27] For the same reasons as stated in the footnote 26, the identity relation = is here, strictly speaking, *the identity qua intension*.

(7) or (8)? One may think that it does so because the method relies upon the causal theory of reference, according to which what 'water' refers to is causally linked to the initial act of demonstrative reference to paradigmatic samples made when (4) was first articulated. Thus, it might be thought that the method is available only if we allow 'water' and 'H_2O' to be referential.

I disagree. It's true that if we treat a concept as a predicate, we cannot hold that the causal chain that is initiated by a demonstrative reference to its paradigmatic sample culminates in a referential use of that concept. However, as an alternative, nothing prevents us from holding that the chain culminates instead in its predicative use. For example, it's perfectly possible for (4) to restrict the predicative application of 'water' in all possible worlds only to those objects that bear the *same substance* relation to the given paradigmatic sample of water since it contains the demonstrative term 'this sample' which rigidly designates it. Just like referential uses can be causally linked to an object picked out in the initial 'reference-fixing' statement, predicative uses can also be causally linked to an object picked out in the initial 'predication-restricting' statement. Hence, the PK-method is compatible with the predicative interpretation of concepts. Since my objective here is to examine whether the PK-method is also applicable to the definitions of non-natural kind concepts many of which are standardly thought to be predicative, to keep things simple, I will henceforth treat 'water' as a predicative concept.

Chapter 3: The Implementation Problem for the PK-Method

3.1. PK-Method as a Three-Step Procedure

We now have a rough idea of what defining agents must do to successfully to define a concept through the PK-method:

> (Step 1) They must identify one or more paradigmatic instances of the property X that the defined concept expresses and tentatively identify X by a statement that specifies a sameness relation between any instance of X and the given instance(s).
>
> (Step 2) They must form a metaphysical belief that will inform them about what having X consists in, i.e., about the general essentialist criterion for individuating objects as X.
>
> (Step 3) They must use this criterion to learn, in virtue of the possession of what specific properties the paradigmatic instances qualify as X. These properties will together amount to the real definition of X.

In this chapter, I will argue that it's difficult to see how linguistic agents can successfully execute these three steps in the course of defining many philosophical concepts. My toy example of a concept for which the PK-method seems unsuitable is 'just society' (henceforth written as 'JS'), where 'just' stands for the comprehensive property of being just *as a whole in all relevant respects*. This seems like an excellent example for the purpose of testing the applicability of the PK-method to philosophical concepts. After all, 'JS' is arguably one of the most important philosophical concepts for which many definitions have been already proposed

in the literature. Therefore, if the PK-method is a useful method of defining concepts as envisaged by CM, defining 'JS' by it should be possible.

3.2. Problem 1: The Identification of a Paradigmatic Instance

Let me start with Step 1. This stage of the PK-method must be completed for the whole defining process to get underway. This is so because if defining agents don't have at least a rough grasp of what object(s) the defined concept applies to, they don't know what sort of property they should direct their focus towards in Step 2 and Step 3. There are two basic ways in which a paradigmatic instance of JS may be identified. One way is to look for an actually existent society to which we are strongly disposed to apply 'JS'. However, there is a strong reason to think that 'JS' is a concept which has, strictly speaking, an empty extension in the actual world, as there is no society which is just in every relevant respect.[28] Even the proponents of non-ideal theories of justice would agree about this, for even they tend to think about justice in a counterfactual way.[29] Specifically, they try to build into the definition of 'JS' only such requirements that are feasible to observe in the actual world, i.e., the requirements that an actual society could meet in a wide range of nearby possible worlds. Therefore, most theorists of justice can be plausibly expected to look for a counterfactual paradigmatic instance of JS. They differ only in how far the possible world in which their paradigmatic instance of JS is located is from the actual world: the more a theorist is drawn to the ideal theory of justice, the more remote an instance she will think of.

[28] For sure, simply because 'JS' has an empty extension in the actual world doesn't make it any less important or worth defining. This is a general point: in areas like philosophy, we often study idealised concepts that are actually empty but still important for our theoretical system.

[29] Theories of justice tend to be more or less ideal along a continuum depending on how strong the feasibility constraints that they impose upon justice are. For an exemplary utopian theory, see Cohen (2003, 2008). For exemplary realistic theories, see Waldron (1999) or Sen (2006). For a theory that combines the aspects of both extremes, see Rawls (1999a, 1999b).

Now, how can defining agents select an imaginary society that is a paradigmatic instance of JS? The most intuitive way seems to be to imagine a society that has some stereotypical characteristics that they associate with JS. For example, they may imagine a society which is governed by laws that are applied to its members impartially, where some elementary rights are guaranteed to everyone, where those of the same talent and ability have the same prospects of success regardless of their social origin, where the rights of minorities are protected, where wealth isn't enjoyed only by a few members of the society at the expense of its majority, etc. These characteristics would together amount to the above-discussed operational definition of JS. As explained in §2.2., this isn't the real definition of JS; its role is merely heuristic: defining agents would rely upon it when identifying the property that 'JS' expresses before they know its real definition. In practice, this means that they imagine a world containing a society that meets their operational definition of JS, identify it as a paradigmatic instance of JS, demonstratively pick it out as 'this society' and accept (9).[30]

(9) □ ∀x (x is a JS ↔ x is dthat [the same type of society as this society is]).

We can see that (9) exemplifies the same identificatory mechanism as Premise 1 since it's supposed to describe a sameness relation between whatever is a JS and a paradigmatic instance of JS.[31] However, is this mechanism at least as reliable here as it was in the original 'water'

[30] It isn't obvious whether it's possible to refer to a causally disconnected counterfactual entity, but I will charitably assume that it is. Detailed accounts of how it might be possible can be found in Hawthorne and Manley (2012, pp.4-35), Salmon (1998) or Thomasson (1993).

[31] This might not be the only way in which the operational definition may contribute to the identification of the given property as arguably it can also do it directly without demonstrative reference to a specific paradigmatic instance of JS (Salmon, 1981, pp.169-170). The basic idea is that the operational definition could be formulated somehow as (10).

(10) □ ∀*x* (*x* is JS ↔ it's an instance of dthat [the property whose instances are commonly S]),

case? This is dubitable since we can seldom imagine a possibility so vividly that we gain a complete picture of what this possibility involves. This difficulty is almost ubiquitous, as human imagination is generally too limited and course-grained to single out a particular possible world (or even a set of possible worlds) with all the relevant details.[32] Yet, in the absence of the relevant details about what a society that meets their operational definition of JS might look like, the defining agents cannot rule out that this society is in fact impossible or incompatible with being just as a whole. Consider, for example, the following two possibilities.

> (Possibility 1) Some stereotypical characteristics in the operational definition of JS are such that, in all nearby possible worlds, if a society has them, it causes severe harm to some of its members.

> (Possibility 2) Because of the unavoidable conflict in interests among different people, two or more stereotypical characteristics in the operational definition of JS are *in principle* mutually incompatible, and therefore no human society could possibly have them together.

If either of these possibilities was confirmed, it would defeat the agent's operational definition as inadequate. If Possibility 1 was confirmed, at least those theorists of justice who insist that JS be paradigmatically instantiated in a nearby possible world would treat it as the evidence that they failed to identify a paradigmatic instance of JS. Likewise, if Possibility 2 was confirmed, both idealists and realists about justice could be expected to react in the same

---

where 'S' denotes the stereotypical characteristics of JS. I admit that (10) might help us identify the property that 'JS' expresses just as well as (9). Even so, my forthcoming doubts about the identification via (9) concern precisely the risk of S being misleading and thus also apply to the identification via (10).

[32] See Yablo (1993, pp.27-28) or van Inwangen (1998) for details.

way. Consequently, to successfully identify a genuine paradigmatic instance of JS, it seems insufficient to merely imagine a putative society that meets our operational definition of JS because our imagination doesn't normally give us enough detail about what this society would be like to eliminate defeating possibilities like the two above.[33]

Furthermore, it seems that the more remote our imagined society is from the actual world, the more pressing a problem such defeating possibilities present. Notice that even in the original case of 'water', when the speakers identify a sample of the transparent liquid around them as a paradigmatic instance of water, they cannot firmly rule out all the defeating possibilities against this sample being a genuine instance of water. For all they know, it's possible that the atoms of its molecules have some peculiar properties which, at certain pressures and temperatures, change some macroscopic properties of the sample without changing its chemical structure. Suppose that, when this change happens, the sample becomes a pink and solid object but retains its chemical structure.[34] Now, many would argue that if this possibility was confirmed, it would reveal that the speakers have misidentified a paradigmatic sample of water. Still, the possibility doesn't seem very disturbing, which can be presumably explained by the fact that, given all the knowledge that the speakers have about how chemical substances behave in the actual world, it's just too unlikely to be seriously considered. By contrast, we cannot disregard defeating possibilities such as Possibility 1 or Possibility 2 with the same ease. Arguably, unless we can imagine a putative society meeting our operative definition of 'JS' in enough detail, which we normally cannot, we simply don't have enough information about it to rule out the above two possibilities as being too unlikely.

[33] Matters are even more complicated, as the initial speakers who stipulatively introduced 'JS' via (9) were also epistemically fallible agents who could have easily misidentified a genuine paradigmatic instance of the property they wanted 'JS' to express. Therefore, 'JS' most likely doesn't express the property which it was originally aimed to express.

[34] The example comes from Steward (1990, p.390).

To be fair, insofar as our imaginary society is located in a nearby world, we can perhaps disregard Possibility 1 if we gather enough counterevidence against it in the form of empirical information about the effects of the stereotypical characteristics of JS in the actual world that are projectable to all nearby possible worlds. However, this move isn't always available. Many ideal theorists of justice would be reluctant to apply 'JS' to any society in an actual or nearby world because of believing that the principles of justice can be jointly upheld only by a putative utopian society in a remote world. This is an objectionable yet certainly legitimate position, and similar reasoning can be pursued in relation to many other concepts. Unless there is a decisive reason to think that our concepts must always be instantiated in the actual world or at least in nearby worlds, the problem persists: we don't know enough details about what our imagined society would look like, and so we cannot be confident that this society is a genuine instance of JS to the same extent to which we can be confident that a sample of liquid around us is a genuine sample of water.

Another strategy to deal with defeating possibilities is to 'confront them directly'. In other words, we might attempt to revise our operational definition of JS to a form which is more robust and immune to such possibilities. The most common way to do so is to systematically subject our imagination to what Sorensen (1998, pp.135-159) calls 'alethic refuters', i.e., thought experiments in the form of scenarios that present defeating possibilities in concrete detail as counterexamples against what is *prima facie* conceivable. In practice, this means that we test whether our imagined society is a genuine instance of JS by constructing a scenario that fills in some relevant details about the society which follow from its meeting our operational definition of JS. If, in response to this scenario, we judge that the imagined society is either impossible or possible but unjust, the scenario defeats it as a genuine instance of JS. By

repeating this procedure, we keep revising our operational definition of JS until no defeating scenarios can be conceived.

It's undeniable that constructing alethic refuters sharpens our imagination since it prompts us to reconsider whether scenarios that are conceivable to us on first appearances in the sense of being consistent with everything that we believe are in fact possible once we conceive them more vividly and coherently by adding missing details to them.[35] However, appealing to alethic refuters at this stage of the PK-method seems premature. To explain, notice that the alethic refuters of operational definitions must be stronger than those that are commonly used in philosophical inquiry. Those are refuters such as Gettier Cases, which are meant to defeat our conception of something by presenting a possible counterexample to it: the possibility of a scenario in which there is an instance of JTB which isn't knowledge refutes that any JTB is necessarily an instance of knowledge, and hence that the real definition knowledge as JTB'. [36] By contrast, to refute the operational definition of JS, it isn't enough to point to a few counterexamples to it such as to a possible society which meets it but isn't just, or a possible society which is just but doesn't meet it. This is because the operational definition is nothing but the set of stereotypical characteristics associated with JS but doesn't purport to deliver the genuine definition of JS. Consequently, its only effective alethic refuters are those that attack the very possibility of a just society having all of these characteristics such as scenarios confirming Possibility 1 or Possibility 2.

The above point brings the PK-method to a paradoxical situation: alethic refuters were meant to make our faculty of imagination more reliable in identifying an instance of JS by

[35] Chalmers (2002, pp.149-156) distinguishes here between negative and positive conceivability.
[36] For a discussion of how alethic refuters of this sort should be formally regimented, see Williamson (2007, pp.179-207), Ichikawa and Jarvis (2009), Malmgren (2011) or Häggqvist (2009).

prompting us to refine our operational definition of JS used for this identification. But now it seems that constructing an alethic refuter that effectively challenges our operational definition is explanatorily more demanding than to simply present an imaginary counterexample against it. After all, as long as it isn't obvious that the stereotypical characteristics associated with JS entail some presently unconceived form of injustice or are mutually incompatible, constructing convincing alethic refuters against them will require us to inquire into the underlying mechanisms that generate injustice in societies and identify how these mechanisms are involved in these characteristics.[37] However, by doing so, we are already exploring what makes a society just or unjust, i.e., the real definition of JS. This seems like putting the cart before the horse: recall that Step 1 is only meant to be the initial stage of our inquiry during which we tentatively identify the property expressed by 'JS' via its paradigmatic instance(s). If the risk of misidentifying this property, and thus failing to complete Step 1, is so high that we can prevent it only by entering the very analysis of the property, methodologically speaking, it's worth asking whether Step 1 still serves its original theoretical purpose.

In response, my objectors might argue that the problem is overstated here. They might suggest that a slightly modified version of the PK-method can accommodate even those properties like JS, the identification of a paradigmatic instance of which seems insecure. Perhaps JS can be at least partially identified negatively as a property which is uninstantiated by such-and-such actual societies that are obviously unjust.[38] For example, an actual society in which racial discrimination is tolerated is obviously unjust. Of course, there is an indefinite number of other properties beyond tolerating racial discrimination that might indicate that a

[37] One might argue that the very fact that constructing scenarios confirming such defeating possibilities is epistemically demanding justifies us in dismissing these possibilities as too unlikely. But this would be so only if it were significantly more demanding than to disconfirm them. That's far from obvious given the complexity of relations between the stereotypical characteristics of JS.

[38] A similar approach has been adopted by Elder (pp.21-41) who argues that the identity of a property can be best cashed out by contrasting it with its contraries.

society is unjust, and thus the present method won't uniquely individuate the property of being JS by distinguishing it from all its contraries. Still, the proponents of the PK-method might think that the proposed strategy gives us at least an elementary grasp of JS before we start investigating its real definition by asking what the social arrangements of the actual counter-instances of JS would have to look like in order to become just.

I'm sceptical about the above response mainly because even the identification of the counter-instances of JS seems to be based upon the operational definition of JS. But, in that case, the individuation of JS via its counterinstances is just as susceptible to Possibility 1 and Possibility 2 as its individuation via its paradigmatic instance, and thus we might easily misidentify some properties in virtue of which the counter-instances of JS fail to be just. Let's assume though that I'm wrong. Nevertheless, the PK-method still owes us an account of how Step 2 and Step 3 can be successfully completed when we define philosophical concepts such as 'JS'.

3.3. Problem 2: Competing Individuating Criteria

To complete Step 2 and Step 3, defining agents must first discover based on what general criterion societies can be individuated as just or unjust. Afterwards, they must use this criterion to find out in virtue of which specific properties their identified paradigmatic instance qualifies as a JS. These properties will be the essence of JS. The first thing to note though is that completing Step 2 is problematic even with respect to natural kind concepts. Take, for example, biological kind concepts. There doesn't seem to be any single type of underlying property in virtue of which two species are members of the same biological kind. When taxonomists classify species into biological kinds, they are guided by various criteria such as intraspecific

genetic variability, reproductive isolation, ecological niche or the phylogenetic makeup. While, for each biological kind, some criteria are weighted more heavily than others, none of these criteria is systematically privileged to the degree that we could consider it as the general criterion for individuating species as biological kinds (Dupré, 1981, pp.83-90).[39]

To make things worse, it might be argued that the individuating criterion for chemical substances is just as inscrutable. Recall that, according to Premise 2, two samples are the same chemical substance iff they have the same chemical structure. But what does 'the same chemical structure' mean here? As Häggqvist and Wikforss (2018, p.97) point out, the following dilemma is involved in answering this question: If it means only that the given samples have the same molecular formula, then we cannot distinguish between chemical substances that are structural isomers in the sense of sharing a single molecular formula arranged in different spatial structures. For example, while propanol, isopropanol and methoxyethane are distinct chemical substances, they are all structural isomers of C_3H_8O. To accommodate these distinctions, 'chemical structure' must mean chemical composition *arranged in a particular molecular structure*. However, the molecular structure of water varies with different pressures and temperatures both across its solid, gas, and liquid phases as well as within each of these phases.[40] So, this understanding of 'chemical structure' has the opposite problem than the previous one: it's too narrow because it doesn't allow water to remain the same chemical substance at various presses and temperatures. Since there doesn't seem to be any 'middle-ground' interpretation of 'chemical structure' under which we can individuate structural isomers as different substances while allowing that water remains the same substance

[39] See also Hull (1997) or Richards (2010).
[40] See also Needham (2011, pp.8-11).

across different pressures and temperatures, we cannot but selectively alternate between these two criteria depending on what chemical substance we are considering.

