This electronic thesis or dissertation has been downloaded from the King's Research Portal at https://kclpure.kcl.ac.uk/portal/

Concepts as Learning Devices
A Referential and Externalist Theory of Concepts

Jin, Woody

Awarding institution:
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT

Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. https://creativecommons.org/licenses/by-nc-nd/4.0/

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 13. Jan. 2025

Concepts as Learning Devices:
A Referential and Externalist Theory of Concepts

By

Zhengxi Jin

A Dissertation Submitted to
the Graduate Faculty of Arts & Humanities
in Partial Fulfilments of the Requirements for
the Degree of Master of Philosophical Studies

King's College London

September 2022

Acknowledgements

I am profoundly grateful to my main Supervisor David Papineau for his invaluable discussions, suggestions and support. This dissertation benefited enormously from his philosophical insights. And thank him for always being patient with my silly ideas and bad writings, and for his constant encouragement.

I am also deeply grateful to my term Supervisor James Stazicker who has read many of my early writings and given me detailed comments to help me clear my thoughts. I can always know how to make progress after discussions with him.

Thank my parents, partner and friends for always being there for me. They are the reason that keeps me going.

Abstract

This dissertation focuses on an often-overlooked function performed by concepts, which is to facilitate *learning*. This function is often implicit in theorists' glosses of concepts' basic jobs but has rarely been treated as a proper explanatory target. So I will try to make explicit this function of concepts in this dissertation and see what impact it will have on our current theories of concepts.

Part 1 aims to give an overview of the dissertation, explain my main motivation and clarify some terminologies/notions. I shall present five common theoretical desiderata on a theory of concepts to demonstrate that learning has not been treated as a proper explanatory target. I also display the thoughts that led me to see the importance of learning in concepts' functions. I finish Part 1 by drawing attention to the distinction between "concepts" and "conceptions".

Part 2 does a literature review of theories of concepts. I begin by grouping the theories into two camps: descriptivism plus internalism ("interiptivism") and referentialism plus externalism ("refernalism"). I then go through some of the main arguments that have been made by each camp based on the five common desiderata. Among these arguments, I shall highlight two criticisms of refernalism: that it is explanatorily "idle" and that it has trouble with explaining concept acquisition. I advertise that appealing to concepts' learning function can help resolve them.

Part 3 first attempts to make explicit my idea that concepts function to facilitate learning and then proceeds to elaborate on how it can help resolve the two criticisms respectively. I begin by specifying concepts' learning function into two functional roles -- the reidentification role and the signal-to-memory role -- by comparing them with what mental/object files do. I justify the ascription of the two roles by showing how the metaphysical structure of the world makes them apt for enhancing one's

adaptive ability. Taking the two functional roles of concepts as given, I move on to explain how they can allow us to demonstrate refernalism's explanatory value and thus resolve the explanatory idleness criticism. I use memory storage as a case study to show that refernalism is especially suitable for explaining how our mind uses concepts to build models of environmental entities. After that, I turn to deal with the criticism that refernalism has trouble explaining concept acquisition. I show that by reflecting on concepts' learning function we can see that concept acquisition is a form of "meta-learning". I argue that conceptualising concept acquisition as meta-learning can help clear away some traditional confusions regarding concept acquisition and show that refernalism can account for concept acquisition very well.

I conclude by summarising the core idea of this dissertation and raising some potentially related research questions.

Table of Contents

Part One: Preliminaries

CHAPTER 1 | INTRODUCTION AND CLARIFICATIONS

1.1 | An overview

Concepts provide the basis for thinking. Though theorists disagree on the nature of this basis and how it supports thinking, it is hardly controversial that concepts are of central interest to anyone who wants to understand how thinking, or cognition in general, works. As Fodor puts it, "...the heart of a cognitive science is its theory of concepts" (1998, p. vii). So this thesis aims to contribute to our studies of concepts by providing a new theoretical space for theories of concepts to develop. This space might be glossed in a slogan: "concepts are (mental) devices for learning". In this thesis, I will explain the idea behind the slogan and examine its implications on our current theories of concepts.

