Reports from the Department of Philosophy
Vol. 21

CONCEPTS AND REFERENCE

Defending a Dual Theory of Natural Kind Concepts

Jussi Jylkkä

University of Turku
Finland

SERIES EDITORS

Olli Koistinen
Juha Räikkä

Department of Philosophy
University of Turku
FIN-20014 Turku
Finland

ISSN 1457-9332
ISBN 978-951-29-3699-1 (Print)
ISBN 978-951-29-3700-4 (Pdf)

Painosalama Oy – Turku, Finland 2008

Acknowledgements

I wish to thank the philosophy department at the University of Turku, which has provided me a friendly and stimulating working environment. Especially I would like to thank our secretary Hanna-Mari Salonen, who has helped tremendously in dealing with any kind of practical matters.

I am indebted to all those researchers who I have met in conferences and workshops, and with whom I have had the chance to discuss the topics of this thesis, especially (in alpabetical order) Nick Braisby, Daniel Cohnitz, Sören Häggqvist, Antti Kauppinen, Joseph LaPorte, Edouard Machery, Christian Nimtz, Panu Raatikainen, Tobias Rosefeldt, Benjamin Schnieder, Michael Strevens, and Åsa Wikforss.

I would like to thank my professor Olli Koistinen, whose inspiring lectures contributed greatly to my starting studying philosophy in the first place. I am most indebted to my good friend Henry Railo, the co-author of an article in this thesis, without whom conducting the experimental work would have been impossible. Thanks go also to Henri Pettersson for discussions, comments, and travelling company. Above all, I am deeply grateful to my supervisor Jussi Haukioja for his guidance and assistance throughout my work, and for inviting me to take part in his project.

I am indebted to Panu Raatikainen, who reviewed the manuscript of this thesis. I would also like to express my deepest gratitude to professor Stephen Laurence from the University of Sheffield for his rigorous review of the manuscript, and for his being kind enough to serve as the opponent in the disputation.

During the preparation of this thesis I have, together with Henri Pettersson, been working in the research project *Externalism about Meaning and Mental Content* (number 214088) led by Jussi Haukioja and funded by the Academy of Finland. The project has provided me the privilege to write this thesis uninterruptedly for three years.

Finally, I would like to thank my parents, who have always encouraged me to find my own way, and Satu for her love and support.

Turku 2008

Jussi Jylkkä

Abstract

In this thesis I argue that the psychological study of concepts and categorisation, and the philosophical study of reference are deeply intertwined. I propose that semantic intuitions are a variety of categorisation judgements, determined by concepts, and that because of this, concepts determine reference. I defend a dual theory of natural kind concepts, according to which natural kind concepts have distinct semantic cores and non-semantic identification procedures. Drawing on psychological essentialism, I suggest that the cores consist of externalistic placeholder essence beliefs. The identification procedures, in turn, consist of prototypes, sets of exemplars, or possibly also theory-structured beliefs. I argue that the dual theory is motivated both by experimental data and theoretical considerations. The thesis consists of three interrelated articles. Article I examines philosophical causal and description theories of natural kind term reference, and argues that they involve, or need to involve, certain psychological elements. I propose a unified theory of natural kind term reference, built on the psychology of concepts. Article II presents two semantic adaptations of psychological essentialism, one of which is a strict externalistic Kripkean-Putnamian theory, while the other is a hybrid account, according to which natural kind terms are ambiguous between internalistic and externalistic senses. We present two experiments, the results of which support the strict externalistic theory. Article III examines Fodor's influential atomistic theory of concepts, according to which no psychological capacities associated with concepts constitute them, or are necessary for reference. I argue, contra Fodor, that the psychological mechanisms are necessary for reference.

Keywords: Semantic internalism, semantic externalism, causal theory of reference, description theory of reference, psychological essentialism, prototype theory, psychosemantics, concept atomism, informational semantics, rigidity, compositionality.

Contents

Part I: Introductory Essay

Part II: Original Articles

List of Original Publications

In the text, articles are referred to as ‘articles’ with Roman numerals I – III.

I Jylkkä, J. (2008). Theories of natural kind term reference and empirical psychology. *Philosophical Studies, 139*, 153 – 169.

II Jylkkä, J., Railo, H., & Haukioja, J. (forthcoming). Psychological essentialism and semantic externalism: evidence for externalism in lay speakers’ language use. *Philosophical Psychology*.

III Jylkkä, J. (forthcoming). Why Fodor’s theory of concepts fails. *Minds & Machines*.

The author acknowledges with thanks permission granted to reproduce the original articles.

A Note on Terminology and Notation

‘Single quotes’ are used when referring to terms and *italics* are used when referring to kinds, properties, and descriptions. Italics are also used for emphasis. Small caps are used to refer to concepts. By ‘concept’ I mean a mental representation; for instance, the concept CAT is a mental representation of cats. By ‘category’ I mean the set of objects that a concept applies to; for instance, the concept CAT applies or refers to whatever belongs in the category of cats.

Part I: Introductory Essay

1. Introduction

Concepts are mental representations of categories of objects, and they enable us to recognise and make inferences about the objects belonging in the represented category. For instance, my concept CAT determines which creatures I am disposed to count as belonging in the category of cats. If my concept of cat consists of representations of features *meows*, *has whiskers*, and *preys mice*, then I am disposed to categorise as a cat anything that meows, has whiskers, and preys mice. Possessing the concept CAT also enables me to make inferences about cats: I know that if an object is a cat, then it is likely to meow, have whiskers, and prey mice. Concepts make the world an intelligible whole to us. They organise the vast complexity of individual objects and properties around us into easily conceivable categories. For example, when looking out of the window, I see trees, grass, the sky, clouds, and people instead of simply a jumble of individual objects and qualities. Without concepts each object I meet would be completely novel to me; say, I would have to acquaint myself with each and every cat over and over again. I would have to learn of each individual cat that it meows, has whiskers, and preys mice, but I wouldn't be able to generalise my knowledge of individual cats to cats in general. In fact, *cats* as a category would not even exist for me, but rather I would only perceive individual objects that don't have anything in common.

Concepts are often assumed to determine not only which objects we *reckon* as belonging in categories, but also which objects *truly* belong in them. In other words, concepts are often supposed to determine not only categorisation, but also reference. The two notions are not equivalent: people may be disposed to categorise in the extension of a concept an object that the concept does not really refer to, and people may *fail* to categorise in the extension of a concept an object that the concept *does* refer to. For example, I might categorise a cleverly constructed cat-like robot as a (real) cat, even though the creature is not really a cat. Again, I might fail to categorise a real cat as a cat, if it happened to be cleverly disguised and trained to look and behave like a skunk. Reference is a normative relation that concerns what a concept truly applies to, whereas categorisation concerns how speakers as a matter of fact happen to use concepts. Accordingly, reference is typically studied in armchair philosophy and categorisation in experimental psychology.

These are some discrepancies between categorisation and reference, but do the two notions have anything in common? Some theorists have argued that reference and categorisation are wholly separate, and that philosophy and psychology examine concepts from wholly distinct perspectives—philosophical study about reference is *normative* in that it is concerned about the *correct* use of concepts, whereas psychology studies how concepts are *as a matter of fact* used (see Rey 1983; Smith, Medin, & Rips 1984; Rey 1985). Yet other theorists take that concepts, conceived in the psychological sense, determine both reference and categorisation (Laurence and Margolis 1999). This is the view that I will also defend.

1.1. Concepts determine reference

What a concept truly applies to is not distinct from what we take it to apply to. Reference is not some mind-independent fact, but what a concept refers to is determined by our intuitions about how we would apply the term in various situations, actual and possible. Consider, for instance, the case against the *description theory of reference* in the philosophy of language. According to the description theory, a term applies to whatever satisfies the description, or cluster of descriptions, associated with it in the mind of the speakers. For instance, suppose that a speaker associates with the term 'gold' a description such as *the yellow, shiny, malleable, metallic substance*. This description arguably manages to pick out gold (that is, the element Au) in everyday circumstances, but not necessarily: gold could be heated so that it becomes gaseous, and would no longer be yellow, shiny, malleable, nor metallic (Laurence and Margolis' (1999) example). Despite failing to satisfy the description associated with the term 'gold', the term would arguably still *refer* to the gas—after all, it is Au and is produced by heating gold. This counts as evidence against the description theory of reference. How do we know that gaseous Au is gold? Philosophers typically call our knowledge of this fact *intuitive*, but the notion of intuition is of little help here, since there is no agreement on what philosophical intuition specifically is (see e.g. the theme number on philosophical knowledge in *Grazer Philosophische Studien, 74*, 2007). Philosophical theories of intuition aside, we may make a simple observation: *intuition about reference is categorisation*. In intuiting that, in the above scenario, the term 'gold' refers to gaseous Au, I am simply making a categorisation judgement that gaseous Au is gold.

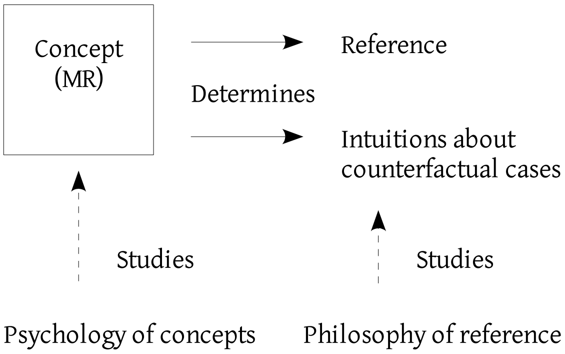
Obviously philosophical intuition is different from typical, everyday categorisation behaviour; it is more considered and immutable. To compare, upon observing a sample of a gold-like substance with the naked eye and categorising it as 'gold', there is always the possibility of error—the sample might be fool's gold or some other non-Au substance. On the other hand, in philosophical thought experimentation we can determine all the facts possibly affecting our categorisation judgement, which enables us to discover the ultimate criteria for category membership. We may make presumptions, independent of contingent matters of fact: *if* this sample of a gold-like substance shares some deep structure with the samples ordinarily called ‘gold’, *then* it really is gold; if it doesn’t, then it is not gold. Philosophising about reference is trying to find out all the facts possibly affecting what we take to belong in the extension of a term, and systematising them into a theory that specifies the necessary and sufficient conditions for the correct application of a term. Building a theory of reference can be compared to experimental study in psychology: we formulate hypotheses about what a term refers to (say, that it refers via descriptions, or causal chains), and test these hypotheses in the face of what we take to belong in the term’s extension in various counterfactual scenarios.

Since concepts determine our categorisation judgements, they also determine our philosophical intuitions about counterfactual cases. If we suppose that our intuitions about counterfactual cases *constitute* or *determine* reference, then we must conclude that concepts determine reference. And it is reasonable, if not even inevitable, to suppose that intuitions do constitute reference; for if they didn’t, they would not serve as evidence for or against theories of reference at all—say, it wouldn’t matter at all for the reference of ‘gold’ whether we take gaseous Au to belong in the extension of the term ‘gold’ or not.

Noticing that concepts determine our reference-constituting intuitions has two important consequences. The first is mostly relevant for the psychological study of concepts. That concepts determine reference makes it intelligible to study concepts from the semantic viewpoint. If a theory of concepts cannot account for reference, then it cannot possibly be a *complete* theory of concepts. A complete theory should both account for reference and also explain why we have such intuitions about philosophical counterfactual cases as we in fact do. The second consequence is important mainly from the perspective of the philosophical study of reference. Traditionally reference is studied through reflecting on our intuitions about possible cases, and theories of reference are typically purely abstract models of these intuitions, not taking a stand on what *causes* these intuitions. That is, philosophical study of reference models our intuitions about reference, but does not explain what it is about *concepts* that cause these intuitions.

Consider, for example, the Kripkean causal theory of reference. On this account, the reference of a natural kind term is fixed in an initial *baptism*—say, the term ‘gold’ is introduced by pointing to some usual samples of gold and stipulating that ‘[g]old is the substance instantiated by the items over there, or at any rate, by almost all of them’ (Kripke 1980, p. 135). After this baptism, the term refers solely to samples of the kind instantiated by the samples pointed at. Supposing that this kind is Au, the term ‘gold’ then refers solely to instances of Au. This entails that, for instance, gaseous Au belongs in the extension of ‘gold’, and thus the account (unlike the description theory) is concordant with our intuition that gaseous Au is gold. But what does Kripke’s account really teach us, or what does it explain? The predictions of the causal theory may be in line with our intuitions about reference, but the theory does nothing to explain *why* we have those intuitions. If concepts are the causes of our reference-constituting intuitions, then a (complete) theory of concepts would explain why we have such intuitions about counterfactual cases as we in fact do, and would thus explain *what makes* terms refer like they do. Philosophical theories of reference investigate merely the *effects* (that is, intuitions) of what in the first place determines reference, whereas psychology of concepts investigates the *causes* of reference. (Figure I illustrates this position.) This raises doubts about the sensibility of the traditional philosophical study of reference.

Figure I. ‘MR’ stands for mental representation.



I do not wish to claim that the philosophical study of reference would be completely illegitimate. It can point towards new aspects about language use by pointing out interesting counterfactual cases (consider the case of gaseous gold), and philosophical theories of reference may serve as guides about why we have certain kind of referential intuitions. For example, the causal theory suggests that at least natural kind concepts may have a causal, or externalistic, element in them. However, I take it that building an adequate theory of reference using purely *a priori* methods cannot succeed. Philosophers are deemed to investigate merely the effects (that is, intuitions) of the real determinants of reference (that is, concepts), and can thus never find out what *really* determines reference. In order to build a complete theory of reference, we must consult experimental psychology.

My proposal could be objected to as follows. I am claiming that psychologists study the causes of reference, whereas philosophers only model the effects of the causes of reference; that is, intuitions. But aren't psychological accounts of concepts based just on intuitions as well: roughly, psychologists present subjects with descriptions or images of objects and ask whether they belong in a category or not. Aren't the subjects' responses to such tasks intuitions as well, comparable to those of philosophers? Just like a philosopher is imagining in her mind a hypothetical case and asks herself what a term applies to in that situation, isn't a psychologist doing just the same thing, except for that the psychologist is using many subjects and an experimental setup? I grant that even the psychologists are deemed to investigate concepts only through their effects (that is, the subjects' categorisation judgements), but I deny that this would make philosophical and psychological study of concepts equal. Consider, for instance, the experimental evidence against the definition theory of concepts. The definition theory holds that a concept is a definition specifying the singly necessary and jointly sufficient criteria for category membership; say, the concept BACHELOR consists of the definition *unmarried adult male*. However, experimental study has revealed that people tend to judge various instances of a category as more or less typical than others; Casanova is more typical a bachelor than a monk, or a car is more typical a vehicle than a tricycle. The judged typicality of a category member explains a range of other categorisation phenomena, such as the speed in categorising objects as instances of the category, or the ease of learning or recalling an object or a subcategory of the concept. These findings suggest that concepts are *prototypes* instead of definitions; that is, probabilistic sets of features where each feature has a specific typicality value. Specific prototype theories to date have become extremely sophisticated. It is highly improbable that a philosopher could ever have ended up in similar models purely on basis of armchair reflection. Psychologists' use of experimental methods gives them a strong advantage over philosophers, who are utilising introspection and thought experimentation, and makes them much more likely to achieve a correct account of reference.

I do not wish to argue in detail here why the psychological, or psychologically motivated study of reference is superior to purely philosophical—that is a metaphilosophical question that cannot be settled in this introductory essay. I propose to endorse a more practical strategy:

let's not immerse ourselves in endless philosophical debates about the study of reference, but rather let's start *studying* reference by investigating the psychological accounts of concepts. This is what I intend to do in this thesis; I will leave methodological and metaphilosophical issues to others.

1.2. Ignorance and error problems, and semantic externalism

Investigating concepts from the semantic perspective is by no means unproblematic. There is especially one class of problems that weakens the ground for the semantic study of concepts; these are the so-called *ignorance and error problems*, and, more generally, *semantic externalism*. The ignorance and error problems and semantic externalism are originally formulated in the highly influential writings of Saul Kripke (1980) and Hilary Putnam (1975), and they are raised against the psychological theories of concepts by Stephen Laurence and Eric Margolis (1999).

The traditional psychological accounts equate concepts with some set of identificatory knowledge about the category members. For example, the definition theory holds that a concept consists of a conjunctive definition which specifies the singly necessary and jointly sufficient conditions for a sample's being categorised as, or belonging in the concept's extension (on this account, concepts are taken to determine both categorisation and reference). According to the definition theory, an object falls in the extension of the concept if and only if it possesses each and every of the features. Now, the ignorance problem is that the identificatory knowledge represented in a concept may be insufficient to delineate the true extension of the concept. We already examined one problem of this kind, namely the case of gaseous gold. The example was raised against the description theory of reference, but it applies equally to the definition theory of concepts. Suppose that the concept GOLD consists of a definition such as *the yellow, shiny, and metallic substance*. Now, the concept arguably refers even to gaseous gold, even though it is neither yellow, shiny, nor metallic. On the other hand, the concept does *not* refer to fool's gold (or some other non-Au gold-like substance) which is indeed yellow, shiny, and metallic. The problem of error, on the other hand, is that the identificatory information a concept consists of may be erroneous. For example, earlier it was believed that gold is a compound, or that whales are fish, or that water is an element, but despite these false beliefs, the terms still arguably referred to gold (Au), whales (certain mammalian sea animals), and water (H_2O), respectively.

The ignorance and error problems can be considered merely symptoms of a yet more profound problem pertaining to the traditional theories of concepts. There is wide agreement in the philosophical circles that some terms (or the corresponding concepts) refer *externalistically*: the reference of these terms is not determined solely by the mental states of the speakers, but also by some external matters of fact (Kripke 1980; Putnam 1975). Externalistically referring terms include at least proper names and natural kind terms; in this thesis I will focus solely on natural kind terms. There is no generally accepted definition for what natural kind terms specifically are, but typically they are taken to include at least terms referring to natural substances, kinds, and phenomena, and biological species. These terms include, among others, 'cat' 'tiger, 'gold', 'water', 'iron pyrites', 'heat', 'light', 'sound', and 'lightning' (Kripke 1980, p. 134). I will take it that if natural kind terms refer externalistically, then so do their corresponding concepts.