The issue gets even exacerbated once we consider the above-discussed example of a pink and solid object with the chemical composition H_2O. It's certainly hard to swallow that such an object is water simply because it's H_2O, which is also why some philosophers think that macroscopic properties also contribute to the individuation of samples into chemical substances.[41] Admittedly, there have been attempts to put pressure on such examples. For instance, one might respond that water can be instantiated only in those worlds that obey the same physical and chemical laws as the actual world but insist that, in such worlds, the above example is impossible because of the tight causal relation between the chemical structure of water and its macroscopic properties.[42] An even more radical response would be to discredit such examples as metaphysically impossible by arguing that macroscopic properties of chemical substances are necessarily determined by their chemical structures.[43]

The weakness of these responses is that the significance of macroscopic properties in individuating samples as chemical substances can be demonstrated also via an actual example: even though most samples of terrestrial water predominantly consist of the isotopic variant $^{16}H_2O$, there are several other isotopic variants of H_2O. The most outstanding one is $^{2}H_2O$, also known as heavy water, some macroscopic properties of which are very different from the other isotopic isomers of H_2O (e.g., it's lethal to some organisms). For that reason, there has been a disagreement over whether it should be considered as an isotopic variant of water or a separate substance (Hendry 2006, p.866-869). However, if it's a different chemical substance, this

[41] See, for example, van Brakel (1986, pp.303-314) or Lowe (2011, pp.12-18).
[42] In fact, this view seems to be defended by the later Putnam (1990, pp.69-70) who claims that asking whether water is H_2O in chemically or physically impossible worlds makes 'no sense at all'.
[43] See Tahko (2015, pp.813-819) for discussion.

cannot be explained independently of its macroscopic properties, as most other isotopic isomers of H_2O that don't differ in macroscopic properties are treated as water without any controversy (Weisberg, 2005, p.340). Consequently, it seems that there are at least three competing candidates for the general criterion for being the same chemical substance: chemical composition *per se*, chemical composition + molecular structure and chemical composition + macroscopic properties.

The fact that we cannot settle upon what being the same chemical substance or the same biological kind consists in should give us pause because it reveals a serious problem in completing Step 2 of the PK-method even when defining natural kind concepts. Why should we believe that complex philosophical concepts such as 'JS' will do any better? Given (9), to complete Step 2, a philosopher defining 'JS' must come to know what conditions two societies must meet to be the same type of society. 'The same type of society' means in the context of the debate about social justice 'the same type of society with respect to how people are treated there in all the relevant respects'. Yet, what respects are relevant is unclear. Plausibly enough, how people are treated in a society can be determined by a variety of complex factors. However, in that case, just like in the case of biological kinds and chemical substances, there will presumably be various competing criteria by which the ways of treating people in a society can be individuated. A few of them include:

- the criterion that tracks how individual people in the society behave to one another on an informal level
- the criterion that tracks how various resources, the distribution of which is controllable, are divided among people within the society

- the criterion that tracks how also incontrollable resources such as natural endowments are divided among them
- the criterion that tracks how legal processes that determine the treatment of people are designed

These criteria might often fail to be jointly satisfied. As an example, there might be two societies where individuals treat one another in the exact same way in terms of their informal actual behaviour. However, while in the first society, they do so spontaneously without any legal guidance, in the second society they do so precisely because of being legally obliged to. Do these two societies belong to 'the same type in all relevant respects with respect to how they treat people'? To resolve such dilemmas, it's necessary to discover how the above criteria compare to one another in their relevance. However, given how complex and diverse phenomena societies are, it's hardly controversial to assume that there are weighted differently for different societies. In other words, some societies might be individuated more based on how resources are distributed to people, while others might be individuated more based on the informal behaviour of their members. What criteria are more relevant might depend on the ingenious characteristics of each society. A similar plurality of competing individuating criteria can be expected to exist also for many other concepts.

It should be clear why this complicates the completion of Step 2. If there are various ways in which these criteria are employed, it's doubtful that there is any systematically privileged criterion (or a set of complementary criteria) which being the same type of society consists in. However, this prevents us also from completing Step 3 because, if we don't know the individuating mechanism which determines whether two societies are of the same type, we can hardly figure out what specific properties make them so by reflecting upon our paradigmatic

instance of JS. After all, we don't know which properties of the instance are weighed more heavily than others in making it just.

Notice that the issue here isn't just an epistemic one. We cannot just assume that if we had a better understanding of how societies behave, we would find a general mechanism for individuating them based on how people are treated in them. After all, we know a good deal of chemistry or biology but haven't settled upon the general mechanism for individuating samples into chemical substances, or organisms into biological kinds.[44] Moreover, even our imaginative faculties don't seem very helpful in settling upon a criterion. What being a chemical substance consists in is an essentialist belief, and, in modal epistemology, such beliefs are usually considered to be an alternative to beliefs acquired by imagination rather than their explanandum.[45]

Additionally, even those epistemologists who think that imagination is the source of modal knowledge seem to presuppose the possibility of essentialist knowledge. For example, Williamson (2007, p.164) who thinks that the imagination is just an application of our capacity to handle counterfactuals in everyday life argues that essentialist beliefs such as that water is H_2O (and presumably also the belief in Premise 2) are part of the background knowledge that we are disposed to hold fixed when making counterfactual suppositions. In that case, whether or not it's conceivable to us that a pink and solid object with the composition H_2O or heavy water are water depends upon on our beliefs about what being a chemical substance consists in, rather than the other way around. As Roca-Royes (2011, p.38) points out, while this suggests

[44] Also, the issue doesn't seem to depend on 'just' being arguably a thick ethical concept whose semantics may contain an evaluative component because biological and chemical concepts for which the issue also arises are clearly non-evaluative concepts.

[45] See especially Lowe (2012) but also Hale (2013, pp.254-256), Jago (2021, pp.1987-1991) and Tahko (2018, pp.100-104).

that essentialist knowledge is a pre-condition for our counterfactual knowledge rather than its output, the question of what makes our essentialists beliefs knowledge remains unsettled. My present objection adds fuel to the fire here because it suggests that it might be impossible to acquire essentialist knowledge in the first place: there might be no general criterion for individuating samples into chemical substances.

Taking stock, in this chapter I have aimed to show that even if the PK-method might be a fruitful method to employ for defining *some* concepts, it seems ill-suited for many other concepts that philosophers define, especially those expressing properties which aren't instantiated in the actual world or those expressing complex properties that aren't attributed to objects by clear individuating criteria. This is a bad result for CM because, as explained in §2.2., the PK-method seems to be the least controversial way of implementing CM. The result is also relevant because RM has been recently criticised on the grounds that implementing the proposed revisions of concepts is often practically difficult.[46] While this is certainly true, my present discussion reveals that CM also faces its own implementation problem.

[46] See Cappelen (2018, pp.72-78), Pinder (2019), Deutch (2020), Koch (2021) or Jorem (2021) for further discussion.

Chapter 4: Reasons for Defining Concepts

4.1. The Wrong Kind of Reason Objection

The proponents of CM might be unconvinced by the argument against the model presented in the previous chapter. They might acknowledge that the argument has shown that specifying the actual meaning of a concept is often difficult but insist that this alone isn't enough to renounce CM because it's still the best available model for defining concepts. This would mean that, insofar as we know the real definition of the property actually expressed by a concept, *as defining agents*, we are rationally committed to defining the concept according to it. The purpose of this chapter is to resist this view by showing that, even in such cases, we may still rationally define it in a different way.

Consider first the following example, which is similar to the one introduced by Simion (2018, pp.920-921). We are defining the concept 'deer'. In our language, 'deer' is a natural kind concept that applies to species such as fallow deer, red deer, elk or roe deer. All these species belong to the biological family *Cervidae*, and thus all the evidence suggests that 'deer' picks out the kind *Cervidae*. We know what the real definition of *Cervidae* is. In fact, it's an exemplary case of what the real definition of a natural kind should look like: it's the conjunction of joint-carving biological properties each of which follows clear and unobjectionable taxonomic criteria and is essential to all members of *Cervidae*. At the same time, suppose that one of the sub-species of *Cervidae*, roe deer, is in a peculiar situation: its population inhabits only our world region and is particularly endangered there. Unless roe deer start to be treated by us more sensitively than the other deer sub-species, they will most likely go extinct before too long. However, neither hunting legislation, nor protection legislation in our region

discriminates between roe deer and less vulnerable deer populations. Moreover, the policymakers in our region don't like conceptual distinctions that are too fine-grained, and thus they are unwilling to draw up legislation that discriminates between the various sub-species that fall under 'deer'. They are willing to include special protective measures about the roe deer in it only if that sub-species doesn't fall under 'deer' but rather an exclusive concept. Consequently, unless we define 'deer' so that it excludes roe deer, the roe deer population won't be legally protected and will die out soon.

Given that, in the above scenario, our goal is to formulate the best possible definition of 'deer', what's the more rational thing to do: to define 'deer' so that it fails to express *Cervidae* but helps to save roe deer by excluding them, or to define it according to the real definition of *Cervidae*? Simion (Ibid., p.922) argues for the latter option, on the grounds that when defining agents decide to define 'deer' with the purpose of saving roe deer, they are responsive to the 'wrong kind of reason'. Two classical examples of the wrong kind of reasons that Simion appeals to are the following:

> (Case 1) An evil demon threatens to severely harm you unless you admire him even though he isn't even slightly admirable. This may give you an all-things-considered justification for admiring the demon, but it's nonetheless the wrong kind of reason for admiring him because it's unrelated to the relevant properties that he must have to be worthy of admiration.[47]

> (Case 2) Someone will offer you one million dollars for believing against all available evidence that you aren't bald. This gives you an all-things-considered justification for

[47] See Crisp (2000, p.459) and Rabinowicz and Rønnow-Rasmussen (2004, p.402) for variants of this case.

believing that you aren't bald. Yet, it's the wrong kind of reason for believing this proposition because it's unrelated to whether the proposition is true.

Simion (Ibid., pp.922-923) tries to draw the following analogy between the above two cases and her case: in both cases, the presented reasons are the wrong kind of reasons for taking respective attitudes because they are irrelevant to their function. The function of admiring an object is to positively evaluate some of its properties; however, the demon has no such properties. Likewise, the function of believing a proposition is to have a mental state that truly represents the world. But the pragmatic reason of getting one million dollars doesn't in anyway pertain to the truth value of the proposition that you aren't bald. Similarly, Simion (Ibid., p.923) seems to argue that defining 'deer' so that it excludes roe deer would help to save the roe deer population from extinction is a non-epistemic reason that is irrelevant to the central function that 'deer' is supposed to serve in virtue of being a representational tool: to represent the world without any epistemic loss about deer as a biological kind. She admits that we might still be all-things-considered justified in defining 'deer' in that way but thinks that the cost of doing so would be a decrease in the value that the concept has for us as a representational tool (Ibid., p.924). As a result, 'deer' would become a *worse concept* than it is when expressing the kind *Cervidae*.

What Simion (Ibid., p.923) concludes based on this argument is only that there must be some epistemic constraints on how it's permissible to define a concept. However, one might attempt to use Simion's argument to support CM in the following way:

> ‘The central function of all concepts is *the epistemic one*: to represent the world without any epistemic loss regarding the phenomena that they are concerned with.[48] ‘Deer’ expresses the kind *Cervidae* as a result of its semantic adaptation to this function. Therefore, applying ‘deer’ to objects based on whether they belong to this kind enables it to fulfil this function optimally. If we re-defined it and made it thereby express a different property or kind, it would start misfunctioning. Therefore, the definition of ‘deer’ that corresponds to the real definition of *Cervidae* trumps any other alternative definition.’

I agree that the central function of concepts is generally to represent the world *in some way*. After all, what concepts do is express properties by means of which we can characterise objects around us. Thus, possessing a concept allows us to make a new distinction: it provides us with a new way to divide objects across possible worlds into classes based on whether they have the property that the concept expresses. Hence, concepts are representational tools in two senses: firstly, they represent properties that individuate objects in the world.[49] Secondly, they can be also said to derivatively represent the objects instantiating properties expressed by them. Nevertheless, I believe that it would be a mistake to think that Simion’s argument can be used to support CM in the above way. This is for two reasons: firstly, even if the central function of concepts is the epistemic one, concepts might still fulfil it better if they expressed a different property than they actually express. Secondly, even though concepts generally have a

[48] ‘The central function of a concept’ means here *the function that is uniquely associated with a concept and fully determines its identity*. In addition, a concept can also have other non-central functions which it might acquire and lose without becoming a different concept. If a concept has several functions determining its identity, its central function is their conjunction.

[49] An important qualification: when writing that ‘concepts represent properties’, I don’t mean only that they express properties as their semantic values but *also that they are actually applied to objects based on whether they instantiate their expressed properties*. The two things would be the same only if a concept’s use fully determined its meaning, but this will be doubted below.

representational central function, it doesn't entail that this function is the epistemic one. In the next two sections, I will elaborate on these two reasons in order.

4.2. The Central Function of Concepts

Let's assume, for the sake of argument, that representing the world without any epistemic loss is indeed the central function of all concepts. Still, there is an explanatory gap between saying that the central function of 'deer' is to represent the world without 'any epistemic loss' and the claim that applying 'deer' to objects based on whether they instantiate the kind which 'deer' actually expresses enables it to fulfil this function optimally. What reason do we have to believe that 'deer', expressing the given natural kind, is epistemically optimal? For example, it might be the case that what a concept in fact expresses is given by our linguistic practice, i.e., by how its competent users are disposed to apply it to objects around them. And, even though, how we use a concept is definitely related to what is its central function, nothing guarantees that we are disposed to apply it in an epistemically optimal way.

To see this point, let's consider what Simion might mean when she writes that a concept functions properly when the way it represents the world 'doesn't translate into any epistemic loss' (Ibid., pp.923-924). Presumably, what she means is that a properly functioning concept represents the world in such a way that its possession and understanding makes it easier for its competent users to acquire true and informative beliefs about the phenomena that the concept is concerned with than if it represented the world in an alternative way. Now, it seems right that some ways of representing the world are more epistemically valuable in this sense than others. To use a trivial example, the definition of 'deer' as *Cervidae* is certainly more valuable for the purpose of acquiring true and informative beliefs about the phenomena that 'deer' is concerned

with than the definition of ‘deer’ as a creature that is considered in the actual world to be a messenger to the gods in the Shintō mythology. After all, only the first definition enables us to acquire true beliefs about the essential properties of the entities to which we apply ‘deer’.

Also, the members of natural kinds such as *Cervidae* tend to be homogenous in the sense of having a large number of common features. As Egré and O’Madagáin (2019, pp.8-9) have noticed, the lower is the variation of relevant features between the instances of the property expressed by a concept, the more projectable those concepts are and thus the more reliable inductive generalisations about them can be made. Hence, ‘deer’ expressing a genuine natural kind property certainly has some epistemic merits. However, I have doubts about whether, when expressing *Cervidae*, ‘deer’ performs its epistemic function optimally. My doubts are general: it isn’t obvious to me that concepts expressing natural kinds make us always acquire more true and informed beliefs about the phenomena that they are concerned with than they would if they expressed an alternative ‘non-natural’ property.

To explain, suppose that our linguistic community has just discovered that the animals to which we apply the concept ‘panda’ share a common ancestor with bear, and hence belong to the family *Ursidae*. Yet, we might think that it would be better if the definition of ‘panda’ was (11), where ‘M’ stands for the manifest characteristics of the panda, ‘G’ for its underlying genetic makeup that make it belong to Ursidae and ‘@’ is the name of the actual world.

> (11) ‘Panda’ expresses the property of being an animal with the manifest characteristics M that is either of the genetic makeup G or is of a different genetic makeup in the worlds other than @.

Because of some peculiar features of the panda which distinguish it from other bears, the hypothesis that it might belong to a different phylogenetic family than *Ursidae* was in fact seriously discussed before the first genetic research about pandas was conducted in the 1980s (Wei et al., 2012, p. 5662). Now, if 'panda' was defined by (11), then it would express a metaphysically strange property that is instantiated either by actual pandas with the genetic makeup G, or by non-actual animals that are manifestly indistinguishable from actual pandas but have a different genetic make-up.[50] This property certainly doesn't provide the real definition of the natural kind that 'panda' actually expresses because, for all we know, that kind is grounded in G. Despite that, defining 'panda' according to (11) might be epistemically valuable for understanding the nature of pandas. Why?

Suppose that we want the competent speakers of our community to be able to articulate counterfactual hypotheses about the dependence relation between some peculiar features of the panda and its membership in the phylogenetic family *Ursidae*, such as, for example, (12).

(12) If there were pandas that weren't ursids, they would have a bamboo-based diet.

Given that the panda is the only ursid with a bamboo-based diet, while the other ursids are either carnivorous or omnivorous, (12) is definitely a hypothesis worth investigating for deepening of our understanding of pandas. But as long as 'panda' expresses the natural kind grounded in G, there is no possible world in which pandas wouldn't be ursids. Hence, the

[50] Strictly speaking, it might be questioned whether these putative animals are really metaphysically possible. The answer will depend on whether manifest physical features of organisms are necessarily determined by their phylogenetic origin. However, it doesn't seem too controversial to assume that it's possible for the sake of discussion.

antecedent of (12) is necessarily false.[51] However, that leads to the unwelcome consequence that the whole counterfactual hypothesis (12) is vacuously true, and thus nothing disputable or worthy of further inquiry is stated by it. By contrast, if 'panda' is defined as (11), the inquiry of competent speakers about (12) becomes a genuinely open-ended inquiry about whether putative animals that are indistinguishable from actual pandas in all other respects than its phylogenetic origin would still have a bamboo-based diet.