My inspiration comes from mainly Millikan's (1998, 2000, 2017) and Papineau's (2006) views on concepts. All I do is signify some of their points, which might have been underestimated in the literature, and explore to what degree we can extend

their important ideas. I will confine my exploration to only “substance” concepts in Millikan’s terms -- concepts of mainly individuals and real kinds. Some interesting, related issues like concepts of abstract entities (e.g. numbers) will have to wait for a larger project in the future.

At the heart of my thesis is the claim that concepts have been designed by natural selection to “lock onto” individuals and real kinds in the environment to facilitate the accumulation of knowledge (i.e. learning) about them for future use. More specifically, I will argue that individuals and real kinds’ being (relatively) enduring, stable and tight property clusters has substantively shaped the way the concept-producing mechanism works: for every individual/real kind it identifies, it produces a single concept to keep track of it and store information about it together in the same memory “chunk”. Had the individuals and real kinds in our environment been more unstable and unpredictable, the concept-producing mechanism would not have worked like this -- we probably would not even have had concepts in the familiar sense. Although I do hope my idea would look interesting to people regardless of their opinions on concepts, I will not remain neutral among different theories of concepts. Specifically, I will argue that if we think of concepts as learning devices in the above sense, we should favour what I call “referential and externalist theories of concepts”. So this thesis can also be seen as a defence and development of this family of theories.

Here is a summary of the structure of the thesis: In the rest of Part 1 (Chapter 1), I will present five common theoretical desiderata on a theory of concepts and suggest that learning should join them as well. I will also distinguish between the notions “concept” and “conceptions” and claim that this thesis shall focus on the former. Part 2 (Chapters 2-3) aims to selectively review our current theories of concepts. Chapter 2 divides them into two camps, “descriptive and internalist theories of concepts” (“interiptivism”) and “referential and externalist theories of concepts” (“refernalism”), and explains their differences. Chapter 3 goes through some of the key arguments given by the two camps for their own theories and against their opponents’, and suggests that adding the learning desideratum can generate new supporting arguments for refernalism. Part 3 (Chapters 4-6) aims to make explicit the learning desideratum and explain how it supports refernalism. Chapter 4 translates the

learning desideratum into two concrete functional roles -- roughly a reidentification role and a memory unification role -- by critically using the mental-file framework, and justifies the ascription of them by arguing that the roles are required for us to efficiently adapt to the world's "clumpy" nature. Chapter 5 argues that given the two functional roles of concepts, we should assign them with referential and externalist content because that will allow concepts to better explain their downstream system's successes. Chapter 6 responds to a lasting challenge to refernalism -- it cannot properly account for concept acquisition -- by exploiting the idea that concepts are learning devices. I shall argue that conceptualising concepts as a form of "meta-learning" can help resolve the challenge.

1.2 | Common theoretical desiderata on a theory of concepts: what is missing?

This section presents a list of common theoretical desiderata on a theory of concepts, extracted from mainly Margolis and Laurence's (1999) and Prinz's (2002) introductory sections. My purpose is to suggest that learning is a crucial desideratum missing from the list. These desiderata are important in that the shape of a theory of concepts will be heavily driven by what kind of desideratum, among others, the theorist thinks of as the most relevant and "weighty". Theorists design their theories with the aim of addressing certain desiderata and use the latter as their reasons to argue for their theories and against others'. I will elaborate on this point in Part 2, where I selectively review our current theories of concepts. But just for illustration, a prominent example is Fodor (1998), where he uses the compositionality of concepts, in the context of the Computational Representational Theory of Mind, as a desideratum to argue for his so-called "atomic" theory of concepts and against some popular theories of concepts in cognitive psychology. In short, adding a new desideratum to the list may change the field and give rise to many new theoretical possibilities. And that is what I will do in this thesis. Before that, let me briefly go through five common desiderata in the literature: cognitive significance, categorisation, acquisition, compositionality and meaning stability.