Externalism about natural kind terms can be illustrated with the help of Putnam's famous *Twin Earth thought experiment*. Putnam asks us to imagine a distant planet, call it 'Twin Earth', where almost everything is exactly like on Earth. Twin Earth is inhabited by people who are psychologically qualitatively identical to us, and who speak a language just like English. There occurs a substance on Twin Earth that is in almost all respects just like our water: it is odourless, colourless, thirst-quenching, life-supporting, fills seas and lakes, and so on. The twin earthlings call this substance 'water'. However, there is one important difference between Earth and Twin Earth: whereas the substance called 'water' on Earth is H_2O, the respective substance on Twin Earth is an altogether different, complex chemical substance, *XYZ*. Finally, place both Earth and Twin Earth in the year 1750 when the chemical composition of the substance called 'water' was not known on either planet. Now Putnam asks: did twin earthlings and earthlings refer to the same substance with their term 'water'? Putnam's answer, widely accepted in the philosophical community, is negative: despite being identical in their mental states (narrowly construed), earthlings referred with their term 'water' solely to H_2O, whereas twin earthlings referred solely to XYZ. So we must conclude that the reference of at least natural kind terms (and their corresponding concepts) is not determined solely by the speakers' mental states (again, narrowly construed).

Externalism might seem to present a problem for the traditional psychological study of concepts: if the mental states of speakers cannot (alone) determine the reference of some of their terms, then neither can concepts if they are conceived of as mental representations (see Braisby et al. 1996, p. 249). I take it that the problem is merely apparent, and that if externalism about natural kind terms is true, it is so because *we make it true*—it is we ourselves that pass over the responsibility of reference fixing to some external matters of fact. Consider the Twin Earth case again. How do we know that in 1750 our term ‘water’ did not in fact refer to XYZ? Answer: we simply intuit. This, however, is of little help, since the core question remains unanswered: *why* do we intuit in this way? After all, we could have had the intuition that (in 1750) XYZ did indeed belong in the extension of our term ‘water’ (some philosophers have defended such a position; see Segal 2000; Crane 1991). What causes our intuition that Twin Earth’s XYZ was or was not water in 1750 is our concept WATER. If the concept causes us to categorise the Twin Earth substance as not water, then externalism receives support; if not, then externalism does not receive support. Externalism stands or falls because of how our concepts are structured. Another matter, one that I will investigate in the present thesis, is what it is about natural kind concepts that causes some terms to refer externalistically.

1.3. The structure of the introductory essay

I will proceed as follows. First, in section (2), I will examine theories of concepts in psychology, and argue that each of them, as typically interpreted, is inadequate as a theory of reference. This is mainly because they all fall prey to the Kripkean problems of ignorance and error. An exception is Jerry Fodor’s account of concepts, where the reference of a concept is determined by specific kind of causal relations between properties and the speakers’ dispositions to token the concept. I will argue, however, that the account is implausible because it does not do any theoretical work in its own right, as it doesn’t explain what makes the reference-determining causal relations hold. Finally, in section (2.5) I will put forward a dual theory of concepts, which arguably solves the ignorance and error problems. On this account, natural kind concept cores are structured roughly as psychological essentialism suggests, with the exception that the essence beliefs involve a belief in an *external* essence determining category membership.

In section (3) I will turn to philosophical theories of reference, which fall mainly into two classes, description and causal theories. I argue that the description theory is undermined by the ignorance and error problems, and that it also conflicts with findings in experimental psychology of concepts. The causal theory, on the other hand, falls prey to the so-called *qua*- and composition problems. These problems manifest that causal relations alone are not sufficient to determine the reference of a term, and that some identificatory knowledge must also be utilised in reference fixing. I argue that refining both the description and causal theory of reference leads to an account practically identical to the dual theory I am defending.

I conclude by, first, discussing some of the major differences between the proposed dual theory and the traditional philosophical accounts of reference. Second, I will briefly investigate a dual theory's applicability to some other concepts than natural kind concepts. Third, I will examine how the dual theory meets the demands typically posited for a theory of concepts, arguing that it may have important advantages over the other theories in some areas. Finally, I will apply the dual theory to the philosophical issue about *rigid predicates*, arguing that the fact that natural kind concepts are essentialistically structured does all the theoretical work expected of rigidity.

1.4. Terms and concepts

As the reader may already have noticed, at times I talk about the reference or use of *terms*, and at others about the reference and use of *concepts.* I will take it that both terms and concepts can refer and can be used in categorisation. The difference between concepts and terms is that the former are mental representations whereas the latter are linguistic objects. A (lexical) concept is associated with its corresponding term in the mind of a speaker, or a translation of the term: for instance, an English speaker associates with the concept CAT the term 'cat', a Finnish speaker the term 'kissa', and a Swedish speaker the term 'katt'. The concept CAT can be shared between speakers of different languages, whereas a term typically is not shared. I will suppose that the reference of a term is determined by its corresponding concept, or, as in the case of natural kind terms, by the concept as a function of some external facts. For instance, the reference of the term 'cat' is determined by the concept CAT as a function of what kind of creatures are actually called 'cats'. The fine details of the term / concept distinction are not important for the topics of this thesis; for a discussion on the distinction, see e.g. Margolis and Laurence (2006).

2. Theories of Concepts

In this section, I examine the most prominent theories of concepts put forward in psychology and cognitive science. I do not intend to provide an extensive review of the theories here, but only focus on how they manage in explaining reference (for a more detailed review, see e.g. Murphy 2004). My discussion in this section partly draws on the valuable review by Laurence and Margolis (1999).

2.1. The classical view

According to the classical view, concepts are definitions, which give the necessary and sufficient conditions for both their semantic and non-semantic application. For instance, the concept BACHELOR might consist of represented features like *unmarried*, *adult*, and *male*; WATER of represented features like *liquid*, *odourless*, *colourless*, *thirst-quenching*, and so on. On this view, a concept C semantically applies to an object x if and only x possesses each of the features represented in C. Likewise, an object x is categorised as C if and only if it is reckoned to possess the features represented in C. On the classical view categorisation is a psychological process of 'checking' whether an object has the features that make it belong in the extension of C.

The classical view is undermined both on theoretical grounds as a theory of reference, and on experimental grounds as a theory of categorisation. The following two sections will be concerned with problems of the first kind, and the third with problems of the second kind.

2.1.1. Plato's and Wittgenstein's problems

Probably the earliest argument against the classical view is the so-called *Plato's problem*, which is that very few concepts in fact have definitions (at least as far as we know). This problem is especially pressing in the case of philosophical concepts. Despite over two millennia of hard reflection, thought experimentation, and argumentation, no one has yet come up with generally accepted definitions for concepts like KNOWLEDGE, JUSTICE, GOODNESS, TRUTH, or BEAUTY, just to name a few. Related to this difficulty is Wittgenstein's problem about family resemblances, which pertains also to ordinary (non-philosophical) concepts, such as GAME. Wittgenstein (1968) points out that some categories are not definable with the help of necessary and sufficient conditions, but their members may share only partially overlapping sets of features, or possibly even no features at all. For example, there is nothing that is common to all games, but various games share only some similarities with each others: some games involve winning and losing, others don't, some involve a ball, others don't, some involve only one person and others many.

These two problems are directed at the classical view's commitment about the *structure* of concepts, namely that they are definitions instead of, say, sets of family-related features. They also affect how plausible the theory is as a theory of reference: if (some) categories cannot be captured in definitions, then a definition cannot possibly determine what belongs in the concept's extension.

2.1.2. Ignorance and error problems

The ignorance and error problems were originally raised by Kripke (1980) against the *description theory of reference*, or *descriptivism*, in short. This account holds that the reference of a term is determined by a description associated with it (in the mind of the speaker). For instance, the term 'bachelor' refers to any object that satisfies the description *unmarried adult male*, or the term 'water' refers to whatever satisfies the description *the odourless, colourless, thirst-quenching substance that falls from the sky and fills lakes and seas* (and so on). Laurence and Margolis (1999) adapt these arguments against the classical view of concepts since, as they note, the classical theory is basically just 'descriptivism applied to concepts' (p. 21). Insofar as the ignorance and error arguments undermine the description theory of reference, they also undermine the classical view of concepts. Kripke's arguments were concerned with natural kind terms and proper names; in this section I will be concerned solely with natural kind terms.

Let us start with the error problems. To take a simple example, many people don't know anything about leprosy but only believe that the disease causes limbs and organs to fall off. This belief is, however, false. Descriptivism implies that in this case the term 'leprosy' does not in fact refer to the disease commonly called 'leprosy', as the disease doesn't actually cause limbs and organs to fall off. This, however, is implausible: if the speakers did not manage to refer to leprosy in the first place, how could they have been wrong about the nature of the disease? It seems that the speakers did manage to refer to leprosy with their term, but only had a false belief about its nature. Other examples of the error problem are easy to come up with: for instance, alchemists believed that gold is a compound that could be manufactured by combining other substances, Aristotle believed that water and fire are elements, the phlogiston theorists believed that fire is a process where a substance is released from matter. Intuitively speaking, despite their false beliefs, these speakers still managed to refer to gold, water, and fire, respectively. It seems clear that they were wrong about the nature of these kinds, but if they failed to refer to them, this couldn't have been so.

Ignorance problems are even more abundant than problems of error. The problem is that the speakers' identifying knowledge of the kind referred to is often insufficient to determine the term's extension. Consider the term 'water', for instance. At the time when the molecular structure of water was not known, the description associated with the term 'water' arguably consisted mainly of water's perceptual and functional features. Such a description fails to pick out uniquely the natural kind we are accustomed to refer to as 'water' (that is, H_2O);

instead, it picks out *any* kind that happens to satisfy the description, irrespectively of its deep structure (such as Putnam's (1975) XYZ). Or, take the leprosy example again. Before the discovery that leprosy is caused by the bacterium *Mycobacterium leprae*, the disease was arguably recognised in virtue of some symptomatic description. There can, however, be diseases which cause symptoms similar to those caused by leprosy, but which are nevertheless not caused by *Mycobacterium leprae*. Thus, the identifying description failed to uniquely pick out leprosy.

2.1.3. Problems in explaining categorisation

The above problems stem from, generally speaking, normative issues about the *correct* use of concepts, in contrast to contingent matters of how people actually happen to *use* concepts. Plato's and Wittgenstein's problems concerned whether categories can be captured in definitions; that is, whether a definition can delineate which all objects truly belong in a category. Kripke's problems, in turn, present a challenge to whether descriptions or definitions can determine reference: it is argued that a speaker can manage to refer despite having erroneous or insufficient identificatory knowledge about the category members. Yet concepts on the classical view are supposed not only to determine reference or true category boundaries, but also be causally responsible for people's categorisation judgements. Whether the account manages in this task is an empirical matter. Let us next briefly examine the experimental evidence against the definition theory.

The weightiest evidence against the classical view in categorisation comes from the so-called *typicality effects* in categorisation (see e.g. Rosch 1999). The classical view entails that every member of a category has an equal footing in the category, but several studies have shown that subjects judge some instances of a category to be more or less typical than others—for instance, apples and plums are judged to be more typical fruits than figs and olives, cars more typical vehicles than tricycles, murder more typical a crime than blackmail, and so on (Rosch 1973, table 3). Rosch and Mervis (1975) showed that the judged typicality of a category member corresponds to its *family resemblance* with other members of the category: the more features a category member shares with other members of the category, the more typical it is judged, and vice versa. The typicality of a category member reliably predicts a range of other psychological phenomena. Most importantly, typicality judgements correlate with subjects' speed in confirming category statements: the higher the judged typicality of an object (or subcategory) x in a category C, the faster a sentence of the type 'x is a C' is confirmed. For instance, a banana is confirmed as a fruit more quickly than a strawberry, a car is confirmed as a vehicle more quickly than a tank, and so on. (Rosch 1973, pp. 136 – 137.) Typical members or subcategories of a category are also more easily learned and recalled, they develop earlier in infancy, and the typicality of a category member determines how sensitive it is to priming (that is, whether advance information about the category facilitates or inhibits categorisation judgements). (See Rosch 1999, pp. 198 – 199.) Even though these findings don't straightforwardly contradict the classical theory, they undermine it because the account has no theoretical equipment to deal with them. On the other hand, these same data support the prototype view of concepts, which will be the topic of the next section.

On the basis of the accumulating experimental data against it, the classical view has been quite unanimously abandoned as a theory categorisation (for an overview of the progress, see Smith and Medin 1981). This data does not, however, straightforwardly undermine it as a theory of reference, or as a theory of the semantic cores of concepts. The definition view can be modified to solve many of the ignorance and error problems, and such a view still has some popularity in philosophy of language (see e.g. Jackson 1998, Lewis 1984). I will examine this view in the section devoted to philosophical theories of reference, now let's take a look at the prototype and exemplar theories which superseded the classical theory.

2.2. Prototype and exemplar theories

We already examined briefly the typicality effects that undermine the definition theory. These same data support the *prototype theory*, originally formulated in the writings of Eleanor Rosch and her colleagues (e.g. Rosch 1999).

The prototype theory equates concepts with (or holds that concepts have) prototypes. Prototypes are sets of represented features just like on the definition theory, but this time the features don't form singly necessary and jointly sufficient conditions for category membership. Instead, an object may belong in a concept's extension even if it doesn't possess each and every one of the represented features. Unlike on the definition theory, the features in a prototype are not equally important for category membership, but each feature is assigned a *typicality value*. The typicality value of a feature in a prototype is the higher the more salient or frequently occurring the feature is among the category members. For instance, most birds fly, sing, are small, and lay eggs, so these features receive relatively high typicality values in the prototype BIRD. According to the prototype theory, an object is categorised in the extension of a concept if it is sufficiently similar to the prototype. Typically a prototype is assigned a *threshold value* which specifies how similar an object needs to be with the prototype in order to be categorised in the concept's extension; on such accounts membership in a prototype category is (aside from borderline cases) all-or-none. However, a prototype can also be taken to lack a threshold value, in which case membership in a prototype category is graded: an object's membership in a category is determined by its similarity to the category prototype (see Osherson and Smith 1981).

To illustrate, let's consider an (simplified) example. Suppose that the prototype CAT consists of the features *has cat shape*, which has the typicality value .9, *meows* (typicality .8), *has whiskers* (.7) and *preys mice* (.6). Further, suppose that the similarity between an object x and a prototype P is calculated according to Amos Tversky's (1977) 'contrast principle' (here I and J are sets of x's and P's features, respectively):

$$\text{(CP) Sim (I,J)} = \mathrm{a}f(\mathrm{I} \cap \mathrm{J}) - \mathrm{b}f(\mathrm{I} - \mathrm{J}) - \mathrm{c}f(\mathrm{J} - \mathrm{I})$$

The constants a, b, and c allow different weights to be assigned to the shared features (I ∩ J) and the two sets of distinctive features (I – J) and (J – I); for simplicity's sake I suppose that each constant is 1. The function *f* assigns each features in the sets I and J a typicality value: it assigns .9 to the feature *has cat shape*, the value .8 to the feature *meows*, and so on. Lastly, we may assign the similarity measure some threshold value which any object needs to exceed in order to trigger the prototype; let's suppose that the measure for the prototype CAT is 1.5. Now, suppose that an object x has cat shape, meows, and has whiskers but doesn't prey mice. The similarity between x and CAT is, then (.9 + .8 + .7) − .6 − 0 = 1.8. Since the similarity between x and CAT exceeds the threshold value 1.5 associated with CAT, the object x is categorised as a cat. An object with another set of features can also trigger the prototype: suppose that an object has all the other prototypical features of cats but lacks whiskers. This time the similarity measure is (.9 + .8 + .6) − .7 − 0 = 1.6, so the object is likewise categorised as a cat. (On the other hand, if an object either lacks cat shape or doesn't meow, it is not categorised as a cat.)

The prototype theory differs from the definition theory in two important respects: First, as the above example illustrates, an object need not possess all the features represented in a prototype to be categorised in the concept's extension, but only sufficiently many of them. This makes it impossible to capture prototype categories in conjunctive definitions or descriptions, where each feature is necessary for category membership. Moreover, even if we allow the definitions or descriptions to be disjunctive, they easily become so complex as to make them trivial, for the disjunction must contain all of the different sets of features that can trigger the prototype (consider Wittgenstein's example about the category of games). Second, unlike on the definition theory, on the prototype account objects that fall under a concept's extension do not have an equal footing in the category. A category member can possess different numbers of the features represented in the prototype, and the possessed features can have different typicality values, causing that the category members can be more or less typical of the category. For instance, robins and sparrows are more typical birds than penguins and chickens, since the former possess more of the typical bird features than the latter. Thus, the prototype theory receives support from the typicality effects in categorisation, which undermine the definition view. (For reviews see Smith & Medin 1981; Murphy 2004.)

The exemplar view

Whereas on the prototype view a concept consists of features abstracted from category members, on the exemplar view the concept consists of representations of the typical exemplars themselves. The exemplars may be either some specific instances or sub-categories of the concept. For instance, a subject's concept CAT might consist of representations some particular cats, say, Oliver and Ronja, and possibly also representations of some subcategories of cats, such as the Siamese, feral cat, and so on. On this view, an object is categorised as a cat if it is sufficiently similar to (some of) the represented exemplars. As on the prototype theory, an object need not possess each of the features of the exemplars to be categorised, but only some criterial number of them. (See e.g. Smith and Medin 1999; Murphy 2004, chapter 4.)

Gregory Murphy (2004) distinguishes between two kinds of support for the exemplar theory; one from *exemplar effects* and the other from *exemplar models*. The former is relatively simple to understand: for instance, suppose in all my life I have seen an espresso machine only once and don't have an idea what espresso machines typically look like. Upon meeting a weird looking machine in a department store, I might try to figure out whether it is an espresso machine by comparing it to my memory of the particular espresso machine I have met. Similar examples are easy to come up with. Support from exemplar models involves the specific theoretical layout of the exemplar models, which has proven to be more successful than prototype models in explaining some experimental findings.

It is still open whether the prototype or exemplar theory will prove correct, or whether they can somehow be made compatible. The differences between these two theories are quite subtle, and irrelevant from the present perspective, as they agree on most major issues: both hold that an object is categorised in a concept's extension if it is sufficiently similar to the prototype or set of exemplars.