It seems plausible that articulating counterfactual hypotheses expressed by (12) and their subsequent investigation helps us to acquire many true and informative beliefs about what bearing pandas' phylogenetic origin has on their manifest properties. To be fair, defining 'panda' according to (11) isn't, strictly speaking, required for articulating such hypotheses because the same hypothesis as expressed by (12) can be expressed even by (13).

> (13) If there were animals with the manifest characteristics M that were of a different genetic origin than ursids, they would have a bamboo-based diet.

Yet, when the given hypothesis is expressed by (13) instead of (12) it's no longer obvious that it conveys the information about the dependence relation between pandas' phylogenetic origin and their bamboo-based diet. Consequently, it's more difficult for the competent speakers of our linguistic community to use (13) to arrive at true beliefs about this relation. On the other hand, acquiring these beliefs by means of the hypothesis as expressed by (12) seems much more straightforward. Thus, there seems to be a significant epistemic gain in defining 'panda' according to (11).

[51] Suppose that, plausibly enough, the competent speakers of our community cannot be expected to evaluate counterfactuals according to impossible worlds semantics but only according to the standard Lewis-Stalnaker possible worlds semantics (Lewis, 1973; Stalnaker, 1968).

One might object that the epistemic gain of defining ‘panda’ according to (11) is outweighed by epistemic losses this might lead to. The argument would be that the new definition will make it more difficult for speakers to individuate the natural kind that ‘panda’ originally expressed and thus entertain beliefs about it. But this seems very questionable. After all, as long as competent speakers understand (11), they can be expected to be able to individuate the given natural kind simply by the first disjunct of (11) as ‘the property of being an animal with the characteristics M and the genetic makeup G’.

Admittedly, representing the given kind by this description is less convenient than simply by ‘panda’. Still, we might reason as follows: we only have a limited number of conceptual resources in our language, and thus we cannot have a separate concept for each property worth individuating. Metaphysically strange properties as the one in (11) seem harder to individuate than ‘joint-carving’ natural kind properties such as the property of being an animal with the characteristics M and the genetic makeup G. However, individuating the former property has epistemic utility in investigating animals instantiating the latter property, so we should be able to do so as easily as possible. Individuating a property by a single concept is cognitively easier than doing so by a lengthy description. Therefore, it’s rational to define ‘panda’ according to (11) but individuate the latter property by the above description. This is especially so because we know that the resultant definition of ‘panda’ will contain this description as its disjunctive constituent, which will make it salient to those speakers who understand the concept.

What the above example illustrates is that even if the central function of all concepts is the epistemic one, this doesn’t automatically support CM. The reason is that whether what concepts actually express enables them to optimally perform this function is far from obvious, even in

the case of scientific concepts expressing natural kinds. For other non-natural concepts, appealing to their epistemic function in support of CM seems even more problematic. This is so because, even though the central function of these concepts is to make some properties and their instances in the world salient by representing them, the primary purpose of representing them may be different than to acquire true and informative beliefs about the relevant phenomena.

To explain, as Podosky (2018, p.7) notes, many concepts directly take part in shaping the reality, and thus are epistemically self-fulfilling in the sense that they will make themselves accurate by creating or having a causal effect on the very things that they represent.[52] For example, complex institutional objects such as credit cards, private property, corporations, or political offices seem to be partially grounded in concepts because their existence requires the imposition of a certain collectively recognised status which must be linguistically articulated by those concepts (Searle, 1995, pp.59-78). Therefore, they wouldn't exist if the society didn't have concepts representing the institutional properties which they instantiate or whose instantiation they presuppose.

Furthermore, even some social and human concepts that aren't directly constitutive of the things that they represent may still have considerable 'looping effects' on them (Hacking, 1995). Podosky (2018, pp.6-7) illustrates this via Haslanger's example (2016, pp.120-122) of how the fact that the property of being a caregiver is a part of the definiens of 'woman' contributes to women being unjustly confined to the role of caregivers in society. Even though the possession of reality-shaping concepts like this certainly enables their users to articulate several true and informative beliefs about the relevant phenomena, this doesn't seem to be their

[52] See also Isaac (2021).

central function. Their central function seems to be to directly change the social reality. Epistemic effects on their users seem to be either only the by-product of their functioning, or a mean to fulfilling this function. Another example from Haslanger (2020a, pp.249-250) is instructive here: the concept 'family' gives to its users 'a cluster of mechanisms for processing information about the coordination of domestic life, for example, intimacy, sex, raising of children, economic partnership, intergenerational transfers of traditions and property' (Ibid., p.250). However, this epistemic function is only instrumental to its practical function to help these users in coordination of their domestic lives. If the information that 'family' helps to process was completely useless for the coordination of domestic lives, 'family' would lose its purpose as a concept.

It's also worth noting that, most authors who have attempted to cash out what is the function of concepts would agree that it may not be solely epistemic. Nado, for example, (2021, pp.1520-1522) thinks about conceptual functions similarly as about the functions of other tools: concepts are tools designed to enable us to achieve certain tasks that we couldn't easily achieve otherwise. She stresses that the extent to which concepts succeed in this may be unrelated to their epistemic utility and should be measured more broadly in terms of 'efficacy' (Ibid., pp.1523-1526). Alternatively, Haslanger (2020b, pp.252-255) who supports Cummins' system account of function (1975) argues that the function of concepts should be characterised as the role that they play in various human 'systems' (e.g., a scientific system, legal system, religious system) but the choice of a system to which a concept belongs depends on our purposes and interests. Thomasson (2020, pp.444-445) also seems sympathetic to the system account of function combined with Millikan's etiological account (1987) which says that the proper function of concepts is what evolutionarily contributes to their proliferation. She admits that all concepts are to some extent subject to some empirical constraints coming from the world in

which they are used, but she denies that the central function of all concepts is to serve in an epistemic system (Ibid., pp.450-453). Along similar lines, Prinzing (2018, pp.866-871) argues the function of concepts is the purpose for which they are designed, where 'designed' is understood in the broadest sense which allows for the possibility that concepts are products of some unconscious processes that happen as our language evolves. He explicitly writes that the epistemic purpose is only a special case of all kinds of purposes for which concepts may be designed (Ibid., p.869).

The message of these authors is clear enough: when identifying what is the central function of a concept, we must consider a variety of factors such as what purpose the concept was designed for, to what human system it contributes, or what historically contributed its proliferation.[53] The view that the central function of all concepts is the epistemic one fails to capture the complexity of these factors.

This isn't to say that the epistemic account of conceptual function is completely unsupported in the literature. Most notably, Cappelen (2018, pp.182-188) seems to defend a simpler version of the epistemic account, arguing that the central function of each concept, 'X', is the trivial disquotational function to enable us to talk (and thus exchange information) about the relevant phenomenon X that the concept is about.[54] For example, the central function of 'family' would be to simply enable us to talk about families, etc. Yet, provided that concepts are representational tools, Cappelen's account seems to me as extremely counter-intuitive. After all, most ordinary tools have a central function that is specifiable without the reference to that tool. For example, the central function of a hammer is *to deliver a physical impact to a small*

[53] Interestingly, this theory of conceptual function is very similar to the one that Simion herself (together with Kelp) has recently proposed (2020, pp.989-990).

[54] Brigandt (2010) also defends a version of the epistemic account; however, as Cappelen (2018, pp.184-185) observes, his account seems to be tailor-made specifically for scientific concepts but doesn't generalise.

area of an object rather than to *serve as a hammer*. Similarly, the central function of a computer is *to carry out sequences of arithmetic and logical operations automatically* rather than to *serve as a computer*. Why should this be different for the central function of concepts?

Cappelen (Ibid., p.183) briefly says that disquotational functions are the only non-controversial candidates for the central functions of concepts but it isn't clear why he thinks so. If it's only because disquotational functions are so trivial that there cannot be a disagreement over whether concepts in fact have them, then the argument isn't very compelling because such disagreement is just to be expected given that concepts acquire their central functions based on a complex range of the above-discussed factors.

Neither is it obvious that disquotational functions are the only candidates that guarantee one-to-one correspondence between them and their respective concepts. Consider, for example, the concept 'family'. Arguably, the central function of 'family' can be specified as 'to represent a complex property that is ascribed to groups united by some proper types of ties and the understanding of which provides people a cluster of mechanisms for processing information that helps them to coordinate their domestic lives'. Even two agents who would disagree over what counts as 'proper types of ties', and consequently over what definition of 'family' best enables it to fulfil the function so described, would both accept this function as the common ground of their disagreement. Also, I cannot think of any other word that would have precisely this central function but wouldn't be the same concept as 'family' or of a situation in which 'family' wouldn't be supposed to serve precisely this function while preserving *its identity qua concept*. Thus, there seems to be a one-to-one correspondence between this function and 'family'. Of course, there might be concepts for which it's more difficult to non-

disquotationally identify their central function. Yet, this is more because the central functions of concepts are complex phenomena rather than because it would be *in principle* impossible.

4.3. The Constitutive Purpose of Defining Concepts

The proponents of CM might think that the crux of their argument can be retained even if they concede that the central functions of concepts aren't exclusively epistemic. They might reason as follows:

> 'We acknowledge that the central functions of concepts vary depending on the above-discussed complex factors. Still, it's true that the property that a concept *in fact* expresses is the one which enables it to fulfil its central function optimally when applied to objects based on whether they instantiate this property. Take, once again, the concept 'deer'. Since 'deer' has been designed or evolved to be a taxonomic concept that would contribute to the zoological taxonomy of animals, its central function is to represent a natural kind at the taxonomic rank of family instantiated by hoofed ruminant mammals. Given semantic externalism, this kind is also what 'deer' in fact expresses. Therefore, when defining 'deer', we will make do with asking the descriptive question "What property does 'deer' in fact express?". If we instead asked the normative question "What property should 'deer' express?", we would either mean "What property should 'deer' express given its central function?" or "What property should 'deer' express, *all-things-considered*?" Asking the former question wouldn't add anything to the descriptive question, for what property 'deer' should express given its central function is precisely the property that it *in fact* expresses. Asking the latter question would make us responsive to the wrong kind of reasons, i.e., the reasons irrelevant to the central

function of 'deer'. Moreover, the same reasoning can be easily extended also to any concept even if its central function is non-epistemic.'

And so, the proponents of CM would argue that, when defining any concept, asking the descriptive question about its actual meaning is the most appropriate question to ask.

I think that this argument doesn't work. But before presenting my response to it, let me first do full justice to it by elucidating one point: I *don't* take issue with the fact that the argument presupposes semantic externalism, i.e., the widely accepted thesis that meanings *aren't* entirely dependent on speakers' mental states.[55] Yet, it's worth noticing that the present argument is compatible with two distinct kinds of semantic externalism. According to *weak semantic externalism*, what determines a concept's meaning is how its competent speakers are in fact disposed to apply it, i.e., its dominant use.[56] According to *strong semantic externalism*, what determines a concept's meaning is directly its central function, i.e., the overall purpose for which it has been designed or retained in our language in linguistic history.

Weak externalism can be employed in the above argument for CM only if there is a convergence between what has been the dominant use of our concepts and their central function. However, as a matter of fact, such convergence is epistemically desirable for CM even from the strong externalist perspective. After all, our dispositions to apply a concept to some objects rather than others seem to be the main type of evidence that we rely upon when investigating what is that concept's central function. If there was no link between the concept's dominant use and its central function of concepts whatsoever, this evidence would be unreliable. In that

[55] See Putnam (1975), Burge (1979) or Korman (2006).
[56] This is still an externalist position because even competent speakers of a concept may think that it expresses one property while applying it to objects so that it in fact expresses a different property (Haslanger, 2006, pp.97-98).

case, strong externalism would imply that identifying what a concept means is even more problematic than I argued in Chapter 3.[57]

Is it generally plausible to assume that there is a convergence between a concept's dominant use and its central function? Since the original argument for CM was based on a too narrow construal of the central function of concepts as being epistemic, the assumption was vulnerable to counterexamples of a concept being dominantly used in a way that isn't epistemically optimal. In contrast, by allowing that each concept might serve a different kind of central function, the present argument is certainly in a better position to convince us that the given convergence in fact takes place. The strongest evidence for such convergence is that it seems to take place in the case of most tools other than concepts that have been used for a long period of time.

To explain, the central function of a tool can be considered to be something like its *use plan*, representable as a set of instructions specifying how the tool should be used given the purpose it was designed for. For example, the use plan of screwdrivers is to install and remove screws. Once a tool has been associated with its use plan, two distinct courses of events might follow. One possibility is that we won't be able to efficiently use it for what it was designed for, i.e., its actual use won't match its use plan. In such a case, we will either abandon the tool as useless or transform it into a different tool by revising its use plan. Specifically, if a screwdriver proves to be completely useless for installing and removing screws, we will either stop using it, or we will turn it into a different tool, such as a chisel. A different possibility is that the actual use of

[57] One might think that there are alternative ways of discovering a concept's central function such as rationally interpreting our shared representational tradition concerning the concept (Schroeter and Schroeter, 2015) or constructing a reverse-engineering story of how and why the concept has been designed and developed to its present form (Queloz, 2021; Pettit, 2020, pp.348-354). I don't have space to discuss these methods here, but I suspect that both of them also ultimately rely upon our observations about the dominant use of concepts.

a tool will get adapted to its *use plan* through the improvement of the tool's properties or of our abilities to use it in the relevant way.[58] This means that we refine the relevant features of our screwdriver so that it can be efficiently used for removing and installing screws.

Similarly, whenever we see that a concept has been used by our linguistic community for a longer period (i.e., that we didn't abandon it), we can take it as evidence that the concept has either retained its original use plan (i.e., its central function) which it acquired when it was coined or has acquired a new use plan. In either case, the dominant use of the concept and its use plan can be eventually expected to converge either because of the dominant use getting adapted to the concept's original use plan or because of its use plan being revised in response to what purpose the concept has become dominantly used for. Strictly speaking, the possibility that the concept is dominantly used even after a long period of its existence in a way that makes it fail to satisfy its use plan cannot be excluded but it's relatively unlikely, as it would reveal systematic irrationality or ignorance in the linguistic community.

However, such cases sometimes arise. For example, even though the concept 'witch' has existed ever since ancient times, even in the 15th century, speakers dominantly applied it to actual human beings none of whom instantiated the supernatural powers that the concept was designed to track. Apart from that, discrepancies between the dominant use of a concept and its central function mostly arise either when the concept is newly coined and we haven't yet thoroughly tested it for its usefulness or when its central function is to track a property that is beyond our power of recognition. If the proponents of CM aim to account for these cases, they'd better be strong rather than weak externalists. This is so because, as weak externalists,

[58] See Franssen (2006, pp.46-49) for a more detailed analysis of the conditions that a tool must meet to fulfil its use plan.

they must concede that the question 'What property the concept X should express, given its central function?' adds something to the question 'What property does the concept X in fact express?' in such cases. As strong externalists, they can avoid this consequence, for they may simply hold that X's meaning is fully determined by its central function even if its actual dominant use deviates from it.

An additional reason to prefer strong over weak externalism is that reducing meaning to use is problematic, as can be illustrated on the following example: if there was a wizard that could perfectly camouflage himself as a deer so that nobody could recognise him as a deer-impostor, all competent users of 'deer' would be disposed to apply the concept to him. However, the meaning of 'deer' clearly wouldn't thereby change so that it includes this wizard in its extension. Yet, this is precisely what weak externalism predicts. This just confirms that strong externalism is the most cogent position that the proponents of CM may adopt. Henceforth, I'll grant it to them.

With this in mind, here is my response to CM: the claim that any reason which doesn't conform to the central function of a concept is *the wrong kind of reason* to define that concept is false. The only argument that Simion has offered for this claim is the argument from analogy featuring Case 1 and Case 2. But these analogies don't withstand scrutiny. To see this, let's first clarify what the wrong kind of reasons exactly are. While several criteria for distinguishing between the wrong and the right kind of reasons have been proposed, the least controversial criteria seem to be those that are sensitive to the varying nature of each individual action which a reason might be of the wrong kind with respect to, such as WKR.[59]

[59] A similar criterion has been defended in Schroeder (2010) or Sharadin (2016). Two other notable criteria are:

(WKR) The wrong kind of reasons for performing an action are those that are irrelevant to the internal standard of correctness of that action.

WKR says that, for each action, there are internally correct and incorrect ways of performing it. The incorrect ways are the ones that don't satisfy the action's teleology, i.e., *its constitutive purpose to which anyone performing it gets thereby committed* (Korsgaard, 2008, pp.7-10).[60] This strikes me as a plausible criterion. To illustrate, the constitutive purpose of admiring an object is to positively evaluate its genuinely valuable properties. Consequently, the consideration that the evil demon threatens to torture you unless you admire him satisfies WKR because it doesn't in anyway relate to the constitutive purpose of admiring.

Given WKR, we can interpret Simion (2018, pp.922-923) as arguing that the central function of a concept is also what guides *the constitutive purpose of defining it*. Now, while Simion's original argument is that the central function of each concept is solely epistemic, the present adaptation of her argument for defending CM allows that the central function of some concepts may be also non-epistemic. The reconstruction of this argument then looks as follows:

(WKR*) The wrong kind of reasons are concerned more with the properties or consequences of the action that they motivate as opposed to its object (Parfit, 2001; 2011, App. A; Piller, 2006; Rowland, 2013; Samuelsson, 2013).

(WKR**) The wrong kind of reasons are those that are only indirectly related to actions that they motivate because of some other reasons that one has (Stratton-Lake, 2005; Persson, 2007; Suikkanen, 2009; Raz, 2009).