Cognitive significance. This desideratum is generated by the famous Frege's Puzzle (1893) about co-referential terms: if a term's reference exhausts its meaning, then

why can an identity statement between two co-referential terms, like “Mark Twain” and “Samuel Clemens”, still sometimes be informative or “cognitively significant”? A related phenomenon is that two co-referential terms sometimes cannot substitute each other in propositional attitude reports, like “John believes that Mark Twain is Mark Twain” and “John believes that Mark Twain is Samuel Clemens”, and also in belief-desire psychological explanations (cp. Braun 2000; Schneider 2005). Frege’s puzzle is related to concepts in that if we assume concepts to be what ground the meanings of linguistic terms and propositional attitudes, then a theory of concepts should be able to explain how differences in cognitive significance can arise for co-referential concepts.

Categorisation. Categorisation refers loosely to the capacity to make judgments about an object’s “member-hood”. Prinz (2002) clarifies that categorisation can be further divided into two different capacities: “category identification” and “category production”. The former refers to the capacity to classify objects, whereas the latter refers to the capacity to produce, often verbally, an object’s properties given its categorical membership. Categorisation has been the dominant desideratum within cognitive psychology. But there is a debate on whether philosophers of concepts should take this desideratum into consideration and if not, whether it means philosophers and psychologists are actually theorising over different theoretical entities (e.g. Machery 2009, 2010; Löhr 2020).

Acquisition. It is a natural phenomenon that we keep creating new concepts for things we encounter for the first time, like people we just meet. Therefore, a theory of concepts should be able to derive a plausible theory of concept acquisition. What is more, as Prinz (2002) suggests, given that innate knowledge has received increased acknowledgement since Chomsky (1972), it is not an unreasonable demand that a theory of concept acquisition should offer satisfactory explanations of both onto- and phylogenetically acquired concepts.

Compositionality. Concepts are compositional if they can be combined to generate more complex concepts/thoughts, whose contents are a function of the contents of the constituting concepts. Though compositionality is often directly cited as an independent desideratum, I think it comes with a large bundle of philosophical ideas

behind it championed mainly by Fodor (1975, 1998, 2008), which can hardly be fairly explained in a short section like this. Here is my best effort to summarise the relationship between compositionality and these ideas: a theory of concepts should be able to explain how concepts can be compositional because it will allow the theory to account for two prominent features of thinking -- productivity and systematicity. Productivity is the capacity to produce in principle an infinite number of thoughts, while systematicity is the capacity that if one can entertain a thought, then one can also entertain some other related (in terms of their contents) thoughts. A classic example of systematicity is that if one can entertain "John loves Mary", then one can also entertain "Mary loves John". Compositionality can explain productivity because it presents a way in which an in-principle infinite number of thoughts can be produced by manipulating a finite number of concepts. Compositionality can explain systematicity because it shows that systematic thoughts will share re-identifiable constituents and that a thinker can entertain systematic thoughts by just re-arranging the constituents (following some syntactic rules). In short, compositionality is a desideratum because theorists want a theory of concepts to be able to explain the productivity and systematicity of thoughts, and they assume that the compositionality of concepts is our best option to explain them.

Meaning stability. Meaning stability consists of two components: interpersonal stability and intrapersonal stability (Margolis and Laurence 1999). The former is also often called "meaning publicity". Meaning publicity requires a theory of concepts to be able to explain how different people can share or grasp the same concept, for it is assumed that only so can we further explain how people can successfully communicate and genuinely disagree with each other (Rey 1983) and how belief-desire psychological explanations are generalisable (Schneider 2005). Interpersonal stability requires a theory of concepts to be able to explain how a concept can retain its identity within a person despite her constantly changing beliefs (maybe also other attitudes). Yet this desideratum is often thought of as compromisable, especially by psychologists, for some of them hypothesise that concepts "never repeat themselves" due to some neuroscientific considerations (e.g. Barsalou 1999; Connell and Lynott 2014). But this thesis shall argue against this kind of view, for reasons to be explicated below.