2.2.1. Ignorance and error problems

The prototype and exemplar theories have proved to have great explanatory force in explaining relatively simple and fast categorisation, but they face difficulties in accounting for more considered judgements, and especially reference. (It is notable that some theorists don't take prototypes or exemplars to determine reference at all; see Smith, Medin, and Rips 1984.) From our perspective the main problem for the similarity based views is that objects which do not trigger the prototype, or which aren't sufficiently similar to the represented exemplars, may nevertheless belong in the category. And vice versa, objects that do trigger the prototype or do bear sufficient similarity to the exemplars may nevertheless *not* belong in the category. For instance, gold in gaseous form does not possess sufficiently many of the prototypical features of gold to be categorised as gold, and neither does it resemble the exemplars of gold we have met. However, it still *is* gold, and is arguably even categorised as gold by anyone who knows it possesses the actual deep structure of gold. On the other hand, an object can trigger the prototype or resemble the exemplars of a category but nevertheless not really belong in it. For example, many corals and anemones arguably do trigger the prototype PLANT and resemble the represented plant-exemplars, but are nevertheless animals. Moreover, they are even categorised as animals by any one who knows about, say, their functioning (e.g., that they eat instead of assimilate) or deep structure (e.g., that they consist of animal cells instead of plant cells). Or, to take an example from the psychological literature, a cat cleverly disguised to look like a skunk may trigger the prototype SKUNK and resemble the represented skunk exemplars, but nevertheless fail to be a real skunk. Moreover, it is not categorised as a skunk by anyone who is aware of the make-up process (cf. Keil 1989). Similar examples are easy to come up with.

Cases like these undermine the prototype theory both as a theory of considered categorisation and reference. At least in the case of natural kind concepts, people's categorisation judgements are not driven solely by prototypical features, or by superficial resemblance to some exemplars, but some unperceivable deep features can override them. The prototype and exemplar theories are even less plausible as theories of reference: triggering a prototype or resembling a set of exemplars is neither necessary nor sufficient for category membership: gaseous gold is gold though it does not trigger the prototype of gold or resemble the represented gold exemplars, and fool's gold is not gold though it does trigger the prototype of gold and resemble the represented gold exemplars.

Some prototype theorists have tried to accommodate for these problems by allowing deep features to be included in prototypes (e.g. Hampton 1998, p. 138). For instance, the prototype or exemplar set of flowers could include a representation that some vital fluids flow inside them, that they turn towards the sun, that they die if not given sufficient sunlight, water, and nutrients, and so on. This makes it possible for a subject to distinguish a silk flower from a real one, but does not preclude more profound possibilities of ignorance and error: for instance, the apparent flower may not eventually belong in the division *Magnoliophyta* like the plants we ordinarily call 'flowers', but might have a different lineage and genome. As Kripke and Putnam have showed, almost *any* specific beliefs about a natural category may turn out to be false.

The ignorance and error problems show that concepts cannot be *merely* prototypes or sets of exemplars. However, these problems by no means undermine the theories as accounts of quick, rough-and-ready categorisation. The theories receive strong support in explaining categorisation of this kind, and it doesn't matter at all if such categorisation judgements can be false—it suffices to confirm the theories that, in those specific kind of categorisation tasks the theories model, people *as a matter of fact* do use concepts in the way predicted. Evidence for the prototype or exemplar views can even be found in some non-human animals (see e.g. Aydin & Pearce 1994; Werner & Rehkämper 2001), which suggests that we are dealing with a profound, hard-wired cognitive process. There is no reason to reject the theories completely because of the ignorance and error problems.

The most plausible response of the prototype and exemplar theorists to the ignorance and error problems is that they are not investigating considered categorisation judgements or reference at all, but rather only a relatively constrained class of categorisation judgements: quick categorisation based on readily perceivable features. How categorisation is made based on some deeper features is a different matter, which the prototype and exemplar theorists need not be concerned with; there can be several, qualitatively different categorisation processes. A strategy like this is endorsed by dual theorists, who hold that concepts have two separate components, an *identification procedure* and a semantically constitutive *core*: the former drives quick and unconsidered categorisation judgements and doesn't determine reference, whereas the latter determines more considered categorisation judgements and is semantically constitutive. Laurence and Margolis 'suspect that a model of this sort has widespread support in psychology' (1999, p. 46).

I will defend a variety of the dual theory in section (2.5). Before that, let us take a look at what's possibly the most important argument against the prototype theory, namely the claim that the theory cannot account for compositionality. Though this argument pertains mostly to the prototype theory, similar considerations can also be raised against the exemplar view.

2.2.2. The composition problem

To begin with it should be noted that if we endorse a dual theory where prototypes aren't semantically constitutive, the composition problem need not arise at all—it suffices for conceptual combination that the semantic cores compose, not the identification procedures. We cannot, however, ignore the composition problem solely by appeal to the dual theory, as even though *some* concepts arguably do have separate cores and identification procedures, not all concepts need to.

Accounting for compositionality is a crucially important task of a theory of concepts. We can form a practically unlimited number of complex concepts out of our limited set of primitive concepts—we can form novel concepts of, say, golden chair, wooden bicycle, miniature horse, monster apple, and so on. Complex concepts are formed out of more primitive ones: say, the concept MONSTER APPLE consists of the concepts MONSTER and APPLE, GOLDEN CHAIR of the concepts GOLDEN and CHAIR. A theory of compositionality for concepts is expected to explain two distinct matters. First, we expect the theory to explain how concepts compose *on the semantic level*, or how the extension of a complex concept is determined as a function of its constituent concepts. Second, we expect the theory of compositionality to explain how concepts are composed *in thought*, or how subjects form complex mental representations out of simpler ones.

The *classical theory of compositionality* explains compositionality in two stages. The first stage is purely semantic. Andrew Connolly, Jerry Fodor, and Lila and Henry Gleitman write that, in the first stage of conceptual combination,

> all you get from your concepts and combinatorics is output denoting relations among sets, properties, or individuals (depending on the ontology assumed). Thus RED HAIR designates the set of instances of hair whose colours are instances of red. (2007, p. 4.)

For instance, if we endorse a set-theoretic approach to conceptual combination, we may stipulate that the concept RED HAIR designates the intersection of the extensions of the concepts RED and HAIR. This explains how the semantic value of the complex expression RED HAIR is determined, but does nothing to account for how we form the representation of red hair. A specific problem here is to explain why some instances of red hair are more typical than others. For instance, in forming a thought of red hair, some hues come into mind more easily than others—say, orange red hair is more typical than fire engine red hair. The classical theorists account for this by relying on another stage in conceptual processing, which involves 'the application of a further set of pragmatic-inferential processes that draw on general knowledge of the world' (Connolly et al. 2007, p. 4). For instance, although red hair can be of any hue of red, I know from everyday experience that it typically is of just some specific hues. We will come back to the classical view after examining the possible problems that the prototype theory faces in accounting for compositionality.

The main problem in prototype compositionality stems from the influential critique by Daniel Osherson and Edward Smith (1981). Osherson and Smith take the prototype theory to entail that category membership can come in degrees, and because of this conceptual combination cannot be modelled in traditional theories based on classical logic. Instead, Osherson and Smith suggest that conceptual combination of prototypes be modelled on *fuzzy set theory* (e.g. Zadeh 1965). Osherson and Smith argue that this approach has some deeply counter-intuitive consequences. Take, for instance, the complex concept STRIPED APPLE. Osherson and Smith take this concept's extension to be determined by the fuzzy intersection of the extensions of the concepts STRIPED and APPLE. This intersection, in turn, is determined by the so-called *Min Rule*, according to which an object is a striped apple to the minimum of the degrees that it is striped and that it is an apple. This leads to a contradiction, as demonstrated by Osherson and Smith in the following way.

Take a typical striped apple x. x is arguably more typical as a striped apple than as an apple, since prototypical apples aren't striped. In other words, x has a higher degree of membership in the category of striped apples than in apples:

(1) $c_{\text{striped apple}}(x) > c_{\text{apple}}(x)$.

(Where *c* denotes degree of category membership.) But now, the Min Rule implies that the striped apple x is a striped apple at most to the extent that it is an apple:

(2) $c_{\text{striped apple}}(x) \leq c_{\text{apple}}(x)$.

We have ended in a contradiction: intuitively, the striped apple x is a good instance of a striped apple, but it is a bad instance of an apple. But according to the Min Rule, x can be a striped apple only to the extent it is an apple. (Osherson and Smith 1981, pp. 43 – 45.) Similar examples are not hard to come up with: a guppy is atypical as a fish and atypical as a pet, but nevertheless typical as a pet fish (example from Osherson and Smith 1981); a king's golden throne is atypical as a sample of gold and atypical as a chair, but nevertheless typical as a golden chair; and so on.

Response strategies to Osherson and Smith's problems can be divided into two classes: some have suggested further logical or semantic strategies to deal with the problems of vague concepts. For instance, Hans Kamp and Barbara Partee (1995) suggest that prototype concept combination be modelled on the basis of *supervaluation theory*, on which a vague expression can be 'precisified' so that it attains a definite extension (see Kamp and Partee 1995, p. 148 ff.). A wholly different approach to the problems is to try to provide not a semantic or logical, but a *psychological* model of prototype combination. The goal of such a theory is to account for how a complex concept inherits the typical features of its constituent prototypes—say, why PET FISH inherits from the concept PET features like cuteness and relatively small size but not features like having fur and waving tail. Connolly et al. (2007, p. 6) title the most prominent theories of this variety *default to the stereotype* accounts.

One especially promising variety of the default to the stereotype account is the *selective modification* model (Smith, Osherson, Rips, & Keane 1988). On this account, the feature representations in a prototype may take different values depending on which concept the prototype is combined with. For instance, the concept APPLE may contain feature dimensions such as colour, shape, and texture, on which the prototype can take various values, such as plain, striped, checked, or dotted on the texture dimension. A complex concept is formed through selectively modifying the value of a specific feature dimension of the prototype; say, the concept STRIPED APPLE is formed by changing the value of the texture dimension in the APPLE prototype from plain (or whatever is the typical texture of apples) to striped. This account has the advantage of accounting for why an object can be more typical as a striped apple than as either striped or an apple: it is typical a member of the concept APPLE where the texture dimension is selectively modified. Similarly, in forming the concept PET FISH, the concept PET affects feature dimensions of FISH such as *size* and *cuteness*, but leaves the *coat* dimension intact.

Though the default to the stereotype models may be on the right track, they face both theoretical and experimental problems (see Connolly et al. 2007). A specific problem pertaining to the selective modification model is that concepts probably don't have enough feature dimensions to enable exotic combinations. Consider, for instance, the concept MONSTER BANANA. It is doubtful that the concept BANANA contains a *monstrosity* dimension that can be modified by the concept MONSTER. A more general problem is that it may not be possible to model conceptual combination with the help of *any* specific psychological model at all; rather, the process may involve highly sophisticated and versatile processes such as creative imagination, and it may be affected by personal memories (see Wiesniewski and Gentner 1991). For instance, we need not suppose that there is some universal cognitive process that determines how one forms a representation of, say, *pre-school teacher*, but various people's representations of pre-school teachers may be affected by their personal memories of some specific pre-school teachers. These considerations point towards the classical theory of conceptual combination.

For some reason the classical theory of compositionality has not been adopted for prototype concepts, but I do not see why this approach shouldn't succeed. This account faces two kinds of problems. First there are problems pertaining to the first stage in conceptual combination. As we saw, Osherson and Smith (1981) argue that membership in prototype categories is graded, which makes modelling their semantics in classical logic and set theory impossible. In particular, both Osherson and Smith's fuzzy set theory and Kamp and Partee's supervaluation theory are problematic. Again, there are problems pertaining to the second, psychological stage in conceptual processing—in particular, we have reason to doubt that the composition of thought could be modelled in any specific, simple model such as the selective modification model. However, the prospects for a classical theory of prototype composition are not as dim as they might seem.

I take that the crucial stage in conceptual combination on the classical theory is the first one, and that the prototype theory has no special problems in accounting for it. The reason for this is twofold. First, the problems that beset Osherson and Smith, and Kamp and Partee stemmed from considering membership in prototype category graded, but the prototype theory needs not make such a commitment. A plausible variation of the prototype theory is one where prototypes have threshold values, and on which an object is categorised in the prototype's extension if its similarity to it exceeds the threshold value. On such an account, membership in prototype categories is all-or-none (borderline cases aside). Providing logic and semantics for such prototypes is, then, much less problematic than providing semantics for graded categories: at best we can do with classical logic (if we ignore the borderline cases), or at worst with three value logic (if we want to take borderline cases into account). Second, and more importantly, if some concepts are indeed vague in that membership in them is graded, they are so *irrespectively of the prototype theory*. Consider, for instance, concepts such as BALD, RED, or STEEP. We can take it that, *as a matter of fact*, some people are more or less bald than others, that some objects are redder than others, or that some slopes are steeper than others. Providing the semantics and logic for such vague concepts is a general problem that belongs mainly to philosophy of language, logic, and linguistics, not the prototype theory—some concepts simply are, irrespectively of the prototype theory, vague (see e.g. Keefe 2000). Fodor and Lepore (1996) notice this same point but try to turn it *against* the prototype theory, which is in my opinion illegitimate.

Of course, nothing said thus far helps in solving the second stage in conceptual combination on the prototype model. In specific, even if prototypes could compose according to classical or three value logic, we would still have to explain how speakers form representations of the complex prototypes, and in particular, what determines their typicality values. I dare claim that this stage is not, however, crucially important in conceptual combination. Typically the first, semantic stage suffices for possessing a complex concept, and often a speaker may not have a specific representation of the complex concept at all. The case of Boolean concepts supports this presumption: a speaker can possess a Boolean concept such as NON-CAT even though she cannot form a specific representation (that is, a prototype) of the category (see Fodor and Lepore 1996). I take it that this is quite compatible with the prototype theory. We may suppose that the concept CAT is nothing but a prototype, and that it refers to whatever triggers it. Similarly, the concept NON-CAT is a combination of the logical modifier NON and the prototype CAT, and it refers to whatever does *not* trigger the prototype CAT. Similarly for any other Boolean concepts: a complex concept (C1 & C2), where C1 and C2 are prototypes, refers to whatever triggers both the prototype C1 and C2; the concept (C1 ∨ C2) refers to whatever triggers C1 or C2; and so on.

This approach raises, however, a problem: it seems that on this account, not all concepts are prototypes, after all. For instance, the concept NON-CAT quite clearly is not a prototype, as it lacks a prototype altogether (Fodor and Lepore (1996) call this the *missing prototypes* problem and consider it as counterevidence to the prototype theory). I suppose that it is theoretically possible that a speaker could possess a complex concept like MONSTER BANANA simply in virtue of possessing its constituent prototypes MONSTER and BANANA (and grasping some basic logic), without forming a representation of it at all: the complex concept would refer to whatever triggers both the prototypes MONSTER and BANANA. So, it might well be that the prototype theorist has to grant that (at least some) complex concepts, even if they are formed out of prototypes, simply are not prototypes themselves. I do not consider this a major problem for the prototype theory: prototypes *can* compose, even though what results from the combination need not be prototypes themselves, but logical compounds of prototypes.

More extensive discussion of the composition problem lies beyond the scope of this introduction. However, we may conclude that the prototype theory is by no means undermined by the composition problem, but it has viable strategies of dealing with it. I have argued that one such strategy is to endorse a classical view of conceptual combination. However, even if the prototype theory could account for composition, it still is viable only as a theory of some restricted class of concepts. In particular, it is doomed to failure at least in the case of natural kind concepts, since in their case it falls prey to the ignorance and error problems. Let's examine next whether the latest trend in the psychology of concepts succeeds any better with these problems.

2.3. The theory-theory of concepts

The notion of a *theory-theory* of concepts can be considered as an umbrella term that encompasses a range of theories. Common to these accounts is that they are committed to the claim that (at least some) concepts are structured around, and embedded in, naive mental theories (e.g., Carey 1985, p. 198). We may distinguish between two research trends in the theory-theory: some authors have focused on issues about conceptual development (see especially Carey 1985; 1999), whereas others have focused on issues about categorisation. I will here focus solely on the latter issue, with respect to which the theory-theory endorses *psychological essentialism.*

2.3.1. Psychological essentialism

Psychological essentialism draws from Kripke's and Putnam's considerations in the philosophy of language (see Keil 1989). It is motivated by the fact that people's categorisation judgements are not always driven solely by an object's appearance, but are sometimes affected by the unobservable deep structure of the sample. Psychological essentialism is concerned specifically with natural kind concepts, which I will focus on in this section; other varieties of essentialism will be briefly examined in section (4.2).

According to psychological essentialism, people believe that natural kind members share some hidden, unobservable deep structure, the possession of which they take to be necessary and sufficient for belonging in the category. For instance, it is believed that all cats share some underlying essence, say, genome, lineage, or more generally 'insides', which determines what belongs in the category of cats. The essence is considered necessary for a cat: even if a creature looked and behaved just like a cat, it might nevertheless not be one if it did not possess the cat essence. The essence is also considered sufficient: even if a creature doesn't appear to be a cat or behave like one, it might still be a cat if it possesses the cat essence.

Psychological essentialism typically takes the believed essence to be *causal* in nature: it is some empirically discoverable property or set of properties, which cause(s) the category members' other typical properties. According to Susan Gelman (2003), essentialism posits people with a belief in some 'substance, power, quality, process, relationship, or entity that causes other category-typical properties to emerge and be sustained, and that confers identity' (Gelman 2003, p. 405). The most prominent variety of psychological essentialism is *placeholder essentialism* (e.g. Medin & Ortony 1989; Rips 2001; Gelman 2003, 2004), which holds that speakers need not have any specific beliefs about a category's essence, but only believe that the category members share *some* essence. For instance, a subject may believe that something about a substance's chemical composition makes it, say, water, although she doesn't know what that deep structure specifically is; or she may believe that something about cats' genetic makeup or lineage makes them cats, although she doesn't know what that genome or lineage specifically is. This empty placeholder belief about essence can be replaced with a specific *essence hypothesis*; for instance, the subject may come to believe that water is essentially H_2O, or that the cat genome is some specific genome G.

Psychological essentialism in categorisation receives support mainly from *discovery* and *transformation studies*. For instance, Frank Keil (1989) presented children with discovery scenarios, in one of which an animal that appeared to be a horse turned out to have the 'blood and bones of cows', and came from cow parents and had cow offspring (p. 162). Keil discovered that children tended to categorise the animal as a horse rather than a cow, that is, they favoured the animal's insides and lineage over its appearance in categorising it. In the transformation studies an animal's appearance was transformed, but its insides were kept constant. For instance, Keil described to children a raccoon that had been cleverly disguised to appear like a skunk; it had been painted with stripes and its body had been added a ‘sac of super smelly odour’ (p. 184). In another variation of the transformation study deeper manipulation was used; for instance, a young horse was given an injection that made it grow up to appear like a zebra (p. 222). In both cases children tended to categorise according to the deep structure of the probe, instead of relying on its appearance. It is noteworthy that, in contrast to natural kinds, when the appearance and function of an artefact was changed, the children did in fact take the object’s category to have been changed. For instance, when a coffeepot was made to look like and function as a birdfeeder, the subjects did in fact consider its category to have been changed (p. 184; p. 186 ff.). (For a fairly recent review of experiments supporting essentialism, see Gelman 2004.)