Unfortunately, this isn't the place to elaborate on the difference between the three criteria. Suffice it to say that both WKR* and WKR** are too narrow which makes them ill-suited for those actions that either aren't related to any object (Schroeder, 2010, pp.33-36) or cannot be performed directly in response to any reason (Heuer, 2011, pp.170-71, pp.179-180).

[60] To be precise, Korsgaard talks about the *constitutive principle or norm* of an action, which anyone performing it is guided by. Yet, the basic idea is the same.

(Premise 1) The constitutive purpose of defining a concept is to associate with it a property which is such that applying the concept to objects based on whether they instantiate this property enables it to optimally fulfil its central function, whatever that function is.

(Premise 2) A reason is the wrong kind of reason to define a concept in a certain way iff it's irrelevant to the constitutive purpose (i.e., the internal standard of correctness) of defining it.

(Premise 3) The reason X doesn't conform to the purpose of defining the concept Y.

(C) X is the wrong kind of reason to define Y.

How can this argument be deployed in support of CM? Recall that the proponents of CM may hold that each concept expresses the property that satisfies the description in Premise 1. Therefore, via Premise 1, it follows that the constitutive purpose of defining a concept is to identify the property that it in fact expresses. From the above argument, it can be then derived that the reasons that don't conform to what the concept in fact expresses are the wrong kind of reasons for defining it.

I believe that Premise 1 is false because it rests upon confusion about what the constitutive purpose of defining a concept is, which is caused by reading too much into the analogies with Case 1 and Case 2. Each type of action has its distinctive standard of correctness and defining a concept is importantly unlike admiring or believing. It isn't an action consisting in forming an attitude but is rather like the act of constructing a tool. When we define a concept, we

associate it with a property based on which we think that the concept should be applied to objects around us. According to CM, this must be a property that the concept in fact expresses (insofar as there is one), while, according to RM, it might also be a property that the concept should express even though it doesn't actually express it. Either way, as long as we succeed in building the association between the given property and the concept into the linguistic practice of our community, this will facilitate the application of the concept to objects based on whether they instantiate that property.[61] Hence, by defining a concept, we make a difference to what representational activities the concept can be effectively used for. This seems analogical to the process of constructing any other tool by equipping it with new features that make a difference to what activities it can be effectively used for. Therefore, we will get a better understanding of the constitutive purpose of defining concepts by examining the constitutive purpose of constructing tools in general.

Imagine, for example, that you are an engineer constructing a cutting-edge technological device such as a drone. You are deciding how technologically advanced your drone should be. You know that the central function of any drone is to be an aircraft that can be controlled remotely without any human onboard while being able to do as many things as crewed aircrafts are able to do. Moreover, your technical knowledge and resources allow you to construct a drone that can fulfil this function optimally. At the same time, you know that the more technologically advanced your drone is, the greater application it has in military operations and the more likely it is to be used as a weapon of mass destruction. Now, does it follow that if, after taking into consideration all the reasons for constructing one drone rather than the other,

[61] This is so even if defining a concept is understood in accordance with CM. We must remember that, according to strong semantic externalism, a concept can express one property while being misapplied to objects as if it expressed a different property. When elucidating what property the concept expresses, we make it easier to apply it correctly.

you construct a less technologically advanced drone that doesn't have military application, you have constructed it for the wrong kind of reasons?

If the constitutive purpose of constructing a drone is to enable it to optimally perform its central function, then it does indeed follow that you have constructed the drone in response to the moral (or environmental) considerations that don't conform to its central function and thus are the wrong kind of reasons. I have, however, two reasons to doubt that the antecedent of this conditional is true. Firstly, recall that the constitutive purpose of an action is such that anyone engaged in it has thereby committed herself to that purpose. As Korsgaard (2009, p.29) puts it, the constitutive purpose of an activity is such that it doesn't make sense to ask whether the purpose should be fulfilled from the standpoint of someone engaged in that activity.

However, to enable a drone to optimally perform its central function doesn't seem to be something to which anyone constructing it gets necessarily committed and which cannot be meaningfully questioned from the standpoint of a drone engineer. On the contrary, as long as a drone engineer has full control over its final product, we expect her to care not only about the instrumental value that her constructed drone would have for fulfilling its central function but also about why a tool with that central function is worth having in the first place. That is to say, it seems that before deciding how technologically advanced a drone to construct, drone engineers are committed to making a judgement not only about what makes the constructed drone instrumentally valuable relative to its central function but also about what makes it a valuable tool given its possible applications. This commitment becomes even stronger if they are constructing a drone that will serve as a model for other drones and might eventually lead to the redefinition of the central function of drones in general. Moreover, they seem committed to making this judgement precisely because of being engaged in the activity of constructing a

drone *rather than because of any other practical identity which they acquire prior to that activity*. That is to say, we consider it as a part of the job of drone engineers to construct drones which are *ultimately valuable technological tools* for people. It they fail to do so; we blame them for committing not only a personal but also professional failure.

Secondly, the wrong kind of reasons for an action deviate from its constitutive purpose in the sense of being irrelevant to it. However, there seems to be a considerable difference between the case when, as a drone engineer, you decide to construct a less technologically advanced drone because of having concerns about its possible destructive use and the case when you decide to do so purely for an independent prudential reason, e.g., only because the evil demon threatens to torture you otherwise. In the first case, you make your decision after seriously reflecting upon what will make the constructed drone a valuable tool for us given its possible applications, while in the second case it almost seems like you ignore any consideration about how drones might be effectively used as tools. You just want to avoid the torture and would do the same no matter what drone the evil demon would want you to construct. Hence, it's the second case in which you seem to respond to a reason that has nothing to do with the constitutive purpose of constructing a drone. The view that the constitutive purpose of constructing a drone is to construct it in the full conformity to its central function completely misses this difference by treating both these cases as involving the wrong kind of reasons.

Accordingly, what the above discussion suggests is that the constitutive purpose of constructing a tool like a drone is to enable it to fulfil its central function satisfactorily but only to the extent to which this function makes it an *all-things-considered* valuable tool. Under this proposal, you didn't respond to the wrong kind of reason when you decided to construct a less

technologically advanced drone as prevention against its destructive use. Admittedly, the proposal is in tension with Korsgaard's view (2009, pp.27-34) about the constitutive purpose of constructing tools. Korsgaard argues that anyone engaged in, for example, building a house is necessarily committed to building a good house in the sense that it fulfils its central function, which is to serve as a habitable shelter (Ibid., pp.29-30). However, Korsgaard isn't clear about one subtle point: does 'good house' mean for her 'non-defectively functioning house' or 'optimally functioning house'?

Suppose first that it means the former. The tension between our views is then smaller than it seems. On my view, as a housebuilder, I'm necessarily committed to building a house which serves as a habitable shelter to the extent to which this function makes it an all-things-considered valuable tool for us. This normally commits me to building a house that is a good house in the sense of serving as a habitable shelter to a high enough degree for it to be non-defective. The only exception are rare cases when the built house will be, all-things-considered, a more valuable tool if it cannot be effectively used for its central function. An example of such a case is when I know that, for whatever reason, if my house is habitable, inhabiting it will cause massive pollution to my neighbourhood but that even if it isn't habitable, it will have a different harmless use. In that case, my account plausibly implies that if I build an inhabitable house, it will be a sign of me being a responsible housebuilder rather than of me deviating from the constitutive purpose of housebuilding.[62] Hence, while I reject that when building a house, I'm necessarily committed to building a non-defective house, I still think that in most cases, as a housebuilder, I'm committed to doing so.

[62] Of course, there might also be cases when we cannot construct a tool responsibly; for example, when we know that, whether it's defective or not, the use of the tool will bring about an overriding negative consequence. In such cases, refraining from constructing the tool and thus deviating from the constitutive purpose of its construction might be the best thing to do.

Suppose now that by ‘good house’ Korsgaard means ‘optimally functioning house’. Then the tension between our views is greater: while Korsgaard believes that, whenever being engaged in housebuilding, we are necessary committed to building an optimally functioning house, I believe that we seldom are. Yet, my view still seems true for the reasons discussed above. After all, in most cases when we are building a house, we don’t aim to build the best possible house, and rightly so. Rather, we try to compromise on the optimal functioning of the house whenever this is a responsible thing to do. For example, we decide not to build a house that is an optimally habitable shelter if we know that the maximal comfort of living that such a house would offer is unecological. Again, it seems misguided to say that by doing so we betray the constitutive purpose of housebuilding, as this is just what we are expected to do as responsible housebuilders.

Since concepts are also just a specific type of tools, I believe that everything I have written about the constitutive purpose of tools such as drones or houses should also be applicable to them. Hence, Premise 1 is false, and we should accept (14) in place of it.

> (14) The constitutive purpose of defining a concept is to associate with it a property which is such that applying the concept to objects based on whether they instantiate this property enables it to fulfil its central function satisfactorily but only to the extent to which this function makes it an all-things-considered valuable representational tool.

Let me exemplify how (14) works using the original example of ‘deer’. We know that the central function of ‘deer’ is to represent a natural kind at the taxonomic rank of family instantiated by such-and-such hoofed ruminant mammals. However, when defining ‘deer’ we also ask what makes the concept with this central function all-things-considered valuable as a

representational tool. In the given case, this makes us realise that even though ‘deer’ fulfilling this central function optimally has some epistemic merits, it will eventually lead to the extinction of one sub-species of the given natural kind, namely the roe deer. Additionally, we also know that if ‘deer’ changes its meaning so that it excludes roe deer, their extinction can be prevented. This gives us a significant moral reason to define ‘deer’ so that it expresses the property whose real definition is the same as the real definition of the natural kind that ‘deer’ currently expresses, except that it includes a qualification that makes it exclude roe deer from its extension. For sure, if ‘deer’ is applied to animals based on whether they have this property, it will no longer fulfil its central function optimally. While it will certainly group together most sub-species of ‘deer’ based on objective similarities between them, its exclusion of ‘deer’ will be artificial, and thus it won’t express a genuine natural kind anymore.

However, even if ‘deer’ will fare worse relative to the given central function after being so defined, we still have an alternative word ‘Cervidae’ with the same central function. This is a less known zoological term but expresses the same natural kind property that ‘deer’ originally does, so, ultimately, our scientific taxonomy won’t be significantly impoverished. Moreover, we know that if we don’t change the meaning of ‘deer’, one species in its actual extension will soon disappear from the actual world. This will be unfortunate not only for moral and environmental reasons but also for epistemic reasons because, for sure, studying roe deer helps us to acquire many true and informative beliefs about deer in general. Having considered all these things, we might legitimately conclude that re-defining ‘deer’ in the given way will make it an all-things-considered more valuable concept than it was before.

Insisting that if, in line with our judgement about the all-things-considered value of ‘deer’ as a concept, we define it so that it excludes roe deer, we do so for the wrong kind of reasons

seems just as misplaced as it was in the drone case. Premise 1 implies that no defining agent is committed to making a judgment about the ultimate value of defined concepts in virtue of being a defining agent even though such a judgement is precisely what we expect from responsible defining agents. This is especially so because, as argued above, concepts have representational functions of various kinds apart from the epistemic one, and thus it might easily happen that, what we previously thought to be the most valuable central function that some concept may have, will turn out to be less valuable within a broader picture. Also, it seems that there is a difference between the case when we re-define 'deer' so that it excludes roe deer to avoid the torture by an evil demon and the case when we do so in order to save roe deer population. Only in the second case, our decision seems to be motivated by a serious reflection about what makes 'deer' a valuable representational tool. In the first case, our decision seems to be purely prudential, and thus unrelated to any consideration about possible representational uses of 'deer' whatsoever. (14), unlike Premise 1, can make sense of these intuitions by allowing that the constitutive purpose of defining a concept is to do something more than to blindly conform to whatever central function the concept has, no matter what the ultimate value of that function is.

4.4. Some Objections

In what follows, I will consider three objections against my above argument, hoping that the discussion will elucidate what (14) precisely says.

Objection 1: 'If, after making a judgement about the ultimate value of a concept, we define it based on this judgement, we change the concept's meaning which might also require a change in its central function. This is so especially if strong semantic externalism, according to which

what the concept means is given by its central function, is true. However, by revising its central function, we change the very identity of the defined concept, which might lead to the concept no longer addressing the pertinent topic that it originally addressed. This is an unpalatable consequence because, when defining a concept, we want to preserve some continuity in the topic that the concept addresses.'[63]

Reply 1: I have two things to say in response to this objection. Firstly, even though I agree that preserving the continuity in topic is in general desirable, there might be cases when changing the topic that a concept addresses is instead the best thing to do.[64] For example, one might argue that if the central function of 'race' was to be a biological kind concept tracking sub-species of human defined by their skin colour, this would be obviously a very bad function not only because no such sub-species exists but also for obvious moral reasons. In such a case, the best thing would be to change the central function of 'race', and thus the topic it concerns, to something completely different.

Secondly, I'm not convinced that, in most cases of changing the identity of a concept by changing its central function, the difference between the old and new concept is large enough to disrupt continuity in the topic that the two concepts address. For example, if, in Simion's case, we conclude that the original central function of 'deer' doesn't make it, all-things-considered, the most valuable concept for us and subsequently revise it into a different concept, we don't lose track of the phenomena that the original concept is concerned with. The best way to see this is to notice that even if we revise 'deer' so that it excludes roe deer, the new concept

[63] The objection of this sort has been originally presented by Strawson (1963, pp.505-506) against Carnap's account of conceptual explication (1950, pp.1-18). Since then, it has been discussed in several places. See, especially, Haslanger (2000, p.34; 2020b), Prinzing (2017), Cappelen (2018, pp.97-136), Sawyer (2018) or Shields (2020).

[64] A similar point is made in Prinzing (2018, pp.856-857).

can still perform the old concept's function to a considerable extent. After all, even under its new definition (and central function), 'deer' can still represent the property, which is such that, even though it isn't a natural kind, it still has considerable extensional overlap across possible worlds with the natural kind that 'deer' originally expressed. Hence, there seems to be a criterion for distinguishing whether the continuity in topic before and after re-defining a concept has been preserved:

> (15) If the concept X gets transformed into the concept Y, the continuity in the topic addressed by X and Y has been preserved iff Y's central function and its semantic value enable it to fulfil the central function of X to a significant degree.

Admittedly, (15) isn't a maximally precise criterion since the expression 'to a significant degree' is somewhat vague. This will sometimes lead to disagreements over whether, according to (15), the new and old concepts are still continuous in their addressed topics. Even so, I believe that, in most cases, (15) fares well in terms of its theoretical utility. Specifically, (15) allows us to hold that the continuity in topic can be sometimes preserved even in the hardest cases when there is no overlap in extensions between the old and new concept. For example, when 18th century chemists discovered that combustion wasn't caused by the process of realising by a putative element called 'phlogiston' but rather by a different chemical reaction between a fuel and oxygen, they replaced the concept 'phlogiston' with the new natural kind concept 'oxygen'. Since oxygen is a real chemical element distinct from the non-existent putative element that phlogiston was supposed to be, the extensions of the two concepts never overlap. Nevertheless, it doesn't seem that replacing 'phlogiston' with 'oxygen' made the chemists' inquiry discontinuous: both before and after the replacement they were studying the cause of combustion.

Using (15) we can neatly account for this intuition: the central function of ‘oxygen’ is to represent a chemical element with several general characteristics, one of which is that it’s involved in the chemical reaction causing combustion. Its meaning is then the given chemical element meeting those characteristics. The central function of ‘phlogiston’ is to represent an element, the emission of which causes combustion. Even though ‘oxygen’ doesn’t optimally fulfil this function because it doesn’t denote an element which is released during combustion, it denotes the element which is *most causally active in combustion* through a different chemical reaction (consumption). Consequently, ‘oxygen’ still fulfils the central function of ‘phlogiston’ to a significant degree. Therefore, (15) seems to deal even with the hardest cases when the old and new concept are extensionally exclusive.

Objection 2: ‘Your criticism of the PK-method was that completing it is too demanding because it often isn’t transparent to us what properties concepts express and by what general criteria we can individuate objects as possessing these properties. However, why should it be easier to make an all-things-considered judgement about the value of a concept? After all, making such a judgement requires us to put a wide range of various epistemic, metaphysical, moral or social considerations about the concept and its possible uses on a common scale and weigh them against one another with respect to how they contribute to its all-things-considered value qua representational tool. This looks by no means easy.’

Reply 2: This objection would be appealing only if my argument was that all-things-considered judgements are made about the value of concepts *in abstracto*. In that case, it would be indeed mysterious how we could weigh against one another all possible considerations that might contribute to a concept’s value. However, contrary to that, I follow Chang (2004a, p.6;

2004b, pp.130-133) in understanding all-things-considered judgements as being made not in absolute terms but only relative to specific evaluative situations. That is to say, when making an all-things-considered judgement about a concept's value, we are making it relative to *a specific evaluative situation in which our linguistic community using the concept is situated.* This evaluative situation helps to determine what considerations are at stake in defining a concept as well as their relative weights.[65] Such restriction of relevant considerations by evaluative situations clearly takes place because not every fact is always relevant to what makes a concept a valuable representational tool in our language. For example, whether Port Meadow is flooding is irrelevant to the value of 'deer' qua representational tool in Simion's example, while the future of the roe deer population or phylogenetic characteristics of the actual instances of deer are relevant here. Hence, I don't argue that every possible consideration is relevant in defining a concept but only that the range of relevant considerations is wider than it would be if the only relevant consideration was what the concept in fact expresses, as CM claims.