Most of our current theories of concepts are designed to meet these five desiderata[1]. However, I think an important desideratum is missing from the list, which is learning. I suggest that a theory of concepts should also be able to explain how they can enable us to learn about things in the world. In this section, I will only "flag" this idea and briefly motivate it. I unpack the whole idea in Part 3. Although I said this desideratum is often missing in the literature, it has not gone completely unnoticed. For example, developmental psychologist Bloom claims:

> For the most part, we are lumpers. Our minds have evolved to put things into categories and to ignore or downplay what makes these things distinct… Why does the mind work this way? … Locke eventually comes to a better answer: A perfect memory, one that treats each experience as a distinct thing-in-itself, is useless. The whole point of storing the past is to make sense of the present and to plan for the future. Without categories, everything is perfectly different from everything else, and nothing can be generalized or learned. There is no savings, no information gained… We lump the world into categories so that we can *learn*. (Bloom 2005, p. 46; emphasis mine)

Philosopher Millikan also explicitly states:

> I will claim that the task of substance concepts is to enable us to reidentify substances through diverse media and under diverse conditions, and to enable us over time to *accumulate practical skills and theoretical knowledge about these substances and to use what we have learned*. (Millikan 2000, p. 2; emphasis mine)

And so does philosopher Papineau:

[1] But there are certainly more, such as "scope", i.e. a theory of concepts is supposed to explain many different kinds of concepts, including concepts of natural, social and theoretical entities (Prinz 2002). A desideratum that recently emerges is the so-called "polysemy" of word meaning/thoughts (Vicente 2018; Quilty-Dunn 2021).

> … the point of perceptual concepts is to *accumulate information about certain entities and make it available for future encounters*. (2006; emphasis mine)

To summarise their points, concepts play a crucial role in our everyday gaining of knowledge via making the world around us recognisable and thus learnable. Concepts serve as the basis for cognisers to stabilise their experiences of the world so that they can extract information from it more effectively. Learning should not be a trivial matter. And this also answers why I claimed the intrapersonal stability desideratum cannot be compromised -- concepts need to be intrapersonally stable to play the learning enabling role[2]. Yet common desiderata on a theory of concepts do not reflect a need to explain how concepts can serve as such devices for learning.

It might be argued that the categorisation desideratum already covers this learning matter. But as Millikan (2000) has emphasised, the capacity to recognise/reidentify is a different matter from categorising/classifying. One can take something to be the same over time without classifying it -- to reidentify x is to just *equivalise* x with some entity one encountered in the past. And it is recognition/reidentification that is more closely associated with learning since it makes knowledge accumulation and integration possible. Studies on categorisation rather focus more on how people *apply* what they have learned and on the structure of the learned, but they are different from the issue of how people are able to learn.

So this thesis will attempt to make explicit this learning desideratum and the explanatory goals it poses to a theory of concepts. But some clarifications and preparations need to be made before I proceed with the main task. An immediate, reflective question is: given these desiderata, are they all supposed to be met by a theory of *concepts* (as opposed to a theory of something else)?

[2] A natural reaction is that we only need *some aspects* of concepts to be stable but can tolerate other aspects' fluctuation. This complication invites a critical clarification -- see section 1.3.

1.3 | (Perhaps) not only a terminological distinction: “concepts” vs. “conceptions”

We may observe that the five desiderata sometimes push in opposite directions. For example, for concepts to explain how one can report features of members within a category, they had better be “bodies of information” -- structured mental entities that encode those features. If so, then probably no two people can share a concept because it is often assumed that what features a concept encodes will depend on the subject’s past experiences, but it is very unlikely that two people can have exactly the same past experiences. However, this seems to go against the meaning stability desideratum, which demands concepts to be sharable. Many theorists have also made similar observations. For example, Quilty-Dunn (2021, p. 159) uses “richness” and “compactness” respectively to describe the two contrasting desired features of concepts. Camp (2015, p. 591) similarly reports that concepts are often conceptualised differently as “associative networks” and “rule-governed atoms”.

An immediate question then is: how should we react? Radicalist like Machery (2009, 2010) takes the divergence between concepts’ desired features to show that there are two completely different kinds of theoretical entities both called “concept”, of interest to respectively philosophers and psychologists. He considers studies over the two to be mutually independent and suggests that theorists should not “cross the line”. Most other theorists hold a milder attitude (many of them can be found in the peer review of Machery 2010). They suggest that the divergence only indicates that there are two distinct sub-parts within a larger conceptual system supplementing each other and that we should use some finer-grained notions to distinguish the two. A common practice is to use “concepts” to refer to the entities that we think can be shared and “conceptions” to refer to the entities that underlie categorisation. Millikan illustrates the idea quite nicely with some examples:

> Having understood what the problem is, we can solve it by introducing a technical distinction. I will say that the child has “the same concept” as the chemist, namely, “the concept of sugar,” but that she has a very different “conception” of sugar than does the chemist. Similarly, Helen Keller had very many of the same concepts as you and I, but quite different

> conceptions of their objects. This fits with the ordinary way of speaking according to which people having very different information or beliefs about a thing have "different conceptions" of it… (2000, p. 11)

A good thing about this illustration is that it lets us grasp the difference between "concepts" and "conceptions" without building into any theoretical commitment. Theorists with different views on concepts can still use this distinction to clarify their own theories and discuss others'. A natural implication of this distinction is the division of the desiderata: cognitive significance and categorisation are to do with conceptions, whereas compositionality and meaning stability are to do with concepts.

Following this practice, I submit that my proposed desideratum of learning is to do with concepts rather than conceptions. More substantively, I shall treat concepts as a "functional kind" (Millikan 2017; Godman et al. 2020; Papineau MS), members within which share a number of properties all caused by the same selective pressure(s) (namely natural selection in the case of concepts)[3]. By saying that concepts are learning devices, I am claiming that concepts as a functional kind are bestowed by natural selection with the core function of enabling learning. More specifically, as I will argue, the "clumpy" nature of the world (Millikan, 2017) as the selective pressure has designed the concept-producing mechanism to make each individual concept function as a learning device dedicated to a "natural clump" (i.e. an individual/real kind) in the sense that the former tracks and creates memory chunk for the latter. And this function constitutes a theoretical desideratum that requires a theory of concepts to explain how it is realised (typically in the human perceptual-cognitive system, but I think it can also be reasonably asked about some other animals (Allen 1999) and even machines).

What about the relationship between concepts and conceptions? A commonly seen, relatively loose way of characterising the relationship is that a concept will "unify" a certain number of conceptions by assigning the latter with the same referential value

[3] Typical functional kinds are artifacts like tools. Their common selective pressures are approximately the designs of the tool makers. For example, hammers form a functional kind because they are all designed to perform the function of hammering, which can further explain why they all tend to have a certain shape, weight and so on. So I choose the wording "device" in my slogan to reflect my idea that concepts also form a functional kind like tools.

and restricting the epistemic conditions where the latter get properly applied, e.g. deciding when misclassifications happen (Margolis and Laurence 1999; Edwards 2009, 2010; Löhr 2020). A more vivid model depicting this picture is the so-called "mental file" theory (Murez and Recanati 2016; Goodman and Genone 2020): "mental files" are mental information repositories where conceptions are stored and concepts are "tags" on the files[4]. A more recent, computational model which resembles the mental file framework in some aspects is the "pointer" model (Quity-Dunn 2021), where a concept is a "pointer" that gives the subject access to a certain memory location where conceptions are stored (but cp. Eliasmith 2013[5]).

Given the above big picture (I assume it is coherent), this thesis will attempt to make explicit two specific relationships between concepts and conceptions: 1) concepts play a role in generating new conceptions (as a form of learning) by functioning as the "middle terms" in inferences; 2) concepts function as signals guiding the memory system to store conceptions about the same referent together for future co-retrieval, which is how conceptions get "unified". And I will show that concepts and conceptions interact in the above ways exactly because our world is clumpy -- they are how our cognition exploits the stability and "clusteredness" of natural clumps and their properties. But before I properly proceed with my project advertised here, I would like to take a few pages to briefly review our current theories of concepts. My purpose is to sketch a crude map of the field for me to later situate my project in it and better examine the project's implications on our current thinking about concepts.

[4] This is Fodor's (2008) interpretation, though; other theorists hold different views on the role of concepts in the mental file framework (e.g. Recanati 2012; Lee 2018), but they may not have the same concept-conception distinction in mind as we do here.

[5] He prefers a "compression" metaphor -- conceptions are compressed into concepts like WinRAR/Zip files in our computers.