In the previous section we saw that the prototype and exemplar theories were viable in accounting for quick rough-and-ready categorisation judgements, but that they couldn’t determine the reference of (at least) natural kind concepts due to the ignorance and error problems. They can, however, serve as non-semantic identification procedures of concepts on a dual theory where some other mechanism, the concept core, determines reference. Could the essence beliefs constitute natural kind concept cores? Unfortunately, the present formulation of psychological essentialism is just as prone to the ignorance and error problems as any other of the discussed theories.

2.3.2. Ignorance and error problems

Laurence and Margolis raise ignorance and error problems against psychological essentialism, and argue that they undermine it as a theory of reference. These problems are made more difficult by the fact that the believed essences are supposed to be necessary and sufficient for category membership, unlike in the case of the prototype and exemplar theory where an object need not possess all of the represented features to belong in a category. People often have erroneous and sketchy beliefs about essences, but usually they nevertheless manage to refer. To take some historical examples, Aristotle thought that earth, water, wind, and fire are elements, alchemists believed gold to be a compound, phlogiston theorists believed fire to be a process where a substance is released, and so on. If these features were necessary for category membership, the concepts couldn’t have referred to anything. This, however, is implausible, as it seems that the concepts did refer, but that the speakers only had false beliefs about the essences of their referents. More modern examples of cases of error can be found as well. For instance, skunks were formerly thought to be a subfamily of the *Mustelidae* family, but recent genetic evidence suggests that they probably aren't (Dragoo & Honeycutt 1997). If lineage is considered as an essential property of animal species, psychological essentialism implies that the term 'skunk' didn't refer to anything. Similar examples are not hard to find. Some people still believe that whales are fish, that air consists mainly of oxygen, that tomatoes, nuts, and grain aren't fruits, and so on and so forth. In each of these cases it seems that, despite their false beliefs, the speakers managed to refer—for if they didn't, they couldn't have been wrong about the nature of skunks, whales, air, and so on in the first place.

The psychological essentialists acknowledge these problems themselves and hold that the essence beliefs cannot determine reference. Gelman holds that specific essences are rarely known by lay speakers, and because of this cannot determine reference (2003, pp. 9 – 10). Douglas Medin and Andrew Ortony (1989) also notice that only in a few cases do the specific essence beliefs provide the necessary and sufficient conditions for category membership, because they rarely consist of a precise set of identificatory features (pp. 184 – 185). These observations are most probably true if we presume reference to be determined internalistically by the speakers' identificatory knowledge alone. In section (2.5) I will propose an alternative, externalistic formulation of psychological essentialism, on which the essence beliefs can in fact determine word extensions, and can constitute natural kind concept cores.

Laurence and Margolis reject the essence beliefs as constituting concept cores because of the ignorance and error problems, but endorse Fodor's conceptual atomism with respect the cores. This is the view I will turn to next.

2.4. Concept atomism and informational semantics

Jerry Fodor's theory of concepts is radically different from the traditional approaches discussed above. It holds that concepts are atomistic mental representations which are identified solely in virtue of their syntactical properties, or 'mental orthography' (e.g. Fodor 1998, p. 38). Any specific beliefs or psychological capacities (such as essence beliefs or prototype driven object recognition systems) are only *non-semantically* associated with them, and don't constitute them. On Fodor's view concepts do nothing but carry information about the properties they refer to.

How, then, do the concepts refer, if they don't encode any identificatory knowledge about the category members? Fodor's answer is informational semantics: a concept refers to the property (or properties) which tends to cause it to be tokened. For instance, cats tend to cause us to token CAT, and chairs tend to cause us to token CHAIR. This is the most basic informational theoretic assumption:

> IS1. It is a law that property *p* is disposed to cause the concept C that expresses it to be tokened.

The problem for this simple formulation is obvious: some properties other than the one the concept refers to can cause its being tokened; for instance, cows perceived on a dark night from a distance can appear to be horses, and cause the concept HORSE to be tokened. This problem is usually known as the *disjunction problem*: why does a concept refer to a specific property instead of the disjunction of properties that tend to cause it to be tokened? To solve this problem, Fodor introduces the following second claim:

> IS2. If it is a law that any property *q* other than *p* is disposed to cause C to be tokened, then this lawful relation is asymmetrically dependent on the lawful relation between *p* and C.

Call this the *asymmetry constraint*. It basically states that some of the property-concept tokening –relations (or p/C-relations for short) are more primitive than others, in that the less primitive are dependent on the more primitive, but not vice versa. For instance, in the above example, the cow's being disposed to cause the tokening of HORSE is dependent on the horse/HORSE-relation, because the cow causes the tokening of HORSE *only because it is mistaken for a horse*: if the speaker perceived the animal in daylight and from a shorter distance, she would not call it 'horse'. On the other hand, the horse/HORSE relation is primitive in that it is not dependent on any other p/C-relations: horses trigger the concept HORSE simply because they are horses, not because they are mistaken for something else. (See Fodor 1990, p. 90 ff; Fodor 1998, p. 120 ff. On specific formulations of informational semantics, see also Laurence & Margolis 1999, p. 61; Margolis 1998, p. 350 – 351.)

Fodor's theory arguably survives the problems of ignorance and error, naturally enough, since it holds reference to be causally determined. For instance, the term 'gold' does in fact refer to gold in gaseous form, since the gas has the property to which the concept GOLD is asymmetrically connected to, namely the property of being Au. It is Au that actually tends to cause us to token GOLD, and since gold in gaseous form possesses that property, GOLD refers to it. Or, take XYZ, which is disposed to cause WATER to be tokened although it doesn't really belong in the concept's extension. Fodor solves this problem by claiming that the XYZ/WATER-relation is in fact asymmetrically dependent on the H_2O/WATER-relation: if the speaker knew that the XYZ-sample is of a different kind than the actual water samples, she would, according to Fodor, not call it ‘water’—XYZ causes WATER to be tokened solely because it is mistaken for water (H_2O). Finally, take an example of an error problem: people used to believe that whales are (essentially) fish, but despite this false belief, they arguably still managed to refer to whales. Fodor would solve this problem as follows: despite the false essence belief, it was still all along the property of being a whale which, as a matter of fact, caused us to token WHALE. It doesn’t matter for the concept’s reference that we believed the animals to be fish, only our actual concept tokening dispositions are relevant.

Fodor’s theory of concepts cannot be a complete theory if concepts are thought to do more than just refer. In particular, concepts on Fodor’s account cannot possibly determine categorisation judgements, since they don’t involve any specific identificatory knowledge. Laurence and Margolis suggest that concepts could still have atomistic cores, and that reference could be determined causally as Fodor suggests. At the same time, prototypes, exemplars, and essence beliefs would have a role to play in explaining (among other things) people’s categorisation judgements, but on this view they wouldn’t semantically constitute concepts.

Laurence and Margolis’ suggestion is initially plausible. Prototypes, exemplar representations, and essence beliefs clearly determine people’s categorisation judgements, but they seem to be incapable of determining reference due to the ignorance and error problems. Causal reference fixing solves this problem, but, on the other hand, atomistic concepts cannot determine categorisation. A natural solution to this stand-off is to endorse a dual theory of concepts where the concept cores are atomistic and refer causally, and the identification procedures determine categorisation. Unfortunately for this variety of a dual theory, Fodor’s account of concepts and reference is arguably mistaken.

2.4.1. The problem with Fodor’s theory

How do properties cause concepts to be tokened on Fodor’s account? Properties cannot cause concepts to be tokened directly, especially if concepts are taken to be atomistic and to not contain any identificatory knowledge or specific psychological capacities. Fodor suggests that the p/C-relations are mediated by some *sustaining mechanisms* (psychological or other capacities), but holds that they are semantically irrelevant:

> [...] *all* that matters for meaning is ‘functional’ relations between symbols and their denotations. In particular, it doesn’t matter *how that covariation is mediated*; it doesn’t

> matter what mechanisms (neurological, intentional, spiritual, psychological, or whatever) sustain the covariation. (1990, p. 56, emphasis original.)

To illustrate, we can recognise cows by eyesight, hearing, through indirect cues such as cow-piles, or even utilising an electronic cow-radar. Since any of these mechanisms (it seems) sustains the same p/C-relations, they all determine reference to the same property, namely *cow*. In other words, the p/C-relations are multiply sustainable, and what in the end matters for reference are the purely functional p/C-relations alone.

Which psychological mechanisms do actually sustain the p/C-relations? Quite clearly, they are the mechanisms driving our categorisation judgements: to be disposed to token a concept C in the vicinity of a property *p* is simply to be disposed to *categorise* (an object instantiating) *p* under the extension of C. Experimental psychology shows that categorisation (or concept tokening behaviour) is probably caused by prototypes, exemplars, or essence beliefs. But if it is these mechanisms that determine the reference determining p/C-relations, how can they be irrelevant for a concept's reference? Fodor would reply, as the above quote shows, that even if prototypes, exemplar sets, or essence beliefs do actually sustain the p/C-relations, *the very same p/C-relations could be sustained by some other mechanisms*. But could they? This is a crucially important question for Fodor's account, since if the actual psychological mechanisms were necessary for the p/C-relations, they'd also be necessary for reference, and conceptual atomism would fail: in order to refer with a concept, one would have to possess some specific psychological capacity.

The key to noticing the problem of Fodor's account is that the sustaining mechanisms, as they are actually modelled in the psychology of concepts, are just functional models of our categorisation dispositions. In Article III I investigate various psychological mechanisms that might actually sustain the p/C-relations, and argue that each of them is necessary for the p/C-relations they sustain to hold (with one exception). Informational atomism thus fails, since in order to refer with a concept it is necessary to possess some specific (functional) psychological capacity. Fodor's claim that the sustaining mechanisms are irrelevant for a concept's reference would succeed if by sustaining mechanisms we meant some very specific neural or biological mechanisms; it wouldn't matter whether the p/C-relations of a concept were sustained by, say, a neural mechanism M or a functionally equivalent electric mechanism M'. But the psychology of concepts does not actually investigate any physiological mechanisms. The psychologists examine human categorisation behaviour —which objects they tend to categorise in the extension of a concept in some specific circumstances—and build functional theories that model this behaviour. So, the psychologists model just people's dispositions to token a concept, which Fodor claims to determine the concept's reference, and their work is, thus, anything but irrelevant semantically.

2.4.2. Ignorance and error problems revisited

So, Fodor's claim about the semantic irrelevance of the sustaining mechanisms is probably false, and the mechanisms are needed for the p/C-relations to hold. But doesn't his causal theoretic account still solve the pressing ignorance and error problems, which undermined each one of the theories we examined earlier? It will turn out that the causal theoretic account shouldn't receive this honour, since all the explanatory work in solving the ignorance and error problems is done by the sustaining mechanisms.

Take, for instance, the case of gaseous gold. Fodor's account implies that the concept GOLD refers to gold in gaseous form, since it possesses the property Au which normally causes the concept to be tokened. But why should we take it that the concept is ordinarily caused to be tokened specifically by the property *Au*? The objects that cause GOLD to be tokened in us can be considered to do so in virtue of *two* properties: the property of being Au and *the property of being goldish* (that is, being yellow, malleable, shiny, and so on). At least scientifically ignorant speakers are disposed to token the concept GOLD in the vicinity of *any* substance they reckon as gold, not just Au, so why should the concept's reference be fixed specifically to Au? Fodor's response is that the non-Au goldish substance/GOLD –relation is asymmetrically dependent on the Au/GOLD-relation: *if* the speaker knew that some instance of a non-Au goldish substance is different in deep structure from the bulk of the samples actually reckoned as gold (that is, Au), she would not call it 'gold'. But this only brings us to a second, more profound problem: *why should the speaker care about the samples' deep structures at all*? What reasons does she have not to consider some sample with a deviant deep structure as gold? After all, many concepts are not sensitive to deep structures: say, even if it turned out that all chairs happen to share some molecular structure S, we still wouldn't consider possessing S necessary for being a chair; we could easily manufacture a chair of some substance that doesn't possess S. To take another example, consider Putnam's XYZ. Fodor maintains that WATER doesn't refer to XYZ, because the XYZ/WATER-relation is asymmetrically dependent on the H_2O/WATER-relation; this, in turn, is because *if* a speaker knew that a sample of a watery substance does not share deep structure with the actual watery samples, she wouldn't reckon it water. But again: why should the speaker care about the sample's deep structure at all? Why doesn't she simply conclude that there happens to be two kinds of water, XYZ and H_2O?

The obvious solution to these problems is that the speakers *intend* (explicitly or implicitly) to use natural kind concepts to refer to a kind with a certain, hidden deep structure. Even Fodor acknowledges that speakers cannot possibly refer to natural kinds without some essence beliefs:

> I'm quite prepared to believe that, de facto, until we had [...] the concepts that cluster around NATURAL KIND, there was probably no way that we *could* link to water except [...] via water's metaphysically accidental but nomologically necessary properties [that is, water's appearance]. (1998, p. 158; emphasis original.)

In order to refer to H_2O instead of any watery substance, or Au instead of any goldish substance, we *must*, even according to Fodor, possess essence beliefs. So, it's the essence beliefs which cause us to stand in some specific kind of p/C-relations that do most of the work in fixing reference, not the purely causal p/C-relations themselves. It is *because of* the essence beliefs that we wouldn't apply a concept C to a sample if we knew that it doesn't share deep structure with the bulk of the actual C-samples.

Fodor's theory of concepts fails, as reference cannot be determined purely causally, but rather the reference-determining causal relations must themselves be determined by some specific psychological capacities. For this reason, Laurence and Margolis' suggestion about the dual theory where concepts have atomistic cores cannot do. However, another variation of a dual theory might well be viable—the idea that essence beliefs could form concept cores is not dead and buried yet.

2.5. Towards a dual theory of concepts

The essentialists claimed that essence beliefs cannot determine reference, since essences are rarely known by lay speakers, and beliefs about them are often imprecise and erroneous. This is indeed correct: people's specific essence beliefs cannot possibly specify category boundaries. But the placeholder belief that the category members share *some* hidden essence *can* determine reference, though not in the traditional, internalistic way, but externalistically. My suggestion is that people believe category boundaries to be determined by the hidden, possibly unknown, deep structure or essence shared by the actual category members, *even if no one knows what the deep structure in fact is*. This view, call it *externalistic essentialism*, can be summarised as follows:

(EE) 1. Speakers believe that samples falling under a natural kind concept C share some hidden, empirically discoverable, essence E.
2. The speakers take possessing E to be a necessary and sufficient condition for belonging in the extension of C.

Claim (1) simply reiterates psychological essentialism's claim that people believe in a category essence. The crucial claim in (EE) is (2), which posits the account with semantic force: it holds that the unknown category essence can determine the concept's extension even if no one knows what it is.

Externalistic essentialism is motivated both theoretically and experimentally. Most importantly, the suggestion solves the pressing ignorance and error problems by being externalistic: for instance, gaseous gold belongs in the extension of GOLD in virtue of being Au like all the other samples typically categorised as gold; XYZ is not water since its chemical constitution is different from actual water. Even though a lay speaker would arguably categorise gaseous gold as not gold and XYZ as water, she would be making a mistake, since the gas shares deep structure with the actual gold samples and XYZ is different in constitution from the actual water samples. Moreover, if the speaker knew about these facts, she would be able to arrive at a correct categorisation judgement. However, (EE) cannot be established purely on theoretical grounds, as it is committed to the empirical claim (2) that people in fact take natural category membership to be determined by the category members' deep structure, even if it is unknown. In Article II we argue that externalistic essentialism has experimentally testable consequences which distinguish it from internalistic theories of concepts. We present two experiments, which support the externalistic essentialist view.

It is quite clear that the externalistic essence beliefs cannot *alone* determine the reference of natural kind concepts. The main reason for this is that the essence of a category C is supposed to be found in the actual instances of C, so something apart from essence beliefs must determine what is counted in the category in the first place. This is where the identification procedure comes into the picture: the identification procedure determines which objects are ordinarily reckoned as C, and it is those objects whose deep structure (if they share any) determines the category boundaries. Experimental psychology suggests that the identification procedures typically consist of prototype or exemplar representations, which consist of the category members' typical perceptual features. However, the identification procedure can also contain theoretical beliefs and beliefs about essences, which can also guide our particular categorisation judgements—for instance, we don't categorise anemones as plants despite their plant-like appearance, and we don't categorise dolphins as fish despite their fishy appearance. It is important to distinguish these *specific* theoretical or essence beliefs from the *externalistic* placeholder beliefs. The former can constitute the identification procedure and don't determine reference, whereas the latter constitute the core and do determine reference (together with external facts). Importantly, the specific essence beliefs can be erroneous: for instance, whales were earlier believed to be (essentially) fish, but despite this false essence belief, the speakers still arguably managed to refer to whales in virtue of the animals' *external* essence.

We have ended up with a dual theory, where the semantically constitutive cores of natural kind concepts consist of placeholder essence beliefs, and the identification procedures consist of prototypes, exemplars, and specific theoretical or essence beliefs (as far as we know; they might also contain other representations or capacities). On this dual theory, the reference of a natural kind concept is determined by the concept's essentialist core as a function of what actually triggers the concept's identification procedure:

> (DTR) A natural kind concept C applies to x iff x shares some deep structure, believed to be essential, with most of the samples that actually trigger the identification procedure of C of some relevant speakers.

The 'most of' –specification is introduced to allow for there being some variation in the actual samples. For instance, fool's gold is sometimes called 'gold', but this doesn't make gold's essence to involve both Au and iron pyrites. Also we need to specify who are the relevant speakers whose linguistic behaviour we take to affect a natural kind term's reference: not all speakers' use of a term affects its reference. For instance, a madman might constantly refer with the term 'gold' to nothing but television sets, but this wouldn't make gold's essence include the deep structures of television sets. There might also be some other specifications we need to take into account in elaborating (DTR). It is worth noting that (DTR) doesn't try to be a universal theory of natural kind concept reference, but a theory of how these concepts *typically* refer in lay speaker use; it allows that natural kind concepts can be introduced differently, especially in some technical contexts. (Consider, for instance, the concept UNUNSEPTIUM which was introduced to stand for the element with the ordinal number 117 before the element was yet empirically discovered.)