To avoid confusion though, my proposal doesn't deflate the all-things-considered value of concepts to something which varies across each occurrent world-time situation in which the linguistic community of their users happens to be situated. Concepts are types of tools that are usually used over a longer period, and thus their central function must be durable. Also, just like in the case of any other tool, we care about various consequences to which the use of a concept might lead to not only in the actual world but also in nearby possible worlds. Therefore, an evaluative situation relative to which we evaluate concepts must be temporally extended as well as modally stable across nearby possible worlds in which they might be used.

[65] Note, however, that the claim that all-things-considered judgements are relative to an evaluative situation of a specific community doesn't imply that they must be fully grounded in something as arbitrary as the subjective preferences or interests of that community: what is genuinely valuable for agents in a situation might be something different than what they prefer. See Raz (1997), Scanlon (2000, pp.17-77), or Parfit (2011, pp.58-111) for details.

Now, that all-things-considered judgments about a concept's value are always made relative to an evaluative situation of its users is important for two reasons: firstly, it makes all-things-considered judgements seem significantly easier than if they were made *in abstracto* because it reduces the number of considerations that linguistic agents must weigh against one another only to those that directly pertain to the value that a concept has for them *as a representational tool in a given situation*. Thus, to make an all-things-considered judgement about a concept's value, defining agents don't need to be omniscient.

Secondly, making all-things-considered judgements about items in specific evaluative situations seems to be an integral part of our daily lives; it isn't something exclusive to the activity of defining concepts. Whenever we compare candidates for a job position, the qualities of two different lifestyles or the values of two different forms of art, we rely upon these judgements. Consequently, even if these judgements are difficult, there are unavoidable, we are constantly trained in making them and, arguably, we are often confident in having made them correctly. Therefore, I don't think that the argument that all-things-considered judgements should be also involved in defining concepts carries any additional explanatory burden than the one which we already face whenever we try to justify why we trust these judgements in our daily lives.

That being so, all-things-considered judgements might be helpful also in resolving one of the above-discussed problems involved in the PK-method. Recall from Chapter 3 that what often prevents us from completing Step 2 of the method is that there are several competing general essentialist criteria which might determine what having a property expressed by a defined concept consists in and there is no clear answer on which of them to rely upon. I believe

that this problem amounts to realising that what the central function of the given concept exactly is can be indeterminate. To illustrate, while we know that the central function of 'water' is to represent the same chemical substance as *this* sample has, we don't know based on what criteria to individuate objects as chemical substances. RM, unlike CM, allows that even if there is no such general mechanism applicable to all chemical substances, we might still revise the central function of 'water' to one of the following options:

- to represent the same chemical composition as *this* sample has
- to represent the same chemical composition and macroscopic properties as *this* sample has
- to represent the same chemical composition and molecular structure as *this* sample has

Our revision would be guided by our normative judgment over which of these functions makes 'water' all-things-considered the most valuable representational tool for us. While this unfolds an alternative method of defining 'water', the method doesn't support CM because it isn't immune from normative judgements about the all-things-considered value of concepts.

Objection 3: 'There are two distinct questions which you are conflating. The first question is what representational functions are such that it's all-things-considered valuable to have any concepts constituted by those functions. The second question is what representational function a particular word X should fulfil, all-things-considered, i.e., what concept X should be constitutive of. The first question is sensible because there are clearly some properties which are worthier of conceptualising than others. But you seem to be asking the second question which makes no proper sense. Words are just conventional lexical items, and it doesn't matter which property they represent. Take, for example, the word 'freedom'. As Chalmers (2011,

pp.532-537) points out, there are multiple interesting properties that this word might express: the ability to do what one wants, the ability to ultimately originate one's choice, the ability to exercise self-control over one's own affairs, etc. Each of these properties is worth conceptualising by *some* concept. However, it's pointless to ask expressing which of these properties makes 'freedom', all-things-considered, the most valuable concept, as that's just a matter of convention. It at most makes sense to add subscripts to the original word and distinguish between 'freedom$_1$', 'freedom$_2$',…,'freedom$_n$', making each subscripted word express one of these properties and alternating between these words based on which property we want to talk about. Hence, no significant normative question about the all-things-considered value of concepts can be asked.'

Reply 3: The two questions are interrelated. To explain, as Chalmers himself observes, insofar as we are non-ideal linguistic agents, the question as to what lexical item we employ to have a certain representational function cannot be dismissed as a matter of convention:[66]

> Ideal agents might be unaffected by which terms are used for which concepts, but for nonideal agents such as ourselves, the accepted meaning for a key term will make a difference to which concepts are highlighted, which questions can easily be raised, and which associations and inferences are naturally made. (Chalmers, 2011, p.542)

I fully agree with this idea. A concept can serve its central function only if speakers use it in accordance with this function. This can happen only if speakers recognise that the lexical item constitutive of the concept has this function. Therefore, whether this lexical item will be in fact dominantly associated by speakers with the concept's central function is a consideration that

[66] See also Cappelen (2020, pp.141-146).

matters *a lot* in the evaluative situation of any non-ideal linguistic community that deliberates over what central function the concept should have. In fact, if this wasn't the case, it would make no difference in the above example if, instead of adding subscripts to the original word 'freedom', we just created arbitrary new words expressing each of the properties associated with 'freedom'. But this obviously makes a difference because, for any non-ideal linguistic community, representing a property by a word that they have already been associating with the set of closely related properties is easier than representing it by a completely new word.

I wouldn't be afraid to go even further and argue the following: as non-ideal linguistic agents, we have only limited lexical resources, and thus cannot have a separate word for each property that might be worth conceptualising. However, in most cases, multiplying subscripted homonymous variations of the original word isn't a viable strategy. The strategy would have its rationale only if homonymous variations could be expected not to occur *in the same context of conversation*; otherwise, distinctions between them won't be salient enough to prevent the risk of conflating them. But, in fact, such risk seems to be very real: if properties are in the same semantic neighbourhood of a word, we can expect there to be some continuity in the topics that they address, and thus speakers will have to use homonymous variations when discussing the same topic. To prevent the risk that this will lead to our communication being obfuscated, the most reasonable option is to define 'freedom' by a single property that we judge to be, all-things-considered, the most valuable property in the semantic neighbourhood of 'freedom' to conceptualise. For other properties, the conceptualisation of which is less valuable for us but still valuable enough, we may then coin separate concepts that will bear an explicit lexical connection to 'freedom' but will be lexically distinguishable from it.[67] Accordingly,

[67] Chalmers (2020, pp.9-12) has recently touched upon this strategy, describing it as 'heteronymous conceptual engineering'.

against Objection 3, I hold that the first question and the second question cannot be treated separately: when we ask what properties are worth representing, we zoom in on the properties in the semantic neighbourhood associated with a particular word and ask representing which of these properties will make the given word the most valuable concept for us, all-things-considered.

Apart from the above three objections, there is one more general objection that I consider to be particularly pressing: 'An important implication of (14) is that defining a concept is a meta-semantic activity *par excellence*. Its primary purpose isn't just to facilitate the application of a concept in association with the property it in fact expresses. Its primary purpose is also to make the concept express a property that it should express, all-things-considered, and hence make the semantics of the concept sensitive to normative facts about its ultimate value as a representational tool. This however seems like a very *ad hoc* claim unless there is independent evidence that normative facts about concepts' all-things-considered value are meta-semantically relevant whatsoever, i.e., that they can encroach upon the semantics of concepts.'

I recognise the force of this objection. To draw a rough analogy, in epistemology a lot of effort has been recently put into justifying the thesis that the epistemic status of a belief may be sensitive to non-epistemic facts about morality precisely because such facts aren't what beliefs are standardly considered to be sensitive to.[68] Similarly, normative facts about concepts' all-things-considered value clearly go beyond what the semantics of concepts is standardly

[68] Some examples of the arguments for epistemic moral encroachment include Fritz (2018, 2020), Basu (2019a, 2019b) or Bolinger (2020). To my knowledge, meta-semantic moral encroachment has been explicitly defended so far only by Davies et al. (2019) and Díaz-León (2016). Their account, though, presupposes a contextualist semantics and thus cannot be applied more generally, even to concepts whose meaning is invariant across various contexts.

considered to be sensitive to. Thus, the thesis that concepts' semantics can be sensitive to these facts cries out for further evidence. The goal of the next chapter is to provide such evidence.

Chapter 5: Normative Encroachment in Meta-Semantics

5.1. Temporal Externalism and Two Phenomena

In the previous chapter I have argued that defining agents should define a concept in compliance with its central function only to the extent that this function is all-things-considered valuable. In this chapter, I will respond to the last objection to my argument by showing that normative facts about concepts' all-things-considered value for its users can encroach upon their semantics. The gist of my argument will be that the account according to which such normative facts may sometimes affect concepts' semantics has an explanatory utility in meta-semantics because it presents a promising alternative to an influential meta-semantic theory called Temporal Externalism (TE).[69]

According to TE, the meaning of a concept is sometimes determined by the future linguistic practice of its users. Hence, TE denies that what a concept means at a time necessarily supervenes upon the totality of the events that have taken place up to that time. What makes TE a compelling meta-semantic thesis is that it purports to provide the best explanation of the following two phenomena:

> (Determinacy) There are cases in which the events that have taken place up to a certain time seem to underdetermine the meaning of a concept, but we are still reluctant to describe the concept as semantically indeterminate at that time.

[69] TE is defended by Jackman (1999, 2005, 2020), Stoneham (2003), Collins (2006), Tanesini (2014), Rouse (2014), Haukioja (2020) or Ball (2020b), and more indirectly in Koethe (1982) and Cappelen (2018, p.73).

(Deference) There is a widely accepted thesis in meta-semantics called 'social externalism', according to which we defer to experts in our linguistic community for the meaning of our concepts (Putnam, 1975, p.186; Burge, 1979). The main evidence for such deference is that we are disposed to correct our understanding (and use) of a concept when discovering that experts understand (and use) it differently. This evidence remains equally compelling even if experts are situated in the future.

My objective here is to argue that the way in which TE is standardly thought to explain Determinacy and Deference is untenable. After introducing how TE is thought to do so, I will argue that there is a better way to explain the two phenomena via a theory that still implies a weaker form of TE. According to the proposed theory, normative facts about what makes a concept all-things-considered valuable for its users, which *may also depend upon the facts about the future linguistic practice of its users*, sometimes directly determine its meaning as long as other meta-semantic factors such as its dominant use or its central function underdetermine it.

5.2. Phenomenon 1: Determinacy

In this section, I will elaborate on how TE is supposed to explain Determinacy. Let's consider the following thought experiment which was designed by Wilson (1982, pp.549-551) and refined by Jackman (1999, pp.158-159) together with other similar thought experiments.

(Ave) There is a linguistic community (hereafter called 'Druids') that inhabits an isolated island and speaks an archaic language. Druids have never seen any other object apart from those located on their island and seldom think of merely possible objects.

Therefore, they have only applied their concepts to those objects that they have encountered on their island. As all flying and living objects located on the island are living, the class of flying and living objects is co-extensive on the island with the class of flying objects. Druids possess the concept ‘ave’ and apply it to objects in these coextensive classes. They have several beliefs about the objects that they call ‘ave’, such as that only aves can fly, that all aves are flying objects, or that all aves are living objects but they don’t have any specific meta-linguistic beliefs regarding what properties makes these objects fall under ‘ave’. Suppose further that the story of Druids at one point in time, t, branches into the following two scenarios:

> (Scenario 1) Druids first see a flying object in the sky. In line with their belief that only aves can fly, they call this object ‘ave’. Upon that, the flying object lands on the island and turns out to be a plane. Druids recognise that it’s an unliving object and revise their belief that all aves are living objects. However, they understand themselves as having always meant ‘flying object’ by ‘ave’ while mistakenly having now the corrected belief that all flying objects, aves, are living objects. Accordingly, they keep applying ‘ave’ to all flying objects.

> (Scenario 2) Druids first see a plane on the ground. In line with their belief that all aves are living objects, they call this object ‘non-ave’. Upon that, they see how the object takes off and flies. They realise that their belief that only aves fly is false and revise it. However, they understand themselves as having always meant ‘flying and living object’ by ‘ave’ while mistakenly having now the corrected belief that only flying and living objects, aves, can fly. Accordingly, they keep applying ‘ave’ only to flying and living objects.

The thought experiment is designed with the purpose to entice the following three intuitions in us:

(Intuition 1) The meaning of ‘ave’ in Scenario 1 didn’t change at t.

(Intuition 2) The meaning of ‘ave’ in Scenario 2 didn’t change at t.

(Intuition 3) The meaning of ‘ave’ after t in Scenario 1 is different from the meaning of the same concept after t in Scenario 2.

These intuitions together imply that the meaning of ‘ave’ before t is different in the two scenarios. Jackman argues that TE can neatly explain this result because it allows that, before t, whether ‘ave’ means ‘flying object’ or ‘flying and living object’ is determined by Druids’ future linguistic practice with ‘ave’, which is initiated by their response to the encountered plane at t in each scenario.

Intuition 3 is relatively uncontroversial. The fact that, from t onwards, Druids systematically apply ‘ave’ either to all flying objects in Scenario 1 or only to flying and living objects in Scenario 2 seems like reliable evidence that ‘ave’ in fact means ‘flying objects’ or ‘flying and living objects’ respectively. This is true even for strong semantic externalists who believe that the meaning of ‘ave’ is fully determined by its central function (i.e., its use plan) because, as discussed in §4.3., a concept’s dominant use and its central function can be generally expected to overlap. Intuition 1 and Intuition 2 are less straightforward. One might think that while Druids in the two scenarios don’t think that the meaning of ‘ave’ has changed in their language,

either in one of the scenarios or in both scenarios, they got this wrong. There are two main strategies through which this line of thought might be pursued.

As a first stab, one might think that, before t, the meaning of 'ave' is by default the less restrictive of the two interpretations. Hence, before t, 'ave' would mean 'flying object' in both scenarios. In Scenario 1, Druids wouldn't revise the meaning of 'ave' but only realise that not all flying objects are living objects, whereupon they would abandon their belief that all aves are living objects. By contrast, in Scenario 2 they would come to believe that some non-living objects can also fly without abandoning the belief that all aves are living objects. They would, however, unconsciously revise the meaning of 'ave' to 'flying and living object' so that the new meaning is jointly consistent with these two beliefs. Accordingly, one might conclude that even though Druids fail to recognise this, Intuition 2 is false.

I think that this reading of the thought experiment doesn't present a plausible alternative to TE. To elaborate on this, if in the period before t, 'ave' means 'flying object' rather than 'flying and living object', for sure, there must be something that determines that the first interpretation is by default semantically privileged over the other. What could this thing be? It cannot be Druids' meta-linguistic beliefs about what makes objects fall under 'ave', as they don't have any definite beliefs of this sort. Neither can it be Druids' dominant use of 'ave' manifested through their linguistic dispositions to apply 'ave' to some objects rather than others. After all, the thought experiment is designed precisely so that Druids before t don't have any determinate linguistic dispositions regarding whether or not to apply 'ave' to a flying but unliving object. This is clear from the fact that whether they apply 'ave' to a plane is completely contingent upon whether they first recognise that the plane is unliving or that it flies. Therefore, if anything determines whether 'ave' means 'flying object' or 'flying and living object', it must be

something external to Druids' actual use of 'ave'. An obvious candidate here is the *use plan* of 'ave', i.e., its central function that stems from the overall purpose for which the concept was designed or evolved in Druids' language.

Accordingly, one might provide a meta-semantic story of how, before t, 'ave' was designed or evolved to serve the central function which is best fulfilled by 'ave' meaning 'flying object'. For example, it might be argued that the central function of 'ave' is grounded in the dominant causal source to which Druids' linguistic practice with 'ave' is linked and that this source is the act of calling 'ave' an object that was more saliently just flying than flying and living. Alternatively, it might be argued that the belief that aves are flying objects is more entrenched in the cognitive system of Druids' linguistic authorities than the belief that aves are living objects, and the central function of 'ave' is responsive to this fact.

This move can be however easily blocked because nothing prevents us from refining Ave so that no similar story is available, and thus even the central function of 'ave' before t isn't determinate enough. For sure, we might imagine that 'ave' is a new concept that has been just recently introduced to Druids' language and doesn't yet play an exact role in their linguistic system. Druids simply started calling flying and living objects around them 'ave' without any of the two features being more salient to them. [70] [71] Likewise, we can imagine that all Druids' beliefs about aves are equally entrenched in the cognitive system of their linguistic authorities with none of them being a more semantically significant than others. Unless there is a different factor that affects the central function of 'ave' and cannot be blocked in a similar way, the reading of the thought experiment that denies Intuition 2 doesn't seem very promising.

[70] A similar point is made in Tanesini (2014, pp.6-7).

[71] Relatedly, the property of flying doesn't seem to be a stronger 'reference magnet' (Lewis, 1984; Sider, 2011, pp.43) than the property of flying and living in the sense of being easier to refer to because of being somehow metaphysically privileged.

A different tack is to argue that both Intuition 1 and Intuition 2 are false, as the meaning of 'ave' changes in both scenarios after Druids encounter the plane and revise their respective belief about aves.[72] One might argue that 'ave' is an open-textured concept before t in the sense that its application conditions to some novel objects to which the concept hasn't been applied before, such as planes, are indeterminate.[73] Now, it's questionable how exactly open-textured concepts behave semantically, but a rough idea is that 'ave' before t means something like 'some flying objects', while it's left indeterminate which flying objects the concept precisely applies to. In that case, when Druids classify the plane as 'ave' or 'non-ave' in the two scenarios and stick to their classification even after finding out that it commits them to revising one of their prior beliefs about aves, they change the meaning of 'ave' to a more precise form. The new meaning makes it determinate which flying objects 'ave' applies to.