Part Two: Reviewing current theories of concepts

I shall divide our current theories of concepts into two camps, which I dub the names “refernalism” and “interiptivism” respectively for convenience. “Refernalism” stands for referentialism plus semantic externalism, while “interiptivism” stands for descriptivism plus semantic internalism. So I basically divide the theories based on their claims about the semantic value of concepts and how such value gets determined. Chapter 2 shall unpack these claims and use concrete examples to illustrate the shapes of the two camps. Chapter 3 will selectively present some arguments for/against the two camps based on the five desiderata. I end this part by suggesting that the learning desideratum can give rise to new arguments for refernalism and also new responses to its criticisms, which will be the tasks of Parts 3 and 4.

CHAPTER 2 | "INTERIPTIVISM" vs. "REFERNALISM"

2.1 | Contrasting interiptivism with refernalism

I use this section to present the key contrasts between interiptivism and refernalism.

Both interiptivism and refernalism focus on the *intentionality*, as distinct from the *format*[6], of concepts. They theorise about the type of semantic content we should assign to concepts and the principle by which concepts get their contents. Individual theories within the same camp may still differ in many aspects, but they all share some important theoretical commitments. My aim here is to summarise the commitments of each camp in order to present their contrasts. As I suggested above, interiptivism and refernalism each are characterised by two main theses. Interiptivism commits to both the descriptivist thesis and the internalist thesis:

> *The descriptivist thesis*: a concept's content is composed of two elements: description and reference; for any concept, its reference is fixed by its description in the sense that it refers to whatever entity that satisfies the description.
>
> *The internalist thesis*: a concept's content supervenes on factors within the subject's head/skin[7], which means subjects who are internally identical must have their concepts' contents shared.

Refernalism, by contrast, commits to both the referentialist thesis and the externalist thesis:

---

[6] By "format", I mean the way a vehicle represents its content. There are many discussions of the format of concepts, such as whether it is modality-specific or amodal (e.g. Barsalou 1999), whether it is language-like or map-like (Camp 2007), and whether perception sometimes uses a conceptual format (e.g. Quilty-Dunn 2020a, 2020b). These interesting issues will not be touched in this essay.

[7] It has been a problem how to delineate the boundary of a subject's head/skin in a way that respects the intuitive contrast between semantic internalism and externalism (Gertler 2012). But here I will assume that my crude formulation suffices for our present purposes because those borderline, troublemaking cases will not affect our discussions here.

The referentialist thesis: a concept's content is just its reference.

The externalist thesis: a concept's content supervenes on factors external to the subject's head/skin, which means subjects who are internally identical can still differ in their concepts' contents.

To gloss them, interiptivism takes a concept's content to be primarily what the subject *knows*, whereas refernalism takes it to be primarily a relation the subject stands in to some entity in the world. I now use some specific theories to illustrate how the commitments can be fleshed out, starting with interiptivism.

2.2 | Interiptivism

The descriptive thesis (I use it interchangeably with "descriptivism") can find its root in the descriptive theory of reference originally proposed by Russell (1905) to explain the meaning of proper names and later natural kind names. To sketch his main idea: a name, N, refers to an object, O, if the user of N associates a set of descriptions, D, with N, and O satisfies D. A classic example is the name "Aristotle": when I use this name, it refers to Aristotle because I associate with it the descriptions "the person who was the student of Plato and teacher of Alexander", and Aristotle, among all other persons in the world, is the only one that satisfies those descriptions.

I take it that the so-called "conceptual role semantics" (Block 1986; Harman 1999) is a way to cash out descriptivism in the philosophy of mind. The main idea of this family of theories is that the content of a cognitive state (so it applies to concepts) is determined by its causal/inferential relations to some[8] other mental states. So we are not talking about literal, sentence-like descriptions but a summary of what the state "does". We could think of this summary as a unique position or node embedded in a complex web of mental states. And the content of a cognitive state will thus be determined by the structure of relations it bears to some other states on this web. For example, a cognitive state which bears causal/inferential relations to the states

[8] It is often assumed that not all causal/inferential relations, but only a selective number of them, will figure in the content of a cognitive state.

representing seas, rain, drinking and so on will be likely to represent water. To sum up, conceptual role semantics fleshes out descriptivism by specifying descriptions into a cognitive state's causal/inferential relations to some other states. Such theories are naturally coupled with semantic internalism (the internalist thesis) because relations between mental states are often taken to be paradigmatic factors within the head/skin. So in this sense, conceptual role semantics is a form of interiptivism.