The most important point left open in (DTR), and also in the formulation of externalistic essentialism (EE) above, is *what the reference determining deep structure specifically is*. In particular, I have left it open whether this deep structure coincides with the scientific or metaphysical essence, if any, of the samples actually referred to. I will return to this issue in section (4.1). I will next examine theories of reference in philosophy, which, like the considerations presented in this section, motivate the dual theory.

3. Theories of Reference

3.1. Description theory

The description theory of reference was examined already, in connection with the definition theory of concepts. The description theory holds that the reference of a term is determined by a description, or cluster of descriptions, associated with it in the minds of the speakers. The description specifies the necessary and sufficient criteria for belonging in the extension of the term. For instance, the term 'bachelor' is associated with the description *unmarried adult male*, and the term refers to anything that satisfies that description; 'water' is associated with a description such as *the colourless, odourless, thirst quenching substance that falls from the skies and fills lakes and oceans* (and so on). Unlike the classical theory of concepts which was committed to conjunctive definitions, the description theory of reference allows the descriptions to be disjunctive.

3.1.1. Plato's and Wittgenstein's problems

The description theory faces problems both generally and especially in the case of natural kind terms. The definition theory of concepts was undermined by Plato's and Wittgenstein's problems, which were that very few terms can be given necessary and sufficient application conditions. The description theory can, however, hold that the reference determining descriptions are disjunctive, or that reference is determined not by a single description, but rather by clusters of descriptions. Such descriptions could (in principle) be complex enough to capture categories such as JUSTICE or GAME, which are impossible to capture in conjunctive definitions. So, Plato's and Wittgenstein's problems don't undermine the modern description theories of reference as they do the definition theory, or at least not as directly.

3.1.2. Ignorance and error problems and the descriptivist's solutions

The most serious problems for the description theory are, once again, those of ignorance and error. These problems pertain to proper names and natural kind terms, but I will here focus solely on natural kind terms. For instance, before the molecular structure of water was discovered, the descriptive content associated with 'water' arguably picked out *any* substance that appeared watery to us, including H_2O, XYZ, and so on. However, according to Kripke's and the other externalists' intuitions, the term still referred solely to H_2O, if that was in fact the watery substance in our surroundings. Similarly, the descriptive content associated with the term 'gold' by scientifically ignorant speakers doesn't pick out gold in gaseous form, but the term still arguably refers to it even in their use, as gaseous gold is of the same natural kind as the samples ordinarily called 'gold'.

These are cases of ignorance, but error scenarios are even more pressing for the description theories, as in them even the actual extension of a term might turn out empty. Kripke's example is about gold. Suppose it turned out that gold isn't actually yellow, metallic, shiny, and so on, but green and murky; imagine that our perception of gold has been disturbed by some peculiarities in the atmosphere (or some neurological disease, or an evil demon, or what have you). Suppose also that we didn't yet know about the deep structure of gold, but instead referred to gold solely in virtue of its perceptual properties. In this case descriptivism entails that the term's extension is empty, as no substance whatsoever satisfies the description(s) associated with the term 'gold'. However, it seems that we *did* refer to some substance with the term 'gold', namely the substance that appeared to us goldish—we were just wrong about its properties. (Kripke 1980, pp. 116 – 119.)

An obvious strategy to try to solve the ignorance and error problems would be to introduce scientific properties in the descriptions, but this will not do for two reasons. First, also scientific descriptions can be erroneous, as illustrated in section 2.3.2. Secondly, it is plausible that speakers can refer to natural kinds even in the case no one in their linguistic community knows about the reference-determining deep structures.

As to the ignorance problems, the description theorist's best solution is to make natural kind term reference dependent on the nature of the actual samples. On the standard reading, descriptivism entails that a natural kind term applies to *whatever* satisfies the descriptive content associated with the term, but on the actualised version the term applies solely to whatever *actually* satisfies the description(s) (see e.g. Jackson

1998, pp. 205 – 206). However, the actuality-operator alone is not sufficient to guarantee that a natural kind term refers to a specific actual natural kind. Consider, for instance, the term 'water' and the pre-scientific description associated with it, *the colourless, odourless, thirst quenching substance that falls from the skies and fills lakes and oceans*. There are (at least) *two* kinds that actually satisfy this description, namely the natural kind H_2O and the non-natural kind *watery substance*. We need to specify that a natural kind term's reference is fixed to the actual *natural* kind instantiated by the samples that actually satisfy the description associated with the term. (See Article I.) This solves many of the ignorance problems: XYZ is not water, as it is not of the same natural kind as the actual watery substance; gaseous gold is indeed gold, as it is of the same natural kind as the actual goldish substance.

Unfortunately, this suggestion doesn't help in solving the error problems. As Kripke demonstrates, gold might not satisfy any of the descriptions we associate with 'gold' *even in the actual world*. Our perception of gold might be distorted and our beliefs about gold might be erroneous. Yet the term 'gold' could still refer via a causal chain to whatever we reckon to be gold, or what appears to us goldish. There is probably no way for the descriptivist to escape this problem other than by loosening the demand that the actual samples have to in fact *satisfy* the descriptions associated with the term. For example, the descriptivist might hold that the actual gold samples need not truly be yellow, shiny, and malleable, but only to be *reckoned* to be such. These refinements lead to the following, modified descriptivist account:

> (D) A natural kind term *t* refers to the natural kind instantiated by (most of) the samples actually reckoned to satisfy the description(s) associated with *t*.

We may title (D) a *psychologistic* formulation of descriptivism. It solves the error problems: an object can belong in a term's extension if it is reckoned to satisfy the descriptions associated with the term, even if it does not truly satisfy them.

These refinements solve the most pressing problems of the descriptivist account, but bring along novel problems.

3.1.3. Counter-argument from experimental psychology

There are two stands the descriptivist can take with respect to how the descriptions are associated with terms. Either the descriptions are conceived of as represented in the mind, and associated with terms in the minds of the speakers. Alternatively, following Frege, the descriptions can be considered as non-mental abstract objects, which are ‘grasped’ by individual speakers. The Fregean view faces serious problems in trying to account for what kind of entities the abstract descriptions specifically are, and how they can be ‘grasped’ by physical beings like us. I suspect that the former, psychologistic variation of descriptivism is the orthodox position among contemporary descriptivists (see e.g. Reimer 2003). This raises a serious threat to descriptivism: in holding that speakers associate descriptions with their terms, the view is committed to an empirical claim about human psychology. And there are strong reasons to doubt the correctness of this view.

We saw above that experimental data crucially undermines the classical theory of concepts, which equates concepts with definitions. The description theory, however, has two advantages over the classical view. First of all, it need not hold that reference be determined by *conjunctive* descriptions; instead, the descriptions can be disjunctive or there can be clusters of them. Second, since the description theory is primarily a theory of reference, not categorisation, the experimental data concerning typicality effects in categorisation don’t undermine it in the same way as they do the classical theory. Despite these differences, the experimental data arguably undermines also the description theory, both as traditionally formulated and especially in its psychologistic variety (D).

The descriptions associated with a term are traditionally thought to consist of the *identificatory information* which determines the term’s extension. On the orthodox, non-Fregean reading of descriptivism, this information is supposed to be represented in the speaker’s mind, associated with the referring term. The information is typically taken to consist of perceptual features such as, in the case of ‘water’, *the wet, thirst-quenching, transparent liquid that flows in rivers and taps...* (e.g. Wikforss 2008, p. 159). But experimental psychology quite convincingly shows that identificatory information of this kind is *not* in fact represented as descriptions of any kind, but probably as prototypes or exemplar sets. This problem is emphasised if we consider the psychologistic formulation of descriptivism (D). This view is committed to the claim that speakers actually *categorise* objects in virtue of descriptions, but this view is even more directly in conflict with the experimental data—it is not descriptions, but prototypes or exemplars which determine people’s categorisation judgements.

A quite natural response strategy for the descriptivist would be to claim that her notion of description is an *abstract account* of people’ identificatory knowledge, independent of how the information is specifically represented. For instance, people might represent identificatory information about water in the form of exemplars of particular instances of water, but the represented information could nevertheless be *described* with the description *the wet, thirst-quenching, transparent liquid that flows in rivers and taps…* The descriptivist might hold that she's not making a psychological claim about how people actually represent identificatory information, but only giving an abstract description of that identificatory information. Unfortunately for the descriptivist, this strategy doesn't succeed, as the identificatory information represented in terms of prototypes or exemplars cannot be even adequately described in terms of descriptions, even if we allow the descriptions to be disjunctive.

Identificatory knowledge, if represented as prototypes or exemplars, is extremely hard, if not impossible, to capture in descriptions. To illustrate, imagine a prototype P which consists of represented features F_1 – F_{10}, each assigned the typicality value 1 Suppose that the similarity between P and an object is calculated according to Tversky's contrast principle (CP in section 2.2), and that the threshold value for P is 1.9. Now, consider which kind of objects can trigger P: an object which has features from F_1 to F_6 but lacks features F_7 – F_{10}; *or* an object which has features F_1 – F_5 and F_{10} but lacks features F_6 – F_9; *or* an object which has features F_2 – F_7 but lacks F_1 and F_8 – F_{10}; and so on. A disjunctive description of the objects that can trigger a prototype quickly becomes extremely complex. So, a description of the identificatory knowledge associated with, say, 'water' as *the wet, thirst-quenching, transparent liquid that flows in rivers and taps…* can be counted as an oversimplification at best, and simply wrong at worst.

In sum, it is hard to make traditional, non-psychologistic descriptivism compatible with how people in fact recognise objects, or how the identificatory information is represented. Moreover, the account falls prey to the error problems, which are arguably fatal for it. The psychologistic variation of descriptivism (D) survives the error problems, but it contradicts with how people as a matter of fact categorise objects: people don't actually recognise objects in virtue of descriptions, not even disjunctive descriptions, but, according to best of our knowledge, in virtue of prototypes or exemplars. So, (D) is simply false.

The descriptivist's best solution is to refine (D) into a form where any commitments to specific psychological matters are minimised. The descriptivist could hold that a term's reference is fixed through some *identificatory capacities*:

> (D') A natural kind term t refers to the natural kind instantiated by (most of) the samples that actually trigger the identificatory capacity associated with t.

(D’) solves the Kripkean problems of error and at the same time is compatible with experimental psychology, as it doesn’t take a stand on how the identificatory capacities are realised. (A similar suggestion is put forward by the descriptivist Martin Davies (2004, pp. 115 – 116) who suggests that a term’s reference could be fixed by some ‘sub-personal device’ associated with the term. He doesn’t, however, elaborate on this idea.)

It is a terminological matter whether (D’) is a form of descriptivism at all anymore. The proposed account suggests that a natural kind term’s reference is fixed to whatever actually triggers the identificatory capacity associated with the term, and the triggering of the capacity is a causal relation. However, the account is still descriptivist in the sense that the speakers’ identificatory knowledge has an important role in fixing the reference of a term, although the identificatory knowledge cannot be considered strictly as descriptions.

Even (D’) still has, if not strictly a defect, at least a deficiency. The proposed descriptivist account relies on the actuality operator to fix a term’s reference to a specific natural kind, but doesn’t explain *why* terms are actuality dependent in this way, or *what makes* them such.

3.1.4. Actuality operator vs. essence beliefs

The role of the actuality operator in (D’) is to guarantee a natural kind term’s reference to a specific natural kind instead of any kind that happens to fit the identificatory knowledge associated with the term. As already noted, the actuality operator alone is not, however, sufficient to guarantee reference to a specific *natural* kind, as the samples actually referred to by a natural kind term instantiate at least two kinds—for example, the samples referred to as ‘water’ instantiate the natural kind water (H_2O) and the non-natural kind *watery substance*. The most obvious way to guarantee a natural kind term’s reference to a specific natural kind is simply to stipulate, as in (D’), that the term refers to a natural kind. But *what makes* the term refer thus? The problem is emphasised if we consider scientifically ignorant speakers, who may not possess the concept of natural kind—do they still manage to refer to natural kinds?

It is plausible to suppose that natural kind terms refer to kinds with a certain actual deep structure because of speaker intentions or beliefs. These intentions or beliefs *cause* it that some terms are used actuality dependently—we *could have* used natural kind terms irrespectively of what turns out to be the deep structure of the samples actually referred to. For example, we might categorise an instance of XYZ as water and not change our judgement even upon learning that the sample is different in deep structure from actual water. In this case natural kind terms (or at least terms with an appearance similar to our natural kind terms) would refer like nominal kind terms, such as ‘chair’. Even if all chairs turned out to share some molecular structure X, we wouldn’t care about this fact in categorising objects as chairs—an object lacking X but satisfying all the hallmarks of chairs would be a chair. What are the speaker intentions or beliefs that make natural kind terms refer to natural kinds?

I suggest that what makes natural kind terms actuality dependent are the speakers’ placeholder essence beliefs: people believe natural category members to share some hidden, empirically discoverable deep structure, whose possession they take to be necessary and sufficient for belonging in the category. We have ended up in a dual theory just like (DTR): the reference of a natural kind term is fixed to (the objects possessing) the deep structure shared by the samples that actually trigger the identification procedure associated with the term.

3.2. Causal theories of reference

We already dealt with Fodor’s informational semantics, which can be considered a causal theory of reference, though, coupled with conceptual atomism, it serves also as a theory of concepts. In this section I will focus on the causal theories originally due to Kripke and Putnam.

Both Kripke’s and Putnam’s formulations of the causal theory are quite sketchy. Kripke suggests, rather metaphorically, that the reference of a natural kind term is fixed in a ‘baptism’. For instance, Kripke suggests, the term ‘gold’ might have been introduced by pointing to a set of samples ordinarily called gold and stipulating that ‘[g]old is the substance instantiated by the items over there, or at any rate, by almost all of them’ (1980, p. 135). Ever since this ostensive ‘definition’, the term ‘gold’ applies solely to the substance instantiated by the samples. While Kripke’s suggestion is only metaphorical, Putnam’s proposal is more realistic. He suggests that natural kind term reference is fixed through ordinary use of the term. For instance, the reference of the term ‘gold’ is fixed to the substance instantiated by the samples the term is ordinarily used to refer to.

I will examine the problems of the causal theories only briefly here, but they are investigated in more detail in Article I. The most pressing problems of the causal theories are the so-called *qua*- and composition problems (see e.g. Brown 1998).

3.2.1. The *qua*- and composition problems

On the causal theories, the reference of a natural kind term is fixed to the kind instantiated by the samples called by that term. The *qua*- problems stem from the fact that the samples called by a term typically instantiate many kinds—for instance, the samples called 'gold' instantiate (among many others) the kinds *valuable substance*, *shiny substance*, *element*, *Au*, and the isotope ^{197}Au, and the problem is to specify why the reference of 'gold' is fixed solely to *Au*. (Brown 1998; Devitt and Sterelny 1999.) The composition problem is that the samples used in reference fixing are typically impure, and the causal theorist needs to explain why the reference of a natural kind term is fixed just to the pure substance. For example, the samples actually called 'water' typically contain salts, minerals, chlorine, and other impurities, but still we take the term 'water' to refer solely to H_2O, not to some disjunctive kind *H_2O plus salt x or mineral y or chlorine or….* (Brown 1998.)

These problems can be taken to indicate that purely causal relations alone aren't sufficient to determine the reference of a natural kind term, and that we have to allow for some, at least minimal, descriptive specification of the referent as well. We need to specify that a natural kind term doesn't refer to *any* of the kinds actually called by the term, but to some specific kind. For example, we need to specify that the term 'water' refers solely to the natural kind that appears watery to us, instead of any compound, solvent, or liquid that the samples called 'water' actually instantiate in addition to H_2O. Michael Devitt and Kim Sterelny (1999) suggest that this be done by some descriptive content associated with the term (for a more elaborate suggestion of this kind, see Stanford and Kitcher 2000). However, I have argued (Article I) that these suggestions, call them causal-description hybrid theories, fall prey to the error problems that undermined the description theory: it might turn out that, say, water is not in fact watery, but its appearance is distorted somehow.

These problems are solved, as in the case of the description theory, by relying on some identificatory capacities associated with the terms. Jessica Brown (1998) puts forward a suggestion like this and holds that a natural kind term's reference is fixed to the natural kind that actually triggers the *recognitional capacity* associated with the term. For instance, 'water' does not refer to any compound or liquid, since not any compound or liquid triggers the recognitional capacity for water, but only H_2O does. Again, water does not refer to the impurities of the actual water samples, since they are not causally responsible for (most of) the perceptual properties in virtue of which we recognise water: even if water lacked the impurities, it would still appear watery and trigger the recognitional capacity. We may reasonably suppose that this ‘recognitional capacity’ is identical with the identification procedure of the concept associated with the term.

Refining the causal theories leads, thus, to an account that is practically identical with the refined description theory (D’): a natural kind term’s reference is fixed to whatever actually triggers the identification procedure associated with the term.

3.2.2. Indexicality

Kripke suggests that the reference of a term is fixed in an initial causal baptism to a specific natural kind in our actual surroundings. Likewise, Putnam suggests that the reference of a natural kind term is fixed through being constantly applied to some instances of a natural kind in the speakers’ surroundings. Here we may ask, just like in the case of the description theory and the actuality-operator, *why* are natural kind terms actuality dependent, or *what makes* them such? We could have used natural kind terms just like we use nominal kind terms, irrespectively of the deep structures of the actual samples—so why do we use natural kind terms like we do? The answer is, once again, externalistic essentialism: we believe natural kind members to share some hidden, empirically discoverable deep structure, and take the terms to refer in virtue of this deep structure even when it is not known. Natural kind term reference is not determined causally by coincidence—we ourselves *make* them refer causally, or in virtue of the unknown deep structure of the actual samples. This leads us to an account just like the dual theory (DTR): the reference of a natural kind term is fixed to whatever shares some relevant deep structure with the samples that actually trigger the identification procedure of the concept associated with the term.

In this section I have argued that refining both the description and causal theories of reference leads to a dual theory like (DTR). The description theory falls prey to the problems of ignorance and error, and it also conflicts with experimental psychology. To solve these problems, it has to be refined into the form (D’). This still leaves the problem of accounting for why natural kind terms are actuality dependent, or why they refer in virtue of some unknown deep structure of the actual samples. The obvious answer is that we ourselves make them refer thus; we intend to use natural kind terms to refer in virtue of some unknown deep structure. On the other hand, the causal theories of reference fall prey to the *qua*- and composition problems, which necessitate introducing some identificatory capacities on the account. However, as on the description theory, one matter is left unexplained on the causal theory: why do natural kind terms refer just to some actual natural kind instantiated by the samples called by the term? The answer is externalistic essentialism, which implies that we ourselves intend to use natural kind terms refer causally. Thus, refining both the description and causal theories leads to a dual theory like (DTR).