The main problem with this strategy can be summarised as follows: the appeal to open-textured concepts is controversial in the context of the standard referential semantics, which assumes that every representational concept has a definite extension across all possible world-time pairs.[74] This isn't the case for open-textured concepts because their intension (i.e., the function from world-time pairs to their extensions) is partially undefined at least for some possible world-time pairs. In general, indeterminacy of this sort isn't semantically desirable, which can be seen from the fact that some authors argue that even those concepts that are vague in the sense of having borderline cases of applications, the case for the indeterminacy of which

[72] This interpretation is defended by Brown (2000, pp.184-185).
[73] This characterisation of open-textured concepts comes from Waissmann (1945, pp.119-134) and Hart (1961, pp.120-132).
[74] See Gauker (2007) for further discussion.

is most compelling, have, in fact, definite extensions.[75] Given this, it seems reasonable to prefer a semantic theory that eliminates indeterminacy as much as possible.

However, if 'ave' before t is semantically indeterminate, then almost any concept is also semantically indeterminate. For most concepts, we can find hard cases of novel objects where it's unclear whether the concept applies to them. This, as a result, over-generates indeterminacy. As Jackman complains (1999, p.172), this and similar accounts fail to give us a criterion by which to distinguish between the sense in which 'ave' in the thought experiment is indeterminate and the other, more striking cases of indeterminacy. Such cases may include not only vague concepts, but also possibly concepts that are like 'ave' before t but cease to exist before the speakers have a chance to decide whether to apply them to novel objects, or the concepts which are defined based on confusion about how many objects they individuate (e.g., a proper name mistakenly applied to a pair of twins during the initial baptism, etc).

By contrast, TE can avoid the over-generation of indeterminacy by distinguishing between the following two types of concepts.

> Underdetermined concepts (UC): the concepts whose applicability to some objects is determined at some time not only by earlier events but also by later events.
>
> Indeterminate concepts (IC): The concepts whose applicability to some objects is indeterminate at any time by all the earlier and later events.

[75] E.g., Sorensen (2001) and Williamson (1994). See also Jackman (2004) who argues against the indeterminacy of vague concepts using TE.

Temporal externalists can make use of this distinction by arguing that only IC are genuinely open-textured concepts that lack definite extensions at least in some world-times, and thus are semantically problematic in the above-discussed sense. By contrast, UC do have a definite extension at any world-time, but these extensions are sometimes determined retrospectively. Hence, the semantics of UC isn't incomplete. Accordingly, they may argue that most concepts, including 'ave', are such that the unclarity of their application conditions to some novel objects makes them UC but not IC.

5.3. Phenomenon 2: Deference

Let me now consider the second reason for adopting TE: TE seems able to best account for Deference. To explain, ever since Burge (1979, pp.77-79) presented his arthritis thought experiment against the individualistic account of meaning (and mental content), it has been widely accepted that the meanings of concepts depend in part on what experts in our linguistic community mean by them.[76] Burge considers two type-identical speakers of English who are situated in almost type-identical worlds, with the only difference being that, while in the first world (actual world), the experts in their linguistic community apply the concept 'arthritis' only to an inflammatory disease of joints, in the second world, they also apply it to an inflammatory disease of thighs. In the first world, when the speaker utters 'I have arthritis in my thigh' to her doctor, the doctor corrects her for using 'arthritis' incorrectly and the speaker accepts the correction. In the second world, the speaker uttering the same sentence uses 'arthritis' correctly.

[76] See also Burge (1989).

Burge (Ibid., pp.78-79) argues that the best explanation for the first speaker's disposition to accept the doctor's correction is that she uses 'arthritis' with the same meaning as the doctor, i.e., she defers to experts over what the concept means. Therefore, the speaker can be said to have believed that she had arthritis in her thigh. The same applies to the second speaker, for even she is disposed to correct herself if she misused 'arthritis' and was corrected by her doctor. However, the beliefs of the two speakers differ in truth-value: one is false; the other is true. Since everything else is type-identical in both worlds, the only way to explain this difference is to admit that the semantic content of both speakers' utterances and beliefs about arthritis is at least in part determined externally by experts in their community.

The relevant part of Burge's argument for TE is the one about the speakers' dispositions to defer to experts on what concepts mean. Recall that, to establish TE, it's enough to show that there are *some* cases when the meaning of a concept is at least partially determined by future linguistic practice. Temporal externalists standardly argue that such cases are the ones when present speakers are disposed to correct themselves about their understanding (and use) of a concept in response to evidence about how future speakers will understand (and use) it. On this account, temporal externalism is just a special type of social externalism about meaning which pertains to the cases when present speakers consider future speakers to have greater expertise about a given concept, and thus are disposed to defer to them on its meaning.[77]

When do such cases arise? Brown (2000, pp.186-187) argues that they never arise because future linguistic developments are full of contingencies and arbitrary decisions, and thus we aren't justified in thinking that they give us any relevant evidence about what our concepts

[77] This position can be found in Jackman (1999, p.173; 2005), Ball (2020a, pp.54-55; 2020b, pp.1058-1067), Ebbs (2000, pp.247-52) or Haukioja (2020, p.926).

mean. For example, if taxonomists want to find out what the concept ‘zebra’ means, they will choose to gather the evidence about the objects in its extension rather than about how the future speakers will understand (and use) the concept.

Jackman (2005, pp.370-373), however, responds to Brown by arguing that she overstates her point because it applies only to the cases when conceptual change is driven by unreliable sociological/psychological factors. Jackman instead thinks that conceptual change is often explained by reliable and rational factors, for we may often expect future speakers to have greater expertise on the concept’s meaning and use. Accordingly, he concludes that all cases of temporal deference which TE is concerned with are explainable by non-arbitrary factors: speakers first defer to the linguistic practice of future speakers on what a concept means, treating them as experts on this concept whose understanding (and use) of it is rationally motivated. Once they identify what these reasons are, they deepen their understanding of what the concept has meant all the way along.

5.4. Towards an Alternative Explanation of Determinacy and Deference

In this section, I will first argue that the way in which Jackman explains Determinacy and Deference by TE faces some serious difficulties. Afterwards, I will show that there is a better way to explain the two phenomena via a theory that still implies a weaker form of TE. As already outlined in §5.1., the theory will be that normative facts about a concept’s all-things-considered value for its users may sometimes directly determine a part of its meaning.

Let’s first return to Deference and carefully consider the following question: what does Deference tell us about what determines the semantic value of concepts? It seems that

Deference disconfirms both semantic internalism, according to which the meanings of concepts are fully determined by the mental states of their present users, as well as *weak semantic externalism*, according to which they are fully determined by their present dominant use. After all, if either of these theories were true, contrary to Deference, we would be never disposed to correct our understanding (and use) of a concept in response to the evidence that the future speakers of our linguistic community understand (and use) it differently.

At the same time, it would be preliminary to conclude that Deference establishes that future speakers' understanding (and use) of a concept is directly involved in determining its meaning. To see this point, recall that, in his response to Brown, Jackman claims that Deference particularly concerns the cases where speakers defer on a concept's meaning to future speakers because of believing that the future speakers' linguistic practice is more reliable:

> 'The future usage that TE takes to be relevant to what we mean is restricted to that usage which reflects epistemic improvements to our web of belief.' (Jackman, 2005, p.374)

Hence, Jackman seems to admit that we aren't always disposed to correct our understanding (and use) of a concept in response to the information that future speakers understand (and use) that concept differently, but only under the special conditions when future speakers seem to have a better grasp of the concept's meaning than we do. However, in that case, it's an open possibility that what we are ultimately disposed to defer to isn't the future speakers' understanding (and use) of the concept, but rather *a different factor determining the concept's meaning* that future speakers, whom we are disposed to defer to, just comply with. In other words, while future speakers' linguistic practice with a concept may sometimes serve as

reliable evidence of the concept's meaning, contrary to TE, this may not be because it partially determines the concept's meaning, but simply because the future speakers are able to use the concept in compliance with its independent meta-semantic grounding.

Temporal externalists can be expected to resist the above possibility, at which point they might appeal to Determinacy. As suggested in §5.2., it seems possible to tweak Ave so that all external factors situated before t, which might be thought to be meta-semantically relevant to 'ave' by affecting its central function, become ineffectual in 'breaking the tie' and determining whether 'ave' before t means 'flying object' or 'flying and living object'. Doesn't this leave us with the explanation of Ave according to which it's Druids' future practice with 'ave' which determines its meaning before t as our best option?

I doubt so. The problem with this explanation is that Druids at t certainly cannot be regarded as experts on the use of 'ave'. To illustrate this, imagine that a trustworthy prophet came to Druids before t and told them how their future selves would use 'ave' from t onwards. Still, insofar as Druids are minimally rational linguistic agents, it's doubtful whether they would be disposed to revise their understanding of 'ave' in response to hearing this. After all, whether their future selves will apply 'ave' to the given plane isn't in anyway rationally motivated but *completely contingent upon* whether they first see it in the sky or on the earth. Their use of 'ave' at t is a paradigmatic example of an uninformed and arbitrary linguistic practice that Jackman himself thinks we don't defer to.

In response, temporal externalists might argue that how strong Druids' dispositions to defer to their future linguistic practice are depends upon whether there is any alternative external resource to which they could shift the burden of determining whether 'ave' means 'flying

objects' or 'flying and living object'. They might suggest that, as long as the setup of Ave is such that no alternative resources are available before t, Druids' deferential dispositions towards the future use of 'ave' are particularly strong even if they have no reason to believe that the given use will be reliable.

Nevertheless, I *do* think that there is one other resource that Druids can defer to before t. Imagine that we would ask Druids before t whether it's in their interest to represent such a property by 'ave' that is worth representing, i.e., which is such that it's all-things-considered valuable for them to make it salient by conceptualising it as 'ave'. It seems that insofar as Druids are agents that take any normative point of view to the world by adopting evaluative attitudes towards things around them, it would be hard for them to deny that they care about the value of 'ave' as a representational tool. After all, if they really thought that 'ave' was a valueless concept, they wouldn't have any reason to introduce it to their language and use it to conceptualise a property in the first place. Also, if, for strange reasons, 'ave' expressed a property which is such that anyone who entertains it under the guise of 'ave' undergoes terrible suffering and dies afterwards, they wouldn't be disposed to revise 'ave' so that it no longer expresses this property. Hence, if Druids denied having any normative concern about 'ave', that would be such a striking manifestation of their normative indifference towards the world that we could only react to it in two ways: either we wouldn't believe their denial or we would start suspecting that they lack elementary normative reflectiveness about the world that is characteristic of human agents, which would also make us doubt whether they are human agents like us at all.

For sure, it's in principle possible to revise Ave so that Druids would be indifferent about everything related to the value of 'ave' and just blindly following whatever linguistic

convention that happens to govern the application of ‘ave’ to objects around them. Suppose though that Druids admit that they aren’t completely indifferent about the ultimate value that ‘ave’ has for them as a representational tool. In that case, we could ask them:

> ‘Assume that the central function of ‘ave’ isn’t fully determinate. Yet, there is a normative fact about which of the two interpretations (‘flying object’ and ‘flying and living object’) makes ‘ave’ a more valuable concept for you, *all-things-considered*. The all-things-considered value of ‘ave’ depends upon wider considerations such as what role it has in your linguistic system, what both theoretical as well as practical consequences it will lead to if it has one of the above interpretations as its semantic value, what consequences it might lead to in nearby possible worlds in which it could be used, and many others. You have two options: one option is to defer on the meaning of ‘ave’ to the normative fact about whether it’s an ultimately more valuable concept for you if it means ‘flying object’ or if it means ‘flying and living object’. Another option is to defer to the use of ‘ave’ by your future selves from the point t onwards. Which option are you more disposed to choose?’

Provided that Druids are normatively reflective human agents like us, they would obviously be more disposed to choose the first option. This amounts to saying that ‘ave’ before t expresses the property representing which makes it the most valuable representational tool for Druids, all-things-considered. If this happens to be the property that Druids’ future selves take ‘ave’ to express from t onwards, this is only by a coincidence and not because their future linguistic practice determines the meaning of ‘ave’, as temporal externalists claim. Let’s call this position Normative Externalism (NE).

There are three responses that the proponents of TE can make to the above argument for NE:

> (Response 1) They might simply tweak Ave so that Druids are indeed creatures that are completely normatively indifferent about the value of 'ave'.
>
> (Response 2) They might argue that neither of the two interpretations makes 'ave' a more valuable concept than the other.
>
> (Response 3) They might just submit to NE while pointing out that even NE implies a weaker form of TE.

In what follows, I will argue that Response 3 is the most rational one to make for temporal externalists.

Firstly, Response 1 doesn't look very helpful. Perhaps normatively unreflective Druids would indeed defer to any future use of 'ave' simply to prevent 'ave' from being indeterminate, even if it was just a contingent use that would be unrelated to what makes 'ave' an all-things-considered valuable concept for them. However, would this really be the result that temporal externalists were aiming for? After all, the only thing that this would mean is that some exotic type of counterfactual linguistic community might decide to defer to their future linguistic practice on the meaning of a concept. That much has been clear from the very beginning. What TE rather purports to show is that *actual speakers like us sometimes defer to their future linguistic practice*. Ave was an interesting case for TE because the story of Druids exemplified the situation that actual speakers in fact face with respect to some concepts: for many concepts,

there are possible objects which we have never thought of, and it's unclear whether our concepts apply to them. In future, these objects might become actual and our linguistic community might come across them. If that happens, is the future response of our community to them what determines whether or not our concepts presently apply to those objects by determining their meaning?

When we tweaked Ave in §5.2 so that even the central function of 'ave' wasn't determinate enough, we still retained the relevant aspects of the thought experiment. If our concepts don't have clear applicability to some novel objects, it often seems to be so because it isn't clear whether they should apply to those objects *given their central function*. However, if we modify Ave so that Druids are completely indifferent about the ultimate value of their concepts, there is no longer any actual corollary to the thought experiment. Actual linguistic agents *do* care about whether their concepts are valuable representational tools at least to some extent: what concepts we have has a considerable impact on what sentences and thoughts we can articulate and communicate to one another, which clearly matters to us. Hence, Response 1 throws the baby out with the bathwater and thus ruins the point of Ave as a thought experiment supporting TE.

What about Response 2? Temporal externalists could argue that even if there are facts about what makes concepts ultimately valuable for their users, in Ave, the evaluative situation of Druids is such that both possible interpretations of 'ave' have equal relative weights, as the all-things-considered value of 'ave' remains the same under both interpretations. There are two ways in which such an equilibrium might be thought to arise. One possibility is that, by default, 'flying object' and 'flying and living object' are equally valuable interpretations of 'ave' in Druids' evaluative situation before t as described by the thought experiment. Another

possibility is that the two interpretations of ‘ave’ originally have different relative weights, but some further factors enter the evaluative situation of Druids and make them offset each other in their value.

The second possibility seems more viable than the first. As long as the setup of the thought experiment is such that ‘ave’ is the only concept in Druids’ language that they apply to flying objects around them before t, the following seems true: if ‘ave’ expresses the property of being a flying and living object, it will be such that whoever understands it (i.e., whoever is able to correctly apply ‘ave’ to objects around her and to explain why it applies to them) must be able to articulate in her language both the distinction between flying and non-flying objects as well as the distinction between living and unliving objects. By contrast, if ‘ave’ expresses the property of being a flying object, whoever understands it must be able to articulate only the distinction between flying and non-flying objects. Thus, since the former property is more fine-grained than the latter, conceptualising it would require Druids’ language to be more discriminatorily sensitive. This would likely prompt them to enrich their language by creating a separate concept expressing the property of being a flying object. But the reverse isn’t true: Druids could perfectly understand ‘ave’ as expressing the property of a flying object while being ignorant of the distinction between living and unliving objects. Hence, it seems that, contrary to the first possibility, ‘ave’ expressing the former property, *other things being equal*, makes it a more valuable concept for Druids because it has the greater potential of eventually making them capable of articulating finer distinctions in their language.

Therefore, to set Response 2 in motion, temporal externalists must flesh out Ave so that Druids’ evaluative situation involves some additional factors that will create a balance between the relative weights of the first and second interpretation of ‘ave’. For instance, they might

make it explicit that, before t, Druids already possess a separate concept expressing the property of being a living object and are thus able to articulate the distinction between living and unliving objects. However, this alone wouldn't be enough to guarantee that the two interpretations are equally weighty in Druids' choice situation. In addition, they would also have to stipulate that there are no other factors involved in Druids' evaluative situation that could tip the balance in favour of one interpretation, i.e., that, insofar as 'ave' preserves its original meaning, there will be no significant circumstances in which one interpretation fits better with what is valuable for Druids than the other.

Perhaps, if Druids knew that this was the case, their disposition to defer to their future use of 'ave' would be strong enough to motivate them to do so. There is, though, a similar problem to the one pressed against Response 1. It's far from obvious that it ever actually happens that all factors involved in our evaluative situations are such that all possible interpretations of an underdetermined concept like 'ave' are equally valuable for us. Moreover, given that these factors often include the complex facts about future (or possible) consequences that the concept will (or might) lead to under its various interpretations, even if such situations occasionally arise, it seems that we normally fail to recognise that we are in them. In such cases, we still unsuccessfully attempt to defer to the normative facts about which interpretation of a concept is ultimately the most valuable for us. Our concept then ends up indeterminate because no interpretation makes it more valuable for us.[78] Consequently, Response 2 also requires revising Ave to a form that doesn't show that we *in fact* sometimes defer to the future use of our concepts.