Another family of theories in the philosophy of mind that may fall under interiptivism is the so-called "phenomenal intentionality theories" (Horgan and Tienson 2002; Kriegel 2013; Mendelovici 2018). Phenomenal intentionality theorists subscribe to the core thesis that the content of a mental state[9] is determined by the subject's phenomenal experience of this state. And the majority of these theorists hold that the thesis applies to both perceptual and cognitive states (so also concepts) -- they have sensory and cognitive phenomenology respectively[10]. So on a phenomenal intentionality theory, descriptions will be equated with phenomenal experiences. This makes phenomenal intentionality theories also submit to semantic internalism because phenomenal intentionality theorists typically assume that phenomenal properties supervene on intrinsic properties within the head, like neural properties[11], as distinct from environmental properties.

In cognitive psychology, most theories of concepts fall automatically under descriptivism because it is a consensus among cognitive psychologists that concepts are to be identified with structured knowledge representations, which can be assimilated into descriptions. They only disagree on what "type" of descriptions concepts are. Popular hypotheses include prototypes (e.g. Rosch and Mervis 1975), exemplars (e.g. Medin and Schaffer 1978) and theories (e.g. Murphy and Medin 1985). For a concept of x, a prototype is roughly the subject's statistical, probabilistic knowledge about the features that instances of x *tend* to have, but they may not be necessary or sufficient conditions for being x; an exemplar is roughly the subject's

[9] There are debates among these theorists on whether the thesis should be limited to only *conscious* states (e.g. Mendelovici 2018).
[10] Some theorists argue for a reduction from cognitive to sensory phenomenology (e.g. Prinz 2009).
[11] Though it has been questioned whether neural properties are thoroughly internal (Figdor 2009).

knowledge about a specific instance of x; a theory is roughly the subject's knowledge about the "deep" features of x, such as the underlying causal mechanisms and functions.

Compared to philosophers, psychologists are less interested in how concepts' contents are determined but more concerned with how people use their concepts to categorise and classify things. But I think it is implicit in their discourse that they assume what determines the reference of a concept, or the range of things a concept "covers", are to do with the subject's intentions or decisions[12], which are paradigmatic internal factors. For example, here is how Millikan (2000, p. 46; emphasis mine) summarises psychologists Ward and Becker's (1992) view: "Made explicit, the idea here seems to be that experience with a natural kind may inspire the category, but the category extent is determined by the thinker's potential *decisions* on exemplars". In this sense, psychologists not only have in mind a descriptivist assumption, but also an internalist one.

According to Edwards (2013), descriptivism (which I think applies to interiptivism as a whole given what I said) about concepts has been the dominant view in cognitive psychology and the philosophy of mind. Millikan similarly points out that in psychological literature, the descriptivist (again, also interiptivist) assumption about concepts has "managed to go completely unchallenged" (Millikan 2000, p. 43). Millikan thinks this is because psychological research on concepts has been focused solely on conceptions, rather than concepts themselves. As a result, the proponents of refernalism (e.g. Fodor 1975, 1998; Margolis and Laurence 1999; Millikan 2000; Millikan 2017; Schneider 2005; Edwards 2009, 2013) attempt to bring our focus back on concepts *per se* and the roles they play in our mental life.

2.3 | Refernalism

Contrasting interiptivists, refernalists typically hold that the semantic value of a concept is exhausted by its reference. Reference is a kind of relation held between

[12] There are exceptions. For example, Machery, who holds a psychologist understanding of concepts, seems to have in mind informational semantics (2010, p. 235).

representational items (so including concepts) and things in the world. We might picture this relation as an arrowed link pointing from representational items to things.

It is not the case that refernalists reject the existence of descriptions, however theorists may characterise them, which play important roles in various psychological processes. They only deny that we should identify concepts with these descriptions and that they directly[13] figure in the contents of concepts. But refernalists do accept that concepts and descriptions are connected in a non-semantically-constituting way, e.g. via association.