4. General Discussion

In this section I will briefly examine some consequences and applications of the proposed dual theory (DTR). I don't intend to provide an extensive discussion of the issues here, but rather, point towards future refinements and applications of the theory.

4.1. Real versus believed essences

The proposed dual theory (DTR) is in most respects equivalent to the refined versions of the description and causal theory, but there is one important difference between the accounts.

The dual theory holds that a natural kind concept has an externalistic essentialist core which determines its reference as a function of what actually triggers the concept's identification procedure. The essentialist core consists of a belief in some 'substance, power, quality, process, relationship, or entity that causes other category-typical properties to emerge and be sustained, and that confers identity' (Gelman 2003, p. 405). What specifically is this deep structure? Clearly it is some empirically discoverable structure or entity, but psychological essentialism leaves it open whether it has to coincide with the category's essence *in the metaphysical or scientific sense*. Externalistic essentialism does hold that the essence has to be some deep structure *in fact* possessed by the actual samples, but it doesn't have to be essential in the strict scientific or metaphysical sense. The essentialists are keen to emphasise that psychological essentialism isn't committed to any metaphysical claims about essences, but that the view is strictly a doctrine about people's essentialist beliefs (e.g. Gelman 2004, p. 405; Gelman & Wellman 1991). The traditional externalistic theories of reference, on the other hand, explicitly hold that the reference of a term is fixed to the specific natural kind (if any) instantiated by the samples the term is actually used to refer to. (I am supposing here that if a kind is natural, then its members share some scientific essence, such as H_2O in the case of water or Au in the case of gold.)

This is a subtle but important difference: psychological essentialism entails that speakers can use a natural kind term in virtue of some deep structure X of the samples referred to even if X cannot strictly speaking be considered as an essence. An important domain where this difference has substantial consequences is biological species, which probably don't have essences in the strict sense at all (see e.g. Hull 1965; Ereshefsky 2007). But irrespectively of the philosophical controversies about essentialism concerning species, various studies have demonstrated that people still *use* species terms essentialistically, or dependently on some unobservable deep structure. For instance, a transformation of an animal's appearance doesn't change its category membership as long as its lineage and insides are kept intact, or an animal appearing to be a member of a species K may be categorised as a non-K if its insides are discovered to differ from that of the other K-members', and so on (Keil 1989). It doesn't matter at all even if the deep features determining people's categorisation judgements cannot be considered as the animal's essence in the strict, metaphysical sense: people still take membership in biological species to be determined not solely by perceptual features, but also by some hidden deep features.

This difference gives externalistic essentialism and (DTR) an advantage over the traditional theories. The traditional theories hold that if the actual samples don't instantiate some natural kind, the term either doesn't refer at all, or it is not a natural kind term. The former option is clearly unacceptable—say, the term 'cat' refers to cats even if cats weren't strictly a natural kind. What about the latter alternative? This would be to claim that species terms function like nominal kind terms, which apply to anything that satisfies the identificatory knowledge associated with the term. But this clearly is not the case: biological species terms apply similarly as any other natural kind terms, namely (at least partly) externalistically. For instance, even if no one in our linguistic community could distinguish cats from cleverly constructed robot cats, the term 'cat' would arguably still refer solely to cats, not robots (supposing that all the actual cats are in fact mammals). In other words, the reference of biological species terms is not determined solely by the identificatory information associated with them, but also by some external facts about the actual samples' deep structures. The traditional accounts have trouble explaining this phenomenon, whereas psychological essentialism has a natural explanation for it: even if the samples referred to don't strictly speaking instantiate a natural kind, people can nevertheless have essentialist beliefs about them and refer with the terms externalistically.

It is important to notice that externalistic essentialism doesn't imply that speakers can *ignore* scientific facts about the samples referred to. If the samples actually referred to with a term don't share any deep features whatsoever, they arguably cannot be referred to externalistically—there simply is no deep structure in virtue of which the term could refer. The reference determining deep features have to be some real features of the actual samples, externalistic essentialism only leaves some *freedom of choice* with respect to which of the properties are taken to determine reference. For instance, cats might actually share lineage L, genome G, or physical makeup M, and it is up to the speakers (possibly some experts) to decide which of these deep structures determine the extension of the term ‘cat’.

We may suppose that *if* the actual samples called by a term in fact share some real, scientific essence (like H_2O in the case of water), *then* this deep structure is what determines the term’s reference also on psychological essentialism. To repeat, psychological essentialism holds that people believe natural category members to share some ‘substance, power, quality, process, relationship, or entity that causes other category-typical properties to emerge and be sustained, and that confers identity’ (Gelman 2003, p. 405). If the actual category members do in fact share some deep structure of this kind, whatever that is, then it determines the category’s boundaries even according to psychological essentialism. For instance, water’s typical properties are, as a matter of fact, caused by its consisting of two hydrogen and one oxygen atoms. If a speaker knew about this (and had sufficient cognitive conditions and scientific competence), then she should most naturally consider being H_2O as water’s essence. (In practice, of course, lay speakers often don’t have an opinion about what the category essences are, as that is beyond their scientific competence, but they may defer to experts; c.f. Putnam 1975.)

Is it possible that a speaker takes the essence of a category to be something other than it really is? In other words, does the proposed variety of psychological essentialism fall prey to a variety of the ignorance and error problem? As suggested above, it is plausible that if the category in fact possesses some scientific essence, then the speakers will endorse that as the category essence, but at least in principle a speaker could endorse some other deep structure as the category essence. For example, suppose that, while the scientific essence of water is H_2O, a speaker believes it to be the property of containing hydrogen. Doesn’t psychological essentialism imply that in this case the speaker refers, incorrectly, to any substance containing hydrogen with her term ‘water’, instead of just water (that is, H_2O)? In order to see whether the speaker *truly* refers to any substance containing hydrogen with her term ‘water’, we have to make sure that her decision is *considered*. First, we have to make sure she understands that containing hydrogen doesn’t alone cause the typical features of water, but that it is being H_2O that does so. Second, we have to make sure that the speaker is sufficiently intelligent to understand relevant facts about atoms and molecules, and how they cause perceptual properties. Possibly the speaker would also need to consult scientists about the subject matter. If the speaker met these requirements *and still* held that the essence of water is simply to contain hydrogen, then we'd have to conclude that she indeed refers with her term 'water' to *anything* containing hydrogen. This is not a case of ignorance and error, but only of non-conventional use of the term 'water'.

So, in case the actual samples referred to with a term do possess some scientific essence, psychological essentialism implies that (typically) the term refers in virtue of that deep structure. But what about cases when the actual samples don't strictly speaking share a scientific essence, like biological species? In this case I'm prepared to maintain that the reference of the category term is (at least partly) underdetermined: there is no mind-independent deep structure that could be unambiguously considered as *the* essence of the category. However, it is still possible that in the course of scientific progress we *decide* to define the category essence as some specific deep structure, in which case the category term obtains a definite extension. (A suggestion of this kind is developed by Joseph LaPorte (2004).)

4.2. Non-natural kind concepts

Thus far I have mostly been concerned with natural kind concepts, but what about the vast range of other concepts—does some variety of the dual theory encompass them? More specifically, do other terms besides natural kind terms have semantic cores and distinct identification procedures, and how are the cores, if any, structured?

4.2.1. Well-defined technical concepts

Arguably at least some mathematical, logical, or other such technical concepts may have definite cores. For example, the geometrical concept SQUARE can be defined as a rectangle which has four right angles and parallel sides (in Euclidean geometry); the mathematical concept RATIONAL NUMBER can be defined as a number which can be expressed as a ratio of two integers; the logical concept NEGATION can be defined as *not*; and so on. Some experimental findings suggest that even well-defined concepts like these may have separate identification procedures. Sharon Armstrong and Lila and Henry Gleitman (1983) found typicality effects concerning what they take to be well-defined categories, namely EVEN NUMBER, ODD NUMBER, PLANE GEOMETRY FIGURE, and FEMALE (the last category may not be considered well-defined, but let's ignore that). For instance, subjects tend to judge the number 4 more typical as an even number than 106, or square more typical as a plane geometry figure than ellipse (p. 276, table 1).

Accordingly, subjects categorise the more typical samples in their corresponding categories more quickly than the less typical samples (p. 286, table 4). These findings clearly don't show that concepts like EVEN NUMBER or PLANE GEOMETRY FIGURE *are* prototypes, but they can be interpreted as showing that even these concepts have prototype-structured identification procedures; the semantically constitutive core is still definitely definable. (Granted, lay speakers don't often know the definitions of mathematical concepts, in which case the concept cores are arguably determined by the experts on the field; c.f. Putnam's (1975) suggestion about the division of linguistic labour.)

The identification procedures of well-defined concepts, if they have any, are semantically totally irrelevant. For instance, that the number 4 is judged more typical as an even number than 106 doesn't mean that 4 would have a higher degree of membership in the category EVEN NUMBER—numbers are even if and only if they are dividable by two, and *any* such number is just as much an even number as any other. A variety of the description or definition theory might well be viable as an account of the cores of well-defined concepts, while their identification procedures might best be modelled on the prototype or exemplar account.

4.2.2. Artefact concepts

Some recent experimental studies indicate that artefact concepts are essentialistically structured (see e.g. Bloom 1996; Kelemen 1999; Kelemen & Carey 2007; but see also Sloman and Malt 2003). The essence of an artefact is suggested to be the artefact's *intended function*—for instance, chairs are intended to be used for sitting, doors for shutting doorways, lamps for giving light, and so on. Crucially, these features are supposed to override the category members' typical perceptual features in determining category membership. For example, a modern, abstractly shaped chair is a chair because of its intended chair-function even if nobody (except for its manufacturer) would recognise it as one. An object may even lack the typical function of an artefact category if it is nevertheless intended for that function: the modern chair may be impossible to sit on because it is too weak or inconveniently shaped. (C.f. Bloom 1996, pp. 2 – 3.)

An important aspect of artefact kind essentialism is that the essence of a category can be distinct from any of the prototypical features of the category: triggering the identification procedure is neither necessary nor sufficient for belonging in the extension of an artefact concept. For example, an object may appear to be a chair but nevertheless not be one; say, it is a piece of art made of plaster. Again, an object may *not* appear to be a chair but still be one—it might be a peculiarly sculptured modern chair. These points suggest that artefact concepts consist of distinct identification procedures and semantic cores (though see Sloman and Malt 2003).

4.2.3. Social concepts

Essentialist tendencies among lay speakers have been found also in other domains besides natural kinds and artefacts, particularly some social categories. Nick Haslam, Louis Rothschild, and Donald Ernst (2000) list as categories high in essentialism gender, ethnicity, race, and disability, among others. Haslam's et al. and others' experimental studies suggest that people tend to conceive some social categories as natural kinds and think that, say, human race is determined by some hidden deep features, which cause the features typical of that race—contrast this with the non-essentialist belief that race is a social construct (cf. Haslam et al. 2000, p. 114 ff.). However, the notion of essentialist, semantically constitutive cores might be problematic in the case of social concepts, as it is unclear whether categories such as ethnicity or race can be delineated in terms of any real deep properties such as genes. For this reason it is also unclear whether social concepts refer externalistically like true natural kind terms —for example, could there be a person that would satisfy all of the hallmarks of being white but would nevertheless not be white in virtue of some unknown deep properties such as genes? Probably not. Finally, even if lay speakers did use social concepts essentialistically, many experts do not: there is an ongoing, heated debate about the naturalness of categories like race or ethnicity. (It is also noteworthy that the researchers themselves studying essentialism about social concepts don't take their view to make any semantic commitments, but the view is generally considered to have explanatory value mainly in the study of social cognition.)

4.2.4. Philosophical concepts

We have thus far encountered three types of concept cores. First, natural kind concept cores consist of externalistic essence beliefs; for instance, if water is actually H_2O, then the term 'water' refers to all and only H_2O. Second, artefact concepts refer in virtue of the category members' intended function; for instance, the term 'chair' refers to whatever is intended for sitting (or whatever chairs are actually intended to be used for). Third, mathematical, logical, and other technical definite concepts refer in virtue of a description or a definition; say, a square is a rectangular polygon with four sides.

A special case are philosophically interesting concepts, such as those of knowledge, justice, virtue, good, beauty, and so on. Philosophers have spent millennia on trying to provide necessary and sufficient application conditions for these concepts, without much success. Do these concepts have any kind of definite cores? There are probably two alternatives with respect to how the cores might be structured; either they are definitions, as in the case of mathematical or other technical terms, or they are externalistic, as in natural kind concepts (it is highly improbable that, say, justice or knowledge would have some intended function like artefacts do). Trying to find analytic definitions for these concepts is the traditional way: for instance, it is suggested that knowledge is justified true belief, or that an act is just if and only if it maximises utility, and so on.

The definition strategy has been criticised recently. For instance, William Ramsey (1992) suggests that the over two millennia long history of failures to provide generally accepted definitions for philosophical concepts suggests that these concepts may not have definitions at all. Ramsey also draws on experimental psychology and suggests that philosophical concepts might be prototypically structured, and thus lack definitions. This line of thought has been replied to by claiming that even though philosophical analysis couldn't ever provide necessary and sufficient conditions for the application of philosophical concepts, the analysis is still fruitful and interesting—philosophical analysis has deepened our understanding of what is (and is not) knowledge, justice, virtue, and so on (see Sandin 2006). Against Ramsey it can also be noted that it is unclear whether philosophical concepts are indeed prototypes: there are no experimental studies that have addressed this issue. Moreover, the long history of analysis of philosophical concepts can be interpreted to count *against* the view that they are prototypes: if philosophical concepts were indeed prototypes, it would probably be very *uninteresting* to try to define them (try to define the prototype of chair, game, bachelor, and so on). The fact that philosophical concepts have interested many great minds for over two millennia can be interpreted as showing that these concepts must involve some interesting, complex theoretical elements. This interpretation is not far-fetched, as the theory-theory might well encompass, if not all concepts, at least a very wide range of them (see Murphy & Medin 1999).

What about the strategy of conceiving philosophical concepts as natural kind concepts? Hilary Kornblith (e.g. 2002), following Quine (1969) argues that *knowledge* is a natural kind, and that epistemology should investigate this kind itself instead of our concept of it. Kornblith compares knowledge with aluminium: In trying to learn about aluminium, it is uninteresting to analyse our *concept* of aluminium, such as that aluminium is typically grey and silvery and used for building airplanes. What is interesting is to investigate aluminium *itself*—what its deep structure is, how it reacts with other substances, and so on. Analogously, according to Kornblith, epistemologists shouldn't be analysing our concept of knowledge, but knowledge itself.

Maybe the main problem of naturalised epistemology is that it is unclear whether knowledge is in fact a natural kind. The whole endeavour of naturalised epistemology rests on the presupposition that knowledge is a natural kind, but I know of no theorist that would have extensively examined the vast literature in cognitive psychology and animal ethology to find out *what kind of a natural kind* knowledge might in fact be, if any. Rather, the discussion has revolved around the philosophical implications of considering knowledge as a natural kind. Yet many or most of the philosophical arguments against naturalised epistemology rest just on what kind of a natural kind knowledge is taken to be.

The strategy of conceiving of philosophical terms as natural kind terms is even less plausible in the case of concepts like JUSTICE, VIRTUE, or GOOD. The long history of failure to define philosophical concepts suggests that these concepts also lack definitions. It is improbable that they are prototypes or exemplar sets either, since the long history of philosophical study suggests that they involve much more complex, theoretical aspects than simple prototypes or sets of exemplars would. An account of a prototype or exemplar concept could be given basically in terms of a list of the features which are more or less typical of the category members, but it seems that philosophical concepts cannot be captured this easily. How, then, are philosophical concepts structured? If the theory-theorists are correct in claiming that even artefact concepts involve theoretical elements, it is well possible that so do philosophical concepts (note that this does not necessarily mean that they should have definite concept cores). In the face of this, the prospects of the traditional philosophical study of these concepts don't appear quite as poor as Ramsey argues. Even though it is improbable that philosophers could ever reach simple definitions for (most) philosophical concepts, they can clarify and deepen our understanding of them, and bring out conceptual problems.

4.3. The dual theory and the desiderata for a theory of concepts

Thus far I've examined mostly how concepts refer and determine people's categorisation judgements, but concepts are supposed to have many other tasks in addition to these two. The presented dual theory is mainly a theory of reference, but it also takes a stand on how natural kind concepts are generally structured. Thus, it cannot be evaluated simply in terms of how successful it is in explaining reference; we must also take into account the other tasks of concepts, and the desiderata for a theory of concepts.

Laurence and Margolis (1999, p. 72; see also Prinz 2002, chapter I) claim that concepts, or a theory of concepts, should provide an explanation for (at least) the following phenomena:

- Fast categorisation
- Considered acts of categorisation
- Semantic application
- The licensing of inductive inference
- Analytic inference
- Concept acquisition
- Compositionality
- Stability

Let us briefly look at each of these requirements and examine how the proposed dual theory tackles with them. My main emphasis will again be on natural kind concepts, but I will also briefly touch upon other concepts. I will have to leave the discussion in the following sections somewhat sketchy, as an extensive investigation of the topics lies outside the scope of this introductory essay.

4.3.1. Categorisation and reference

As suggested in section (2.5), the identification procedures typically involve prototypes or exemplars, which enable quick, rough-and-ready categorisation, but they can also include more theoretical beliefs. For instance, even though we typically recognise plants by their appearance, we nevertheless know that, despite their resemblance to the plant prototype or exemplars, corals and anemones are not plants. (As already noted, it is important to distinguish these *specific* essence beliefs from the reference-determining, *externalistic* essence beliefs.) Any categorisation judgement based on the identification procedure can go wrong: an instance of a watery substance might turn out not to share deep structure with the actual water samples; or it might turn out that corals and anemones are not animals after all. The concept cores are the ultimate arbiters of categorisation, and determine reference. Whatever truly is water is determined by our externalistic essence belief about water, which makes the term apply to whatever shares some relevant deep structure with the actual water samples. Accordingly, a categorisation judgement based on the externalistic essence belief cannot go wrong (supposing the speaker meets certain requirements). Say, if I know that an instance of a watery substance shares some relevant deep structure with the actual water samples, and if I am sufficiently intelligent and competent in science, then I cannot be wrong about the sample's being water.