[78] Since such cases are relatively rare, this no longer gives rise to the over-generation of indeterminacy discussed at the end of §5.2.

This leaves temporal externalists with Response 3 which amounts to accepting NE. This response makes them acknowledge that normative facts about what makes a concept (and thus its central function) all-things-considered valuable may sometimes directly contribute to determining its meaning. This is because we defer to these facts on that part of the concept's meaning which is underdetermined by its dominant use and central function. Accordingly, normative encroachment on concepts' semantics is possible, which answers the last objection in Chapter 4. Temporal externalists might content themselves with this response because it still implies a weaker form of TE: even though the future use of a concept by a linguistic community doesn't directly determine its meaning, the normative facts that do so might still often depend on what future linguistic practice of the given community will be like. The reason is that what value a concept has often depends on what role it plays in the overall diachronic linguistic practice of the community of its users. Hence, a difference in what that practice will be like in the future might lead to a difference in what makes the concept valuable for its users.

One last qualification: it's true that Response 3 implies that at least one of the three intuitions that Ave is supposed to entice is false, as it implies that, before t, 'ave' means either 'flying object' or 'flying and living object' depending on which of the two interpretations is all-things-considered more valuable for Druids. There are then two options about what might happen when one of Scenario 1 and Scenario 2 gets actualised at t. The first option is that the original meaning of 'ave' gets overridden by the new dominant use of 'ave' in one of the scenarios. In that case, either Intuition 1 or Intuition 2 is false. The second option is that 'ave' still retains its original meaning in both scenarios despite that meaning being incompatible with its new dominant use. In such a case, Intuition 3 is false. I'm inclined to think that the first option is true, i.e., that the meaning of 'ave' adapts to its new dominant use even if it makes the concept less valuable for Druids. But I don't see why it's too much of a problem if either of Intuition 1

or Intuition 2 is false. After all, both these intuitions stem mainly from the fact that Druids don't notice that there is any change in the meaning of 'ave'. However, given that Druids aren't rational linguistic agents who are reflective about the meaning of 'ave', it's only to be expected that, even if 'ave' changes its meaning in one of the scenarios, they fail to notice it.

Chapter 6 - Conclusion

The objective of this thesis has been to argue that the model according to which the primary purpose of defining a concept in philosophy (and perhaps more broadly) is simply to report its actual meaning irrespective of wider normative considerations about what meaning makes it a valuable representational tool shouldn't satisfy us. I will now briefly summarise how each of the last three chapters addressing a separate issue contributed to this objective.

In Chapter 3, I attempted to show that the least controversial method of defining concepts according to CM faces a serious implementation problem in two respects. Firstly, those concepts that are uninstantiated in the actual world seem to be such that we cannot reliably identify a paradigmatic instance of the properties they express, which inhibits the completion of Step 1 of the method. Secondly, and more seriously, most concepts, including natural kind concepts, seem to express properties which are such that there is no privileged general criterion for determining what their possession consists in, which inhibits the completion of Step 2 of the method. Unless an alternative method of implementing CM that fares better than the PK-method is available, the problem suggests that there might be an ultimate limit to how far we can get by defining concepts according to CM.

In Chapter 4, I argued that, even setting aside the implementation problem, CM distorts what considerations are relevant to the act of defining a concept. I defended RM against the objection that defining a concept in response to a normative judgement about its all-things-considered value qua representational tool is to define it for the wrong kind of reason. My conclusion was that such judgements are fully consistent with the constitutive purpose of defining concepts. At the same time, my advocated account implies that, to the extent that the

central function of a concept makes it, all-things-considered, a valuable representational tool for us, we should aim to define the concept so that it can serve this function satisfactorily. The account also allows that the continuity between the central function of an old and new concept is desirable in most cases of conceptual revision. Therefore, it generally supports gradual and moderate revisions of concepts rather than rapid and radical revisions unless the all-things-considered value of the original concept is extremely negative.

In Chapter 5, I provided evidence for the thesis that the semantics of concepts is sensitive to wider normative facts about their value. The evidence was that the thesis can best account for some meta-semantic phenomena that are standardly thought to support TE. I recognise that the scope of my argument was limited because it demonstrated the sensitivity of a concept's semantics to such normative facts only in special cases when it's underdetermined by other meta-semantic factors. Having said that, I believe that my argument has given some initial impetus to the question on whether there is any encroachment of wider normative facts about concepts' value upon their semantics. This question certainly deserves more attention than it has received so far, especially if we want to find out whether conceptual engineering, pace RM, has any potential to lead not only to proposals on how the meanings of concepts should change but also to their actual implementation.[79]

This brings us to the end of this thesis. I hope that the readers are now more convinced that wider normative considerations about the value of concepts *do* matter when we are deciding how to define them in philosophy.

[79] An especially interesting question is whether, for the revision of a concept to succeed, it's enough if the concept's users defer on its meaning to an authority who proposed the revision or whether they must also have direct understanding of why the proposed revision makes the concept more valuable (Pollock, 2019, pp.94-95).

Bibliography

Appiah, K.A. (1996). Race, Culture, Identity: Misunderstood Connections. In: Appiah, K.A., and Gutmann, A. (eds.), *Color Conscious: The Political Morality of Race*. Princeton, NJ: Princeton University Press, pp.30-105.

Aristotle. (1933). *Metaphysics, Volume I: Book 1-9*. Translated by H. Tredennick. Cambridge, MA: Harvard University Press.

Aristotle. (1960). *Posterior Analytics, Topica*. Translated by H. Tredennick, and E.S. Forster. Cambridge, MA: Harvard University Press.

Audi, P. (2012). A Clarification and Defense of the Notion of Grounding. In: F. Correia, and B. Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*. Cambridge: Cambridge University Press, pp.101-121.

Ball, D. (2020a). Revisionary Analysis without Meaning Change (Or Could Women Be Analytically Oppressed?). In: H. Cappelen, D. Plunkett, and A. Burgess (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press, pp.35-59.

Ball, D. (2020b). Relativism, Metasemantics, and the Future. *Inquiry*, 63(9-10), pp.1036-1086.

Basu, R. (2019a). Radical Moral Encroachment: The Moral Stakes of Racist Beliefs. *Philosophical Issues*, 29(1), pp.9-23.

Basu, R. (2019b). The Wrongs of Racist Beliefs. *Philosophical Studies*, 176(9), pp.2497–2515.

Berto, F., and Jago, M. (2019). *Impossible Worlds*. Oxford: Oxford University Press.

Blum, L. (2002). *"I'm Not a Racist, but . . ."* *The Moral Quandary of Race*. Ithaca, NY: Cornell University Press.

Bolinger, R. J. (2020). Varieties of Moral Encroachment. *Philosophical Perspectives*, 34(1), pp.5-26.

Boyd, R. (1991). Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds. *Philosophical studies*, 61(1), pp.127-148.

Braddon-Mitchell, D. (2004). Masters of Our Meanings. *Philosophical Studies*, 118(1/2), pp.133-152.

Braddon-Mitchell, D., and Nola, R. (eds.). (2009). *Conceptual Analysis and Philosophical Naturalism*. Cambridge, MA: MIT Press.

Brigandt, I. (2010). The Epistemic Goal of a Concept: Accounting for the Rationality of Semantic Change and Variation. *Synthese*, 177(1), pp.19-40.

Brown, J. (2000). Against Temporal Externalism. *Analysis*, 60(2), pp.178-188.

Brun, G. (2020). Conceptual Re-Engineering: From Explication to Reflective Equilibrium. *Synthese*, 197(3), pp.925-954.

Burge, T. (1979). Individualism and the Mental. In: P.A. French, T.F. Uehling, and H.K. Wettstein. (eds.), *Midwest Studies in Philosophy IV: Studies in Metaphysics*. Minneapolis: University of Minnesota Press, pp.73-121.

Burge, T. (1989). Wherein is Language Social? In: G. Alexander (ed.), *Reflections on Chomsky*. Oxford: Blackwell Publishing, pp.175-192.

Burgess, A., and Plunkett, D. (2013a). Conceptual Ethics I. *Philosophy Compass*, 8(12), pp.1091-1101.

Burgess, A., and Plunkett, D. (2013b). Conceptual Ethics II. *Philosophy Compass*, 8(12), pp.1102-1110.

Cappelen, H. (2017). Why Philosophers Shouldn't Do Semantics. *Review of Philosophy and Psychology*, 8(4), pp.743-762.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Cappelen, H. (2020). Conceptual Engineering: The Master Argument. In: H. Cappelen, D. Plunkett, and A. Burgess (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press, pp.132-150.

Cappelen, H., Plunkett, D. and Burgess, A. (eds.). (2020). *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Cargile, J. (1991). Real and Nominal Definitions. In: J.H. Fetzer, D. Shatz, and N.G. Schlesinger (eds.), *Definitions and Definability: Philosophical Perspectives*. Netherlands: Springer, pp.21-50.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: The University of Chicago Press.

Chalmers, D. J., and Jackson, F. (2001). Conceptual Analysis and Reductive Explanation. *The Philosophical Review*, 110(3), pp.315-360.

Chalmers, D.J. (2002). Does Conceivability Entail Possibility?. In: T. Gendler, and J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford: Oxford University Press.

Chalmers, D.J. (2011). Verbal Disputes. *Philosophical Review*, 120(4), pp.515-566.

Chalmers, D.J. (2020). What is Conceptual Engineering and What Should It Be?. *Inquiry*, pp.1-18.

Chang, R. (2004a). All Things Considered. *Philosophical Perspectives*, 18, pp.1-22.

Chang, R. (2004b). Putting Together Morality and Well-Being. In: P. Baumann, and M. Betzler (eds.), *Practical Conflicts: New Philosophical Essays*. Cambridge: Cambridge University Press, pp.118-158.

Cohen, G.A. (2003). Facts and Principles. *Philosophy & Public Affairs*, 31(3), pp.211-245.

Cohen G.A. (2008). *Rescuing Justice and Equality*. Cambridge, MA: Harvard University Press.

Cohnitz, D., and Haukioja, J. (2020). Variation in Natural Kind Concepts. In: T. Marques, and A. Wikforss (eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variation*, Oxford: Oxford University Press, pp.128-146.

Collins, J. M. (2006). Temporal Externalism, Natural Kind Terms, and Scientifically Ignorant Communities. *Philosophical Papers*, 35(1), pp.55-68.

Correia, F. (2017). Real Definitions. *Philosophical Issues*, 27(1), pp.52-73.

Cowling, S. (2013). The Modal View of Essence. *Canadian Journal of Philosophy*, 43(2), pp.248-266.

Crisp, R. (2000). Value... and What Follows by Joel Kupperman. *Philosophy*, 75(3), pp.458-462.

Cummins, R. (1975). Functional Analysis. *The Journal of Philosophy*, 72(20), pp.741-765.

Dancy, J. (2002). *Practical Reality*. Oxford: Oxford University Press.

Davies, A., Kaplinski, L., and Lepamets, M. (2019). Meta-Semantic Moral Encroachment: Some Experimental Evidence. *Studia Philosophica Estonica*, 12, pp.7-33.

Denby, D. (2014). Essence and Intrinsicality. In: R. Francescotti (ed.), *Companion to Intrinsic Properties*. Berlin: De Gruyter, pp.87-109.

Deutsch, M. (2020). Speaker's Reference, Stipulation, and a Dilemma for Conceptual Engineers. *Philosophical studies*, 177(6), pp.1-23.

Díaz-Léon, E. (2016). Woman as a Politically Significant Term: A Solution to the Puzzle. *Hypatia*, 31(2), pp.245-258.

Dutilh Novaes, C. (2020). Carnapian Explication and Ameliorative Analysis: A Systematic Comparison. *Synthese*, 197(3), pp.1011-1034.

Dupré, J. (1981). Natural Kinds and Biological Taxa. *The Philosophical Review*, 90(1), pp.66-90.

Ebbs, G. (2000). The Very Idea of Sameness of Extension Across Time. *American Philosophical Quarterly*, 37(3), pp.245-268.

Egré, P., and O'Madagáin, C. (2019). Conceptual Utility. *The Journal of Philosophy*, 116(10), pp.525-554.

Eklund, M. (2015). Intuitions, Conceptual Engineering, and Conceptual Fixed Points. In: Ch. Daly (ed.), *The Palgrave Handbook of Philosophical Methods*. London: Palgrave Macmillan, pp.363-385.

Elder, C. L. (2004). *Real Natures and Familiar Objects*. Cambridge, MA: MIT Press.

Fine, K. (1994a). Essence and Modality: The Second Philosophical Perspectives Lecture. *Philosophical Perspectives*, 8(Logic and Language), pp.1-16.

Fine, K. (1994b) Senses of Essence. In: W. Sinnott-Armstrong (ed.), *Modality, Morality and Belief: Essays in Honor of Ruth Barcan Marcus*. Cambridge: Cambridge University Press.

Fine, K. (2001). The Question of Realism. *Philosophers' Imprint*, 1(1), pp.1-30.

Fine, K. (2012). Guide to Ground. In: F. Correia, and B. Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*. Cambridge: Cambridge University Press, pp.37-80.

Fine, K. (2015). Unified Foundations for Essence and Ground. *Journal of the American Philosophical Association*, 1(2), pp.296-311.

Fodor, J. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Franssen, M. (2006). The Normativity of Artefacts. *Studies in History and Philosophy of Science Part A*, 37(1), pp.42-57.

Fritz, J. (2017). Pragmatic Encroachment and Moral Encroachment. *Pacific Philosophical Quarterly*, 98(S1), pp.643-661.

Fritz, J. (2020). Moral Encroachment and Reasons of the Wrong Kind. *Philosophical Studies*, 177(10), pp.3051-3070.

Gauker, C. (2017). Open Texture and Schematicity as Arguments for Non-referential Semantics. In: K. Petrus, and S.J. Conrad (eds.), *Meaning, Context, and Methodology*. Berlin: Mouton de Gruyter. pp.13-30.

Glasgow, J. (2009). *A Theory of Race*. New York: Routledge.

Gupta, A. (2015). *Definitions*. In: E. Zalta (ed.), The Stanford Encyclopedia of Philosophy. [Online] Available at: https://plato.stanford.edu/entries/definitions/#ExpDef. [Accessed on 15.02.2020]

Hacking, I. (1995). The Looping Effects of Human Kinds. In: D. Sperber, D. Premack, and A.J. Premack. (eds.), *Causal cognition: A Multidisciplinary Debate (Symposia of the Fyssen Foundation)*. New York: Clarendon Press, pp.351-394.

Häggqvist, S. (2009). A Model for Thought Experiments. *Canadian Journal of Philosophy*, 39(1), pp.55-76.

Häggqvist, S., and Wikforss, Å. (2018). Natural Kinds and Natural Kind Terms: Myth and Reality. *The British Journal for the Philosophy of Science*, 69(4), pp.911-933.

Hale, B. (2013). *Necessary Beings: An Essay on Ontology, Modality, and the Relations between Them*. Oxford: Oxford University Press.

Hale, B. (2021). Essence and Definition by Abstraction. *Synthese*, 198(Suppl 8), pp.2001-2017.

Hart, H.L.A. (1961). *The Concept of Law*. Oxford: Clarendon Law Series.

Haslanger, S. (2000). Gender and Race: (What) Are They? (What) Do We Want Them to Be? *Noûs*, 34(1), pp.31-55.

Haslanger, S. (2003). Social Construction: The "Debunking" Project." *Socializing Metaphysics: The Nature of Social Reality*. In: F. Schmitt (ed.), Lanham, MD: Rowman and Littlefield, pp.301-325.

Haslanger, S. (2006). What Good are Our Intuitions? Philosophical Analysis and Social Kinds. *Proceedings of the Aristotelian Society Supplementary*, 80(1), pp.89-118.

Haslanger, S. (2012). *Language, Politics, and 'The Folk': Looking for 'The Meaning' of 'Race'. Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press, pp.429-445.

Haslanger, S. (2016). What is a (Social) Structural Explanation?. *Philosophical Studies*, 173(1), pp.113-130.

Haslanger, S. (2020a). Going on, Not in the Same Way. In: H. Cappelen, D. Plunkett, and A. Burgess (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press, pp.230-260.

Haslanger, S. (2020b). How Not to Change the Subject. In: T. Marques, and A. Wikforss (eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variation*, Oxford: Oxford University Press, pp.235-258.

Haukioja, J. (2015). On Deriving Essentialism from the Theory of Reference. *Philosophical Studies*, 172(8), pp.2141-2151.

Haukioja, J. (2020). Semantic Burden-Shifting and Temporal Externalism. *Inquiry*, 63(9-10), pp.919-929.

Hawthorne, J., and Manley, D. (2012). *The Reference Book*. Oxford: Oxford University Press.

Heuer, U. (2011). Beyond Wrong Reasons: The Buck-Passing Account of Value. In: M. Brady (ed.), *New Waves in Metaethics*, London: Palgrave Macmillan, pp.166-184.

Hendry, R. F. (2006). Elements, Compounds, and Other Chemical Kinds. *Philosophy of Science*, 73(5), pp.864–875.

Horwich, P. (2005). *Reflections on Meaning*. Oxford: Oxford University Press.