Similar considerations apply also to some non-natural concepts which might have distinct cores and identification procedures. We may recognise grandmothers in virtue of their having grey hair, wearing glasses, and so on; or we may recognise chairs in virtue of their typical shape. Any categorisation judgement made solely on the basis of the identification procedure can, however, go wrong: an old woman with grey hair and glasses might turn out to have no children, or the chair-shaped object might turn out to be a piece of abstract art. The ultimate criterion for being a grandmother is be female and have grandchildren; to be a chair is to be intended for sitting; and so on.

We may distinguish between more or less defining features, and more or less reliable categorisation judgements, even in the case of concepts which lack strictly distinct cores and identification procedures. For instance, *right acts* may have certain salient properties in virtue of which we can easily and quickly recognise them. However, such quick categorisation judgements may go wrong, as upon further investigation and reflection we may find out that a particular act was not in the end right. For instance, we may make a prima facie judgement that the killing of a murderer is right, but upon reflecting on our moral intuitions and relevant contingent matters of fact, we may end up concluding that the earlier judgement was wrong and that the killing is after all wrong. This might be so even if there eventually is no definite matter of fact whether the deed is really right or wrong.

4.3.2. Inductive and analytic inference

What about the licensing of inductive and analytic inference? Quite clearly, it is the task of the identification procedures to determine the former, and the task of concept cores to determine the latter. For instance, we can inductively infer on the basis of our prototypical representation of water that water is typically odourless, colourless, thirst-quenching, wet, flows in rivers and fills lakes, and so on. These inferences aren't, however, analytic or conceptual: it is not logically necessary that water possess any of these features. The only analytic truth about water is determined by the semantic core of the concept WATER: necessarily, any sample of water possesses the relevant deep structure (if any) shared by the actual samples called ‘water’. This is a necessary truth about water, as the reference of the term ‘water’ is ultimately determined by the actual samples’ deep structure. It is also an *a priori* truth, that is, knowable through reflection: the reference-determining placeholder essence beliefs constitute concepts, which are mental representations and in principle accessible through reflection. Similar considerations hold for other concepts: we may inductively infer of grandmothers, on the basis of the concept’s identification procedure, that they are typically grey-haired and wear glasses; similarly that chairs typically possess legs, a seat, and a backrest. However, grandmothers aren’t necessarily grey-haired, and chairs don’t necessarily have legs, a seat, and a backrest (consider a beanbag chair). Analytic inferences are driven by the core: it is analytic and necessary that grandmothers have grandchildren, and that chairs are intended for sitting.

The division between the analytic and synthetic may be somewhat fuzzy in the case of some concepts, which can be taken to indicate that the cores and identification procedures of these concepts aren’t strictly separate. For instance, it is not clear whether it is an analytic truth about bachelors that they are unmarried adult males: say, the pope seems not to be a bachelor although he satisfies these criteria. Again, some concepts arguably lack cores altogether, and thus don’t entitle strictly analytic inferences at all. Such concepts might include social or philosophical concepts such as RIGHT, JUSTICE, GOOD, and so on, which have proven to be extremely difficult to provide any definitions for. For instance, an act that produces the greatest overall utility is typically right, but sometimes not (as in the case of killing of a scapegoat); good intentions typically elicit good acts, but sometimes not; and so on.

The proposed account of analyticity makes minimal metaphysical commitments; in particular, it isn’t committed to there being any mind-independent, abstract analyticities. On my account, a truth involving a concept C is analytic just in case it is true in virtue of the reference-determining conditions represented in the concept C. Since this account of analyticity does not rely on notions such as synonymy, but instead is purely naturalistic, there are chances that it escapes the Quinean problems of analyticity.

4.3.3. Concept acquisition

The desideratum that concepts should be acquirable addresses mainly Fodor’s theory of concepts, which many, including Fodor himself, take to entail that concepts can *not* be acquired, and that they are innate. This is roughly for the following reason. On Fodor’s account the reference of a concept is fixed through dispositions to token that concept, and accordingly, a subject cannot possess a concept unless she has some specific concept tokening dispositions. But now, a subject cannot be disposed to token a concept unless she already possess it. Fodor's solution to this problem is to hold that almost all concepts are innate, which makes his account deeply implausible in the eyes of many. (However, Margolis (1998) argues that concepts on Fodor's account *can* be learned through acquiring the mechanisms sustaining the concept tokening dispositions.) The proposed dual theory has no trouble accounting for concept acquisition, as prototypes, exemplars, and essence beliefs can be acquired (see e.g. Margolis 1998, p. 359 ff.). This gives the proposed dual theory an advantage over Laurence and Margolis' account, where the concept cores are atomistic and may not be learnable.

4.3.4. Compositionality

As we saw in the section concerning prototypes, the main problems in accounting for compositionality pertain to vague concepts. In positing at least some concepts with definite cores, the dual theory can escape these problems. With respect to these concepts we can endorse the classical theory of compositionality, where the first, semantic stage can be modelled on classical logic and set theory. Again, in the second stage a subject forms a representation of (a typical member of) the complex category, and this process may involve versatile psychological processes. The dual theory does not take a stand on how the second stage is realised, but we may reasonably suppose that this stage is not, in the end, crucial for cognition: we can possess a concept even if we are incapable of forming a specific representation of its typical instances (recall the example about Boolean concepts, such as NON-CAT).

In fact, it seems that whether concepts can or cannot compose is a question independent of how they are structured; that is, if we consider purely extensional or semantic compositionality. For instance, the concept NON-CAT applies to whatever does not belong in the extension of CAT, and it is another question *what determines* what belongs in the extension of CAT—whether it is prototypes, essence beliefs, or whatever. Similarly for other complex concepts, such as MONSTER BANANA, PET FISH, and so on. The constituent concepts determine what belongs in their respective extensions, and the remaining theoretical work is done purely on the logical or semantic level: an object belongs in the extension of the complex concept MONSTER BANANA if and only if it belongs both in the extension of MONSTER and BANANA, and so on. Thus conceived, the problem of compositionality (if there is such) is independent of specific theories of concepts.

4.3.5. Stability or concept sharing

By stability it is meant that concepts should be shareable between different individuals, or by a single individual at different times. The traditional theories of concepts which identify concepts with some set of identifying knowledge have trouble accounting for stability, as the identificatory knowledge can vary between individuals, or within an individual at different times. For instance, a speaker may be ignorant of the deep structure of gold and believe that gold is a compound, whereas another speaker may believe (correctly) that gold is an element. The traditional theories entail that these two speakers don't strictly share concepts, but nevertheless it very much seems that they aren't talking past each other: they are talking about the same substance, gold, and disagreeing about its deep structure. Again, a single speaker may change her beliefs about gold—say, the ignorant speaker can learn that gold is not in fact a compound, but an element. In this case it seems that she learns something about gold, not that the meaning (that is, reference) of her concept GOLD changes. If the concept GOLD had indeed changed meaning, then the speaker couldn't have been wrong about the deep structure of gold, since then she would have failed to refer to gold.

The dual theory accounts for stability readily, as it holds that any specific identificatory knowledge is not semantically constitutive for a natural kind concept. Instead, the reference of a natural kind concept is determined as a function of external facts by the externalistic essence belief. Accordingly, in order to share concepts, two speakers only need to share the externalistic essence belief. For example, if both the speakers disagreeing about the deep structure of gold believe that the ultimate criterion of being gold is to share some relevant deep structure with the actual gold samples, they manage to co-refer.

The case with biological species terms is somewhat different since, as noted in section (4.1), the reference of these concepts may not be totally pre-determined. This is because biological species members may not share any single empirically discoverable deep essence at all, but rather there are a range of properties that could be identified as the essence of the category. As LaPorte (2004) notes, essences of such categories may not be strictly *discovered*, but at least partly *defined*. It is, then, possible that along different theories of the essence of biological species, the meanings of the category terms do in fact change (as Ghiselin 1987 argues). This point about failure of concept sharing and stability may, however, apply mostly to experts, who may be committed to specific views about species essences. Lay speakers need not do so, and can share species concepts. As in the case of other natural kind concepts, they can even have diverging beliefs about the species members as long as they share the placeholder belief that the species members still share *some* deep structure which determines the category boundaries. (Note that concept sharing is not precluded even if species members don't share an essence in the strict sense—concept sharing is enabled by the *belief* that there is some reference-determining deep structure.) For example, one speaker may believe that tigers are spotted and red, another that they are tawny and striped, but they still manage to co-refer in virtue of the essentialist intention to refer with the term 'tiger' to creatures that share some relevant deep structure with the animals actually called 'tigers'.

The dual theory can salvage stability also in the case of artefact concepts, unlike some previous theories. For instance, a description or a prototype theory might identify the concept CHAIR with a set of perceptual features of chairs, thus entailing that two speakers having different perceptual representations of chairs fail to share the concept CHAIR. But this is counterintuitive. It is possible that one of the speakers has only met, say, office chairs during her life, while the other has only met bean bag chairs, but arguably they both mean the same thing by their concept CHAIR—given the opportunity, they would easily learn that there are other kinds of chairs than just office and beanbag chairs. The dual theory can easily account for this fact. It entails that the extension of the concept CHAIR is determined by the essence belief about chairs, and most of the perceptual representations of chairs are irrelevant for the concept's reference. Thus, if the two speakers both believe that chairs are artefacts intended for sitting, they manage to co-refer and share the concept. If they knew that both the bean bags and office chairs are intended for sitting, they'd agree in their judgements about their chairhood.

In the case of concepts which lack cores altogether we may have to allow for some extent of instability. For example, if one speaker believes that punishing a scapegoat is just and another that it is unjust, and neither would change her opinion upon reflection and discussion, then there seems to be no alternative but to grant that they mean slightly different things by their concepts of justice.

4.4. Philosophical applications: rigid predicates

Natural kind terms have had, at least since Kripke, a special position in the philosophy of language. Kripke suggests that natural kind terms are, like proper names, *rigid designators*. Rigidity is easily defined for proper names and other singular terms, but not so for kind terms. Let's start with the former.

Kripke's suggestion about rigidity is best understood in opposition to description theories of reference. A description theory about the proper name 'Aristotle' might hold that the name's reference is determined by the description *the last great philosopher of antiquity*. Kripke argues that this description cannot, however, determine the name's reference, as it might pick out different objects in different possible worlds: in some alternative worlds Aristotle died in infancy, in some others he never became interested in philosophy. Kripke suggests that proper names are rigid designators and refer to the same individual as they actually do in every possible world (in which that individual exists), and nothing else:

> (STR) If a singular term *t* refers to object x in a world w, then *t* refers to x in every possible worlds where x exists, and *t* doesn't refer to any other object in any possible world.

Thus, the name 'Aristotle' refers to the same actual individual even in possible worlds where he doesn't satisfy the descriptions associated with the term.

Arguably the most important theoretical consequence of proper name rigidity is that identity statements involving two actually co-referential rigid designators come out necessarily true. For instance, consider the following statement:

> (1) Hesperus is Phosphorus.

If the terms 'Hesperus' and 'Phosphorus' both actually refer to the planet Venus, then by (STR), the statement (1) is necessarily true: the two terms name the same object in every possible world. Crucially, despite being necessarily true, (1) is *a posteriori*: it was a remarkable empirical finding that Hesperus and Phosphorus are in fact one and the same planet.

Kripke suggests that natural kind terms are rigid just like proper names, and that rigidity makes certain identity statements between co-referring natural kind terms necessary. For instance, rigidity should make statements like 'water is H_2O', 'tigers are feline mammals', or 'gold is the element with the atomic number 79', necessary if actually true. However, thus far no one has come up with a generally accepted definition of rigidity; all proposals face serious problems (see Schwartz 2002; Soames 2003). Stephen Schwartz (2002) has even suggested that we reject the notion of rigidity for kind terms or predicates altogether. I will not go into examining the proposed accounts of rigidity here. Instead, I will argue that rigidity need not have any role whatsoever in explaining the necessity of statements of the above kind, but rather that their necessity is explained by the essence beliefs of the corresponding concepts. (A precautionary note: it is important not to confuse the present proposal with Michael Devitt's (2005) 'essentialist view' about rigid predicates, according to which a predicate is rigid if and only if it applies to an object necessarily, if at all.)

Identity statements versus essence statements

Possibly the most important theoretical task of the notion of rigidity for natural kind terms is to explain why identification sentences involving two co-referring natural kind terms are necessary. This is supposed to be done in somewhat the same manner as in the case of singular terms: just like the sentence 'Hesperus is Phosphorus' is necessary if actually true *solely in virtue of the rigidity of the terms in it*, we'd expect that so is a statement like 'water is H_2O' (see Kripke 1980, e.g. p. 143). There are, however, some important differences between the two statement types, which give us reasons to doubt the reasonableness of this endeavour.

Let us consider more closely the following two statements:

(1) Hesperus is Phosporus.
(2) Water is H_2O.

The most important discrepancy between (1) and (2) is that the former is an identity statement between two individuals, whereas the latter is a predicative statement, attributing water a certain scientific property (see e.g. LaPorte 2004, p. 36 ff.). This is easier to notice if we reformulate (2) as follows:

(2') Water consists of two hydrogen and one oxygen atoms.

Quite clearly, (2') and (1) are not analogous, and thus, we need not suppose that their identity be explained in the same way. This point is strengthened by the fact that the logical form of predicative statements concerning the deep structures of natural kinds are different from that of identity statements between singular terms. Whereas (1) can be easily formulated as a proper identity statement as in (1'), (2) is best captured in (2''):

(1') Hesperus = Phosphorus.
(2'') $(x)(\text{water}(x) \leftrightarrow H_2O(x))$

(On this point, see also Soames 2003, p. 430 ff.) Moreover, even if (2) could somehow be formulated as an identity, many other similar statements which also should come out necessary if actually true cannot. Consider, for instance, the statement 'cats are animals', which, if actually true, should come out necessarily true on the basis of rigidity (c.f. Kripke 1980, p. 125 – 126). This statement clearly cannot be captured as an identity, since even though all cats are actually animals, not all animals are cats. Instead, the form of the statement is a universally quantified material implication:

(3) $(x)(\mathrm{cat}(x) \rightarrow \mathrm{animal}(x))$

So, unlike (1), (2) and (3) cannot be called 'identity statements' at all—to separate the two, let us call the latter *essence statements*. Whereas rigidity is arguably needed to account for the necessity of identity statements like (1), the notion is *not needed at all* in explaining the necessity of essence statements. The latter's necessity stems simply from the fact that they describe the category essences; no theoretical role whatsoever is left for rigidity. For example, it is necessary that water is H_2O because the concept WATER has an essentialist core that makes it apply to a sample if and only if it shares some relevant deep structure with the actual water samples. And if this deep structure is H_2O, then WATER applies to a sample if and only if it is H_2O. Thus, the statement 'water is H_2O' comes out necessary.

The above discrepancies hold between natural kind terms and singular terms, but the same discrepancies hold even *within* a domain of terms. For instance, contrast (1) with the following:

(4) Hesperus is the celestial object with constitution C and origin O.

(Example from Haukioja 2006.) Whereas the necessity of (1) is guaranteed solely by the actual co-reference and rigidity of the terms, (4) is clearly an essence statement. If the essence of a celestial object is to have a certain constitution and origin, and if Hesperus actually has the constitution C and origin O, then it is necessary that (4) is true. The necessity of (4) is explained solely by the fact that it is a description of the essence of the planet, and rigidity has no role whatsoever in accounting for it. On the other hand, we can find examples of identity statements also in the case of natural kind terms. LaPorte (2004, p. 36) gives the following as an example:

(5) Brontosaurus is Apatosaurus.

The terms ‘Brontosaurus’ and ‘Apatosaurus’ were originally thought to refer to distinct species of dinosaurs, but it turned out that they denoted just the same species. The statement (5) is analogical to (1): in both cases we have two terms that refer to the same object or kind. Other examples of identity statements involving kind terms include ‘honeybee is Apis Mellifera’ (LaPorte 2000) and ‘soda is pop’ (LaPorte 2004). The necessity of each of these sentences is guaranteed by the fact that the terms in them denote the same species: ‘Apis Mellifera’ is just the Latin name for the honeybee, and ‘soda’ and ‘pop’ are names for the same beverage. I’m prepared to believe that *some* notion of kind term rigidity might explain the necessity of identity statements like these. A plausible account is given by LaPorte (2000; 2004), who holds that rigidity for kind terms means simply that the kind term refers to (objects of) the same kind in every possible world. However, no notion of kind term rigidity whatsoever is needed to explain the necessity of *essence statements* like (2) and (4). And these are the kind of statements that Kripke and others were originally most interested in discussions about rigidity, not identity statements like ‘Brontosaurus is Apatosaurus’ or ‘soda is pop’.

So, we might need some notion of kind term rigidity to account for identities involving kind terms, such as (5), but we don’t need rigidity to account for the necessity of essence statements like (2) or (3). But why cannot (2) be conceived as an identity statement? After all, the terms ‘H_2O’ and ‘water’ do co-refer, even though the former is a scientific description unlike the latter. Moreover, the logical form of (2) and the identity statement (5) is the same:

(2’’) $(x)(\text{water}(x) \leftrightarrow H_2O(x))$
(5’) $(x)(\text{Brontosaurus}(x) \leftrightarrow \text{Apatosaurus}(x))$

The answer is that, contrary to appearance, (2) and (5) are not strictly analogical. Whereas (5) is necessarily true solely in virtue of the co-reference of the terms in it, irrespectively of what we believe to be the essence of Brontosaurus, the truth of (2) depends on our essence beliefs about water. If a speaker believes that the essence of water is not its actual molecular structure, but instead some set of functional properties such as *potable, life-supporting, odourless, and colourless* (and so on), then the statement (2) is not necessary even if actually true—in the use of this speaker, ‘water’ could refer to something besides H_2O (such as XYZ). (5), on the other hand, is true irrespectively of what we believe to be the essence of Brontosaurus. If the essence is X, then *both* Brontosaurus *and* Apatosaurus are necessarily X; if the essence is Y, then *both* Brontosaurus *and* Apatosaurus are necessarily Y. In any case the essence is shared by

Brontosaurus and Apatosaurus, *simply because they are one and the same kind*. So, essence statements like (2) and identity statements like (5) are not analogical, despite their equivalent logical form.

In sum: rigidity explains the necessity of identity statements involving proper names, and some analogical notion of kind term rigidity (such as LaPorte's) might explain the necessity of identity statements involving kind terms. However, I distinguished from these two classes of statements *essence statements*, whose necessity is explained by essence beliefs about the categories—no role is left for kind term rigidity. Thus, the present proposal departs from the traditional accounts of kind term rigidity, in which the necessity of essence statements is supposed to be guaranteed by rigidity alone.

The claim that rigidity is not needed at all to account for the necessity of essence statements might be contested. It could be claimed that even though rigidity is not alone sufficient to explain the necessity of essence statements, it is nevertheless *necessary* in explaining it. This, the argument would go, is because without being able to refer to the same kind in all possible worlds, we couldn't say anything about a specific actual kind's essential features in the first place, as essential features can only be defined as features possessed by the kind in all possible worlds. For instance, without being able to refer just to *water* (instead of, say, any watery substance) in all possible worlds, we wouldn't be able to conclude that water is essentially H_2O, as the referent of the term 'water' would vary between different worlds.

In my opinion this is to put the cart before the horse. Whatever is the 'same' kind across possible worlds is *determined* by our essence beliefs about it. In other words, 'water' refers rigidly to H_2O just *because* we believe water's deep structure to be determined by the deep structure of the actual samples. We could have used 'water' non-essentialistically, irrespectively of whatever turns out to be the kind's actual deep structure. We could have taken water's essence to be some set of functional properties like being life-supporting, potable, refreshing, and so on. In this case the term would not have referred rigidly to H_2O, but to whatever is life-supporting, potable, refreshing, and so on. Accordingly, the statement 'water is H_2O' would not have been necessary even if actually true. What explains the necessity of the essence statements is our essence beliefs about the kind referred to—that the deep structure of the actual water samples is necessary and sufficient for being water, that being a mammal is part of the cat-essence, and so on.

All the theoretical work concerning essence statements expected of natural kind term rigidity is done by the fact that natural kind concepts have essentialist cores. Most importantly, this view makes essence statements necessary if actually true. For example, (2) is necessary if actually true because we intend to use the term ‘water’ to refer to whatever shares some relevant deep structure with the actual water samples. If that deep structure is actually H_2O, then water is necessarily H_2O. This also explains why essence statements are *a posteriori*: the reference-determining deep structure is external and empirically discoverable. In a related vein, the essentialist view bridges the so-called *modal gap* concerning *a posteriori* necessities. The question is how empirical findings about, say, water’s actual, contingent deep structure can tell us anything about the *necessary* properties of water. Externalistic essentialism grants that the empirical finding *alone* does not warrant any modal conclusions, but it entails that it can do so in conjunction with the essence belief that the actual deep structure of water is necessary for it.

Rigidity is typically supposed to apply mainly to natural kind terms, and not to any other terms. A problem pertaining to accounts of rigidity thus far is that they fail to meet this demand by either under- or overextending the notion of rigidity. For instance, LaPorte’s account of rigidity has been criticised for making such terms as ‘bachelor’ or ‘Mary’s favourite colour’ rigid (e.g. Schwartz 2002), whereas Devitt’s account has been criticised for *not* making terms such as ‘hot’, ‘loud’, and ‘red’ rigid (e.g. Soames 2003). The essentialist view has no trouble for accounting for these cases. Terms like ‘bachelor’ or ‘Mary’s favourite colour’ are not essentialist in that they don’t refer in virtue of any deep structure of the actual samples. For instance, even if all bachelors happened to share some genome, arguably we wouldn’t count possessing that genome as necessary for being a bachelor. On the other hand, terms like ‘hot’, ‘loud’, and ‘red’, which should come out rigid, are all essentialist. Hot objects are hot because they have high molecular motion, loud objects are loud because they emit high intensity sound waves, and red objects are red because they reflect white light on a certain wavelength (granted, in the case of ‘red’ it isn’t literally the *deep* structure of the actual samples that determines reference, but it is still some hidden, empirically discoverable essential property).

Finally, rigidity has been considered to have a role in distinguishing between terms that the description theory of reference applies to and those that it does not (Devitt 2005, p. 144). Externalistic essentialism can make such a distinction: since essentialist terms refer in virtue of some external deep structure of the actual samples, any description can fail to pick out that deep structure, and thus cannot determine reference. (An exception is, of course, if the descriptions are rigidified by stipulating that they apply solely to whatever *actually* satisfies them.)

In sum, since the externalistic essentialist account does all the theoretical work expected of kind term rigidity, I suggest that we follow Schwartz's (2002) recommendation and don't extend the notion of rigidity to kind terms at all. Another option would be to *define* kind term rigidity in terms of externalistic essentialism, but this may not be reasonable, since the notion would not be analogous with the notion of singular term rigidity.

5. Conclusions

At the outset I argued that concepts determine not only categorisation but also reference; thus, a complete theory of concepts should explain both categorisation and reference. I have argued that none of the present theories of concepts succeed in this task, as each of them falls prey to the ignorance and error problems. These problems suggest that we endorse a theory of concepts which is at least partially externalistic. Fodor's theory of concepts is of this kind, but I argued that it cannot do since it doesn't do any explanatory work in its own right, but rather has to rely on the notion of a sustaining mechanism. As an alternative, I proposed a dual theory of concepts, where (at least) natural kind concepts have distinct cores and identification procedures. The cores are externalistic essence beliefs, which determine reference as a function of what actually triggers the concept's identification procedure.

Noticing that concepts determine reference brings purely *a priori* accounts of reference into question. Philosophers typically study our semantic intuitions, which are mere effects of the real determinants of reference, that is, concepts. I argued that each of the existing accounts of reference are inadequate. The description theory falls prey to the ignorance and error problems, and conflicts with findings in experimental psychology. The causal theory, on the other hand, is undermined by the *qua*- and composition problems. Finally, neither account explains why certain terms refer externalistically. I argued that refining the theories of reference leads to an account practically identical to the dual theory I am defending.

The dual theory has some important advantages over the other theories of concepts and reference. The core determines reference and is the ultimate arbiter of categorisation; it enables concept sharing and stability, and may also warrant analytic inferences. The identification procedure, on the other hand, determines fallible categorisation judgements and warrants inductive inferences. Finally, I argued that the fact that natural kind concepts have externalistic essentialist cores may explain a phenomenon about natural kind term reference that has troubled philosophers since Kripke, namely the problem of rigid predicates.

6. Summary of the Articles

Article I examines the relationship between philosophical theories of reference and empirical psychology of concepts from a general perspective, focusing on natural kind terms. I argue that both description and causal theories of reference involve, or need to introduce, certain psychological elements. Refining the causal and description theories with the help of psychology of concepts ultimately leads to the dissolution of boundaries between the two kinds of theories. However, it is argued that even the resulting refined account of reference is incomplete in that it does not account for what makes natural kind terms rigid. In this article, by 'rigidity' I mean the phenomenon that, in every possible world, a natural kind term applies solely to instances of the same natural kind that it actually applies to, even if no one knows the deep structure of the kind. Thus, explaining what makes natural kind terms rigid basically comes down to explaining what makes them refer externalistically. In the conclusion of this article, I put forward the following sketch of a unified account of natural kind term reference:

> (T) A natural kind term *t* applies to an object x iff the DC_t maps x in the extension of *t* as a function of what actually triggers in the relevant speakers the RC_t (p. 168).

In (T), DC_t refers to what in Article I is titled the 'deep component' of a natural kind term *t*, and RC_t refers to what is titled its 'recognitional component'. In the article, I leave it open for future research what the two components of natural kind terms specifically are. Now I can conclude that the deep component of a natural kind term is the externalistic essence belief of the concept associated with the term. The recognitional component, in turn, is the identification procedure of the concept associated with the term, and it typically consists of prototypes or exemplars (and possibly also more theoretical beliefs). The externalistic essence belief determines that a natural kind concept applies solely to samples that share some relevant deep structure with the samples that actually trigger the concept's identification procedure.

Article II is best interpreted as elaborating on the nature of the 'deep component' of a natural kind concept; it is in this article that the notion of an externalistic essence belief is first introduced. I and my co-authors argue that, in contrast to what psychological essentialists maintain, psychological essentialism can be considered as modelling not only categorisation judgements or property inferences, but also reference. We put forward two possible semantic adaptations of psychological essentialism, a *hybrid* and a *strict externalistic account*. On the latter account, a natural kind concept C applies to an object if and only if the object possesses the real, external deep structure of the samples actually called C. On the hybrid account, a natural kind concept has two senses, of which in one, the *externalistic sense*, the concept applies to whatever possesses the real, external essence of the category, and in the other, *epistemic sense*, to whatever satisfies the identificatory knowledge the speaker associates with the term. For instance, if all cats turned out to be robots (Putnam's (1975) example) while they are believed to be mammals, strict externalism implies that cats have all the time existed, and that the term 'cat' has all along referred in virtue of the robot essence, whereas the hybrid view entails that cats did exist in the externalistic sense, in which 'cat' applied to robots, but that cats did *not* exist in the epistemic sense, where 'cat' applied to mammals. The strict externalistic view is motivated by Kripke's and Putnam's externalism, while the hybrid view is motivated by experimental results which suggest that natural kind terms may be ambiguous between two senses, externalistic and epistemic (Braisby, Franks, & Hampton 1996).

We examine an experimental study that claims to undermine externalism about natural kind concepts, due to Nick Braisby et al. (1996). We argue that due to some experimental and theoretical flaws, the study fails to assess externalism. We then go on to present two experiments of our own, the results of which support the strict externalistic reading of psychological essentialism. We conclude that natural kind concepts have an externalistic essentialist core which determines their reference.

Article III focuses on criticising Fodor's theory of concepts, according to which concepts are atomistic mental representations that refer in virtue of specific kind of causal relations between properties and our dispositions to token a concept. Fodor argues that whatever psychological or other mechanisms mediate between properties and our dispositions to token a concept do not semantically constitute the concept, or are necessary in order to refer with it. Instead, the causal relations alone, no matter how they are mediated, suffice to determine reference. Fodor's account thus denies the traditional psychological accounts of concepts any semantic force: they model merely the sustaining mechanisms, which are not necessary for reference, but rather only non-semantically associated with concepts.

I argue that Fodor is wrong in claiming that the mechanisms sustaining the reference determining property – concept tokening relations (or p/C-relations for short) are accidental for reference. I investigate various specific mechanisms that might actually sustain the p/C-relations, namely deferential intentions, essence beliefs, prototypes, and what Margolis (1998) titles *syndromes*. Syndromes are sets of represented category typical features; deferential intentions are dispositions to use a term in accordance with some experts; essence beliefs we are already familiar with. I argue that, of these mechanisms, syndromes, deferential intentions, and essence beliefs are strictly necessary for the reference determining p/C-relations they sustain. Moreover, even though prototypes are not strictly necessary for the p/C-relations they sustain, there are no non-trivial alternatives to them. I conclude that the sustaining mechanisms, as they are studied in experimental psychology, are not irrelevant for reference, but instead necessary.

References

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13*, 263 – 308.

Aydin A., & Pearce J. M. (1994). Prototype effects in categorization by pigeons. *Journal of Experimental Psychology. Animal Behaviour Processes, 20*, 264 – 277.

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition, 60*, 1 – 29.

Braisby, N., Franks. B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition, 59*, 247 – 274.

Brown, J. (1998). Natural Kind Terms And Recognitional Capacities. *Mind, 107*, 275 – 304.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Carey, S. (1999). Knowledge acquisition: enrichment or conceptual change. In E. Margolis & S. Laurence (Eds.), *Concepts. Core readings*. Cambridge, MA: MIT Press.

Connolly, A. C., Fodor, J. A., & Gleitman, L. R. (2007). Why stereotypes don't even make good defaults. *Cognition, 103*, 1 – 22.

Crane, T. (1991). All the difference in the world. *The Philosophical Quarterly, 41*, 1 – 25.

Davies, M. (2004). Reference, contingency, and the two-dimensional framework. *Philosophical Studies, 118*, 83 – 731.

Devitt, M., & Sterelny, K. (1999). *Language And Reality. An Introduction To The Philosophy Of Language*. Oxford: Blackwell Publishers Ltd.

Devitt, M. (2005). Rigid application. *Philosophical Studies, 125*, 139 – 165.

Dragoo, J. W., & Honeycutt R. L. (1997) Systematics of Mustelid-like Carnivores. *Journal of Mammalology, 78*, 426 – 443.

Ereshefsky, M. (2007). *Species*. Retrieved April 14, 2008, from Stanford Encyclopedia of Philosophy. Web site: http://plato.stanford.edu/ entries/ species

Fodor, J. A., & Lepore, E. (1996). The red herring and the pet fish: why concepts still can't be prototypes. *Cognition, 58*, 253 – 270.

Fodor, J. A. (1990). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. A. (1998). *Concepts. Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Gelman, S. A. (2003). *The Essential Child. Origins Of Essentialism In Everyday Thought*. New York: Oxford University Press.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences, 8*, 404 – 409.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: early understanding of the non-obvious. *Cognition, 38*, 213 – 244.

Ghiselin, M. (1987). Species concepts, individuality, and objectivity. *Biology and Philosophy, 2*, 127 – 134.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*, 137 – 165.

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology, 39*, 113 – 127.

Haukioja, J. (2006). Proto-rigidity. *Synthese, 150*, 155 – 169.

Hull, D. (1965). The effect of essentialism on taxonomy: two thousand years of stasis, *British Journal for the Philosophy of Science, 15*, 314 – 326.

Hull, C. L. (1920). Quantitative aspects of the evolution of concepts. *Psychological Monographs*, 28, 1 – 86.

Jackson F. (1998). Reference and description revisited. *Philosophical Perspectives, 12*, 201 – 218.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition, 57*, 129 – 191.

Keefe, R. (2000). *Theories of Vagueness*. Cambridge, UK: Cambridge UP.

Keil, F. C. (1989). *Concepts, Kinds and Cognitive Development*. Cambridge: MIT Press.

Kelemen, D., & Carey, S. (2007). The essence of artifacts: developing the design stance. In S. Laurence and E. Margolis (Eds.), *Creations of the Mind: Theories of Artifacts and Their Representation*. Oxford, UK: Oxford UP.

Kelemen, D. (1999). The scope of teleological thinking in preschool children. *Cognition, 70*, 241 – 272.

Kornblith, H. (2002). *Knowledge and Its Place in Nature*. Oxford: Oxford UP.

Kripke, S. A. (1980). *Naming And Necessity*. Cambridge, Massachusetts: Harvard University Press.

LaPorte, J. (2000). Rigidity and kind. *Philosophical Studies, 97*, 293 – 316.

LaPorte, J. (2004). *Natural Kinds and Conceptual Change*. Cambridge, UK: Cambridge UP.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In S. Laurence & E. Margolis (Eds.), *Concepts: Core Readings*. Cambridge, MA: MIT Press.)

Lewis, D. (1984). Putnam's paradox. *Australasian Journal of Philosophy, 62*, 221 – 236.

Margolis, E. (1998). How to Acquire a Concept. *Mind & Language, 13*, 347 – 369.

Margolis, E., & Laurence, S. (2006). *Concepts*. Retrieved May 9, 2008, from Stanford Encyclopedia of Philosophy. Web site: http://plato. stanford.edu/entries/concepts

Medin, D., & Ortony, A. (1989). Psychological Essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity And Analogical Reasoning*. Cambridge: Cambridge UP.

Murphy, G, L. (2004). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1999). The role of theories in conceptual coherence. In E. Margolis & S. Laurence (Eds.), *Concepts. Core readings*. Cambridge, MA: Cambridge UP.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition, 9*, 35 – 58.

Prinz, J. J. (2002). *Furnishing the Mind. Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.

Putnam, H. (1975). The meaning of 'meaning'. In H. Putnam, *Mind, Language and Reality. Philosophical Papers Volume 2*. Cambridge: Cambridge UP.

Quine, W.V.O. (1969). *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Ramsey, W. (1992). Prototypes and conceptual analysis. *Topoi: An International Review of Philosophy, 11*, 59 – 70.

Reimer, M. (2003). *Reference*. Retrieved April 10, 2008, from Stanford Encyclopedia of Philosophy. Web site: http://plato.stanford.edu/entries/ reference

Rey, G. (1983). Concepts and stereotypes. *Cognition, 15*, 237 – 262.

Rey, G. (1985). Concepts and conceptions: A reply to Smith, Medin and Rips. *Cognition, 19*, 297 – 303.

Rips, L. J. (2001). Necessity and natural categories. *Psychological bulletin, 127*, 827 – 852.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. T. E. Moore (Ed.) *Cognitive Development and the Acquisition of Language*. Oxford, UK: Academic Press.

Rosch, E. (1999). Principles of categorization. In E. Margolis & S. Laurence (Eds.), *Concepts. Core readings*. Cambridge, MA: Cambridge UP.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573 – 605.

Sandin, P. (2006). Has psychology debunked conceptual analysis? *Metaphilosophy, 37*, 26 – 33.

Schwartz, S. P. (2002). Kinds, general terms, and rigidity: a reply to Laporte. *Philosophical Studies, 109*, 265 – 277.

Segal, G. M. A. (2000). *A Slim Book about Narrow Content*. Cambridge MA: MIT Press.

Sloman S., & Malt B. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. *Language and Cognitive Processes, 18*, 563 – 582.

Smith, E. E., Medin, D. L., & Rips, L. J. (1984). A psychological approach to concepts: comments on Rey's 'Concepts and stereotypes'. *Cognition, 17*, 265 – 274.

Smith, E. E., & Medin, D. L. (1981). *Categories and Concepts*. Cambridge: Harvard University Press.

Smith, E., & Medin, D. (1999). The exemplar view. In E. Margolis & S. Laurence (Eds.), *Concepts. Core readings*. Cambridge, MA: Cambridge UP.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science: A Multidisciplinary Journal, 12*, 485 – 527.

Soames, S. (2003). *Philosophical Analysis in the Twentieth Century. Volume 2. The Age of Meaning*. Princeton: Princeton University Press.

Stanford, P. K., & Kitcher, P. (2000). Refining the causal theory of reference for natural kind terms. *Philosophical Studies, 97*, 99 – 129.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327 – 352.

Werner, C. W., & Rehkämper G. (2001). Categorization of multidimensional geometrical figures by chickens (*Gallus gallus* f. domestica): fit of basic assumptions from exemplar, feature and prototype theory. *Animal Cognition, 4*, 37 – 48.

Wisniewski, E. J., & Gentner, D. (1991). On the combinatorial semantics of noun pairs: minor and major adjustments to meaning. In G. B. Simpson (Ed.), *Understanding Word and Sentence*. Oxford, UK: North-Holland.

Wikforss, Å. (2008). Semantic externalism and psychological externalism. *Philosophy Compass, 3*, 158 – 181.

Wittgenstein, L. (1968). *Philosophical Investigations. 3rd edition*. Oxford, UK: Blackwell.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338 – 353.

Part II: Original Articles