Hull, D. L. (1997). The Ideal Species Concept-and Why We Can't Get It. In: M.R. Claridge, H.A. Dawah, and M.R. Wilson (eds.), *Species: The Units of Biodiversity*, London: Chapman & Hall, pp.357-380.

Ichikawa, J., and Jarvis, B. (2009). Thought-Experiment Intuitions and Truth in Fiction. *Philosophical Studies*, 142(2), pp.221-246.

Isaac, M. G. (2021). Post-Truth Conceptual Engineering. *Inquiry*, pp.1-16.

Jabès, E. (1991). *The Book of Questions*. Translated by R. Waldrop. Middletown, Connecticut: Wesleyan University Press.

Jackson, F. (1994). Armchair Metaphysics. In: M. Murray, and J. O'Leary Hawthorne (eds.), *Philosophy in Mind: The Place of Philosophy in the Study of Mind*. Dordrecht: Springer, pp.23-42.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon Press.

Jackson, F., and Pettit, P. (2004). Functionalism and Broad Content. In: F. Jackson, P. Pettit, and M. Smith (eds.), *Mind, Morality and Explanation: Selected Collaborations*. Oxford: Oxford University Press, pp.95-118.

Jackman, H. (1999). We Live Forwards but Understand Backwards: Linguistic Practices and Future Behavior, *Pacific Philosophical Quarterly*, 80(2), pp.157-177.

Jackman, H. (2004). Temporal Externalism and Epistemic Theories of Vagueness. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 117(1/2), pp.79-94.

Jackman, H. (2005). Temporal Externalism, Deference, and Our Ordinary Linguistic Practice. *Pacific Philosophical Quarterly*, 86(3), pp.365-380.

Jackman, H. (2020). Temporal Externalism, Conceptual Continuity, Meaning, and Use. *Inquiry*, 63(9-10), pp.959-973.

Jago, M. (2021). Knowing How Things Might Have Been. *Synthese*, 198(Suppl. 8), pp.1981-1999.

Jenkins, K. (2016). Amelioration and Inclusion: Gender Identity and the Concept of Woman. *Ethics*, 126(2), pp.394-421.

Jorem, S. (2021). Conceptual Engineering and the Implementation Problem. *Inquiry*, 64(1-2), pp.186-211.

Jylkkä, J., Railo, G., and Haukioja, J. (2009). Psychological Essentialism and Semantic Externalism: Evidence for Externalism in Lay Speakers' Language Use. *Philosophical Psychology*, 22(1), pp.37-60.

Kaplan, D. (1989). Demonstratives. In: J. Almog, J. Perry, and H. Wettstein. (eds.), *Themes from Kaplan*. Oxford: Oxford University Press, pp.481-563.

Khalidi, M. A. (2013). *Natural Categories and Human Kinds: Classification in the Natural and Social Sciences*. Cambridge: Cambridge University Press.

Koch, S. (2021). The externalist challenge to conceptual engineering. *Synthese*, 198, pp.1-22.

Koethe, J. (1982). The Stability of Reference Over Time. *Noûs*, 16(2), pp.243-252.

Korman, D. Z. (2006). What Externalists Should Say about Dry Earth. *The Journal of Philosophy*, 103(10), pp.503-520.

Korsgaard, C. M. (2008). *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford: Oxford University Press.

Korsgaard, C.M. (2009). *Self-constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

LaPorte, J. (2004). *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Law, S. (2016). Natural Kinds of Substance. *Australasian Journal of Philosophy*, 94(2), pp.283-300.

Lewis, D. (1970). How to Define Theoretical Terms. *Journal of Philosophy*, 67(13), pp.427-446.

Lewis, D. (1972). Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy* 50(3), pp.249-258.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell Publishing.

Lewis, D. (1984). Putnam's Paradox. *Australasian Journal of Philosophy*, 62(3), pp.221-236.

Livingstone-Banks, J. (2017). In Defense of Modal Essentialism. *Inquiry*, 60(8), pp.1-27.

Locke, J. (1975). In: P.H. Nidditch (ed.), *The Clarendon Edition of the Works of John Locke: An Essay Concerning Human Understanding*. Oxford: Oxford University Press.

Lowe, E. J. (2011). Locke on Real Essence and Water as a Natural Kind: A Qualified Defence. *Proceedings of Aristotelian Society Supplementary Volume*, 81(1), pp.1-19.

Lowe, E.J. (2012) What is the Source of our Knowledge of Modal Truths? *Mind*, 121(484), pp.919–950.

Löhr, G. (2021). Social Constructionism, Concept Acquisition and the Mismatch Problem, *Synthese*, 198, pp.659-673.

Lycan, W. (1991). Definitions in a Quinean world. In: J.H. Fetzer, D. Shatz, and N.G. Schlesinger (eds.), *Definitions and Definability: Philosophical Perspectives*. Netherlands: Springer, pp.111-131.

Mackie, P. (2006). *How Things Might Have Been: Individuals, Kinds, and Essential Properties*. Oxford: Oxford University Press.

Mallon, R. (2017). Social Construction and Achieving Reference. *Noûs*, 51(1), pp.113-131.

Malmgren, A. S. (2011). Rationalism and the Content of Intuitive Judgements. *Mind*, 120(478), pp.263-327.

Margolis, E., and Laurence, S. (2007). The Ontology of Concepts — Abstract Objects or Mental Representations?. *Noûs*, 41(4), pp.561–593.

McNaughton, D., and Rawling, P. (2018). Motivating Reasons and Normative Reasons. D. Star (ed.), *The Oxford Handbook of Reasons and Normativity*, pp.171-194.

Mikkola, M. (2009). Gender Concepts and Intuitions. *Canadian Journal of Philosophy*, 39(4), pp.559-583.

Millikan, R.G. (1984). *Language, Thought, and Other Biological Categories.* Cambridge, MA: MIT Press.

Moore, G.E. (1942) A Reply to My Critics. In: P.A. Schilpp (ed.), *The Philosophy of G.E. Moore*. Evanston: Open Court.

Nado, J. (2021). Conceptual Engineering, Truth, and Efficacy. *Synthese*, 198(Suppl 7), pp.1507-1527.

Needham, P. (2011). Microessentialism: What Is the Argument? *Noûs*, 45(1), pp.1–21.

Nimtz, C. (2004). Two-Dimensional and Natural Kind Terms. *Synthese*, 138(1), pp.125-148.

Nolan, D. (2009). Platitudes and Metaphysics. In: D. Braddon-Mitchell, and R. Nola (eds.), *Conceptual Analysis and Philosophical Naturalism*. Cambridge, MA: MIT Press, pp.268-299.

Nolan, D. (2014). Hyperintensional Metaphysics. *Philosophical Studies*, 171(1), pp.149-160.

Parfit, D. (2001). Rationality and Reasons. In: D. Egonsson, J. Josefsson, B. Petersson, and T. Rønnow-Rasmussen (eds.), *Exploring Practical Philosophy: From Action to Values*, Aldershot: Ashgate, pp.17-39.

Parfit, D. (2011). *On What Matters (Vol. 1)*. Oxford: Oxford University Press.

Peacocke, C. (1992). *A Study of Concepts*. Cambridge, MA: MIT Press.

Persson, I. (2007). Primary and Secondary Reasons. In: T. Rønnow-Rasmussen, J. Josefsson, and D. Egonsson. (eds.), *Hommage á Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*. [Online] Available at: http://www.fil.lu.se/hommageawlodek [Accessed on 08.06.2021]

Pettit, P. (2020). Analyzing Concepts and Allocating Referents. In: H. Cappelen, D. Plunkett, and A. Burgess (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press, pp.333-357.

Podosky, P.M.C. (2018). Ideology and Normativity: Constraints on Conceptual Engineering. *Inquiry*, pp.1-16.

Pollock, J. (2019). Conceptual Engineering and Semantic Deference. *Studia Philosophica Estonica*, 12, pp.81-98.

Piller, C. (2006). Content-Related and Attitude-Related Reasons for Preferences. *Royal Institute of Philosophy Supplements*, 81(59), pp.155-81.

Pinder, M. (2019). Conceptual Engineering, Metasemantic Externalism and Speaker-Meaning. *Mind*, 130(517), pp.1-23.

Plunkett, D. (2015). Which Concepts Should We Use?: Metalinguistic Negotiations and the Methodology of Philosophy. *Inquiry*, 58(7-8), pp.828-874.

Plunkett, D., and Sundell, T. (2019). Metalinguistic Negotiation and Speaker Error. *Inquiry*, 64(1-2), pp.1-26.

Prinzing, M. (2018). The Revisionist's Rubric: Conceptual Engineering and the Discontinuity Objection. *Inquiry*, 61(8), pp.854-880.

Putnam, H. (1973). Meaning and Reference. *The Journal of Philosophy*, 70(19), pp.699-711.

Putnam, H. (1975). The Meaning of 'Meaning'. In: K, Gunderson. (ed.), *Language, Mind, and Knowledge. Minnesota Studies in the Philosophy of Science, Volume 7*. Minneapolis: University of Minnesota Press, pp.131-193.

Putnam, H. (1990). *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge: Cambridge University Press.

Queloz, M. (2021). *The Practical Origins of Ideas: Genealogy as Conceptual Reverse-Engineering*. Oxford: Oxford University Press.

Rabinowicz, W., and Rønnow-Rasmussen, T. (2004). The Strike of the Demon: On Fitting Pro-attitudes and Value. *Ethics*, 114(3), pp.391-423.

Raven, M. J. (2015). Ground. *Philosophy Compass*, 10(5), pp.322-333.

Rawls, J. A. (1999a). *Theory of Justice*. Oxford: Oxford University Press.

Rawls, J.A. (1999b). *The Law of Peoples*. Cambridge, MA.: Harvard University Press.

Raz, J. (1997). Incommensurability and Agency. In: Chang, R. (ed.), *Incommensurability, Incomparability, and Practical Reason*. Cambridge, MA.: Harvard University Press, pp.110-128.

Raz, J. (2009). Reasons: Practical and Adaptive. In: D. Sobel, and S. Wall. (eds.), *Reasons for Actions*. Cambridge: Cambridge University Press, pp.37-57.

Richards, R.A. (2010). *The Species Problem: A Philosophical Analysis*. Cambridge: Cambridge University Press.

Robinson, R. (1963). *Definition*. Oxford: Clarendon Press.

Roca-Royes, S. (2011). Conceivability and De Re Modal Knowledge. *Noûs*, 45(1), pp.22-49.

Rosen, G. (2015). Real Definition. *Analytic Philosophy*, 56(3), pp.189-209.

Rouse, J. (2014). Temporal Externalism and the Normativity of Linguistic Practice. *Journal of the Philosophy of History*, 8(1), pp.20-38.

Rowland, R. (2013). Wrong Kind of Reasons and Consequences. *Utilitas*, 25(3), pp.405-416.

Salmon, N. (1981). *Reference and Essence*. Princeton, NJ: Princeton University Press.

Salmon, N. (1998). Nonexistence. *Noûs*, 32(3), pp.277-319.

Samuelsson, L. (2013). The Right Version of 'the Right Kind of Solution to the Wrong Kind of Reason Problem. *Utilitas*, 25(3), pp.383-404.

Saul, J. (2006). Gender and Race. *Proceedings of the Aristotelian Society Supplementary Volume*, 80(1), pp.119-143.

Sawyer, S. (2018). VI—The Importance of Concepts. *Proceedings of the Aristotelian Society*, 118(2), pp.127-147.

Scanlon, T. (2000). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schaffer, J. (2009). On What Grounds What. In: D. J. Chalmers, D. Manley, and R. Wasserman (eds.), *Metametaphysics*. New York: Oxford University Press, pp.347-383.

Schroeder, M. (2010). Value and the Right Kind of Reason. In: R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Volume 5*. New York: Oxford University Press, pp.25-55.

Schroeter, L. (2004). The Limits of Conceptual Analysis. *Pacific Philosophical Quarterly*, 85(4), pp.425-453.

Schroeter, L., and Schroeter, F. (2015). Rationalizing Self-Interpretation. In: Ch. Daly (ed.), *The Palgrave Handbook of Philosophical Methods*. London: Palgrave Macmillan, pp.419-447.

Searle, J.R. (1995). *The Construction of Social Reality*. New York: Free Press.

Sen, A. (2006). What Do We Want from a Theory of Justice?. *Journal of Philosophy*, 103(5), pp.215–38.

Sharadin, N. (2016). Reasons Wrong and Right. *Pacific Philosophical Quarterly*, 97(3), pp.371-399.

Shields, M. (2020). Conceptual Change in Perspective. *Inquiry: An Interdisciplinary Journal of Philosophy*, 63(9-10), pp.930-958.

Shields, M. (Forthcoming). On Stipulation. *European Journal of Philosophy*. [Online] Available at: https://philpapers.org/rec/SHIOS [Accessed on 11.06.2021]

Sider, T. (2011). *Writing the Book of the World*. Oxford: Oxford University Press.

Sidelle, A. (1989). *Necessity, Essence, and Individuation: A Defense of Conventionalism*. Ithaca, NY: Cornell University Press.

Simion, M. (2018). The 'should' in Conceptual Engineering. *Inquiry*, 61(8), pp.914-928.

Simion, M., and Kelp, C. (2020). Conceptual Innovation, Function First. *Noûs*, 54(4), pp.985-1002.

Sorensen, R.A. (1998). *Thought Experiments*. New York: Oxford University Press.

Sorensen, R.A. (2001). *Vagueness and Contradiction*. New York: Oxford University Press.

Spinoza, B. (1997). *On the Improvement of the Understanding*. New York: Dover Publications, Inc.

Stalnaker, R.C. (1968). A Theory of Conditionals. In: N. Rescher, (ed.), *Studies in Logical Theory*. Oxford: Basil Blackwell, pp.98-112.

Steward, H. (1990). Identity Statements and the Necessary a Posteriori. *The Journal of Philosophy*, 87(8), pp.385-398.

Stoneham, T. (2003). Temporal Externalism. *Philosophical Papers*, 32(1), pp.97-107.

Stratton-Lake, P. (2005). How to Deal with Evil Demons: Comment on Rabinowicz and Rønnow-Rasmussen. *Ethics*, 115(4), pp.788-798.

Strawson, P. F. (1963). Carnap's Views on Constructed Systems Versus Natural Languages in Analytic Philosophy. In: P.A. Schilpp (ed.), *The Philosophy of Rudolf Carnap*. Lasalle, IL: Open Court, pp.503-518.

Suikkanen, J. (2009). Buck-Passing Accounts of Value. *Philosophy Compass*, 4(5), pp.768-779.

Sveinsdóttir, Á.K. (2013). The Social Construction of Human Kinds. *Hypatia*, 28(4), pp.716-732.

Tahko, T.E. (2015). Natural Kind Essentialism Revisited. *Mind*, 124(495), pp.795-822.

Tahko, T. E. (2018). The Epistemology of Essence. In: A. Carruth, S. Gibb, and J. Heil. (eds.), *Ontology, Modality, Mind: Themes from the Metaphysics of E.J. Lowe*, Oxford: Oxford University Press, pp.93-110.

Tanesini, A. (2014). Temporal Externalism: A Taxonomy, an Articulation, and Defence. *Journal of the Philosophy of History*, 8(1), pp.1-19.

Thomasson, A.L. (1993). The Reference of Fictional names. *Kriterion*, 6, pp.3-12.

Thomasson, A.L. (2020). A Pragmatic Method for Normative Conceptual Work. In: H. Cappelen, D. Plunkett, and A. Burgess. (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press, pp.435-458.

Thompson, N. (2016). Grounding and Metaphysical Explanation. *Proceedings of the Aristotelian Society*, 116(3), pp.395-402.

Tichý, P. (1988). *The Foundations of Frege's Logic*. Reprint 2012. Berlin and New York: De Gruyter.

van Brakel, J. (1986). The Chemistry of Substances and the Philosophy of Mass Terms. *Synthese*, 69(3), pp.291-324.

van Inwagen, P. (1998). Modal Epistemology. *Philosophical Studies*, 92(1), pp.67-84.

Waismann, F. (1945). Symposium: Verifiability. *Proceedings of the Aristotelian Society: Supplementary Volume*, 19, pp.119-150.

Waldron, J. (1999). *Law and Disagreement*. Oxford: Clarendon University Press.

Wei, F., Hu, Y., Zhu, L., Bruford, M.W., Zhan, X., and Zhang, L. (2012). Black and white and read all over: the past, present and future of giant panda genetics. *Molecular Ecology*. *21*(23), pp.5660-5674.

Weisberg, M. (2005). Water Is Not H_2O. In: D. Baird, E. Scerri, and L. McIntyre (eds.), *Philosophy of Chemistry: Synthesis of a New Discipline*, Dordrecht: Springer, pp.337-45.

Wildman, N. (2013). Modality, Sparsity, and Essence. *Philosophical Quarterly*, 63(253), pp.760-782.

Williamson, T. (1994). *Vagueness*. London: Routledge.

Williamson, T. (2007). *The Philosophy of Philosophy*. Oxford: Blackwell Publishing.

Wilson, M. (1982). Predicate Meets Property. *The Philosophical Review*. 91(4), pp.549-89.

Yablo, S. (1993). Is Conceivability a Guide to Possibility?. *Philosophy and Phenomenological Research*, 53(1), pp.1-42.

Zalta, E. (2001). Fregean Senses, Modes of Presentation, and Concepts. *Philosophical Perspectives,* 15(Noûs Supplement), pp.335-359.