Refernalism about concepts is similarly inspired by discussions of linguistic semantics in the philosophy of language, specifically Kripke's (1980) and Putnam's (1975) criticisms of the descriptive and the internalist theses. I will present their criticisms in more detail in Chapter 3. What is important for now is that refernalists take seriously Kripke's and Putnam's criticisms and also their positive proposal -- reference is to be fixed by a certain kind of causal-historical relation between subjects and things in the world -- and apply it to theories of concepts.

Unlike description, reference is a much more straightforward notion. There is hardly any substantive debate among refernalists on how to characterise reference. Instead, refernalists are more interested in finding out the right type of causal-historical relation mentioned above. In other words, they want to explain how a concept come to have the referent it has. The motive behind this focus is probably that because of refernalists' commitment to the externalist thesis, they refrain from the talk of descriptions, intentions, or other intentional states, so they need to provide an alternative satisfactory story of how a referential relation can be established without being mediated by those states. This can be seen as belonging to the "content naturalising project" (e.g. Dretske, 1981; Millikan 1984; Papineau 1987; Fodor 1990). It is impossible to review the whole project here. So I will focus on Millikan's (2000, 2017) and Fodor's (1998) theories as two examples because they explicitly theorise over concepts as distinct from mental representation in general.

[13] I say "directly" because refernalists like Margolis and Laurence (1999) allow descriptions to indirectly fix concepts' reference, by establishing what they call "sustaining mechanisms" that sustain the causal relations required for fixing concepts' reference.

Fodor's theory of concepts is in service of his larger picture of how the mind works, namely the Representational Theory of Mind (RTM) (Fodor 1979) and the Language of Thought hypothesis (LOT) (Fodor 1975, 2008). He takes concepts to be unstructured mental symbols computed by the brain that works in a way pretty similar to how words work in natural languages: they can be systematically combined into more complex concepts and thoughts. Fodor considers RTM and LOT to be our best options in explaining the systematicity and productivity of thoughts. A key implication of RTM and LOT on concepts is that they need to be semantically arbitrary, simple and stable so that they can possess enough expressive power and recombinability across contexts. This makes Fodor come to think that a concept's content is just its reference. And he proposes what he calls the "asymmetric dependence theory" (Fodor 1990) to explain how concepts' contents are determined. The main idea of the theory is that a concept refers to x if 1) it will be reliably caused by x and 2) non-x will need to rely on 1) to cause the concept, but not vice versa. A classic toy example is that a concept HORSE[14] refers to horses but not, say, cows because it will be reliably caused by horses and when it is caused by cows, the causation will have to rely on the pre-existing causal relation between HORSE and horses, but not vice versa.

Millikan's (2000, 2017) theory of concepts shares with Fodor's commitment that a concept's content is just its reference, but driven by a quite different theoretical motivation. According to Millikan, concepts, or more precisely "substance concepts" in her terms, are used by the subject to reidentify natural clumps in the environment, namely individuals and real kinds, in order to learn about and thereby better act upon them. So for Millikan, a concept of x is roughly the subject's ability to reidentify x across contexts. She thinks such concepts must be thoroughly referential as opposed to descriptive because they are developed by the subject to "point" to those units in the environment, and such pointing relations can remain despite changes in the subject's beliefs (i.e. descriptions) about the units. Millikan also differs from Fodor in her metasemantic theory of concepts. She takes a so-called "teleosemantic" approach (Millikan 1984) to concepts' content determination: a concept refers to x if it

[14] Following the tradition, I use capital letters to denote concepts.

has the proper function of reidentifying x and enabling information accumulation about x. And a concept's "proper function" is roughly what it has been selected to do by evolution and/or learning.

I hope I have illustrated the rough shapes of the two camps in these two sections. To end this short review, I will now move on to go through some of the arguments for/against the two camps based on the five common desiderata.

CHAPTER 3 | EXAMINING THE THEORIES

As I mentioned in Section 1.2, different theories of concepts are often designed to address different theoretical desiderata. So a common strategy theorists of concepts use to defend their theories and criticise others' is to show how their theories can successfully meet some desiderata, whereas others' cannot. To avoid repetition, I will divide the following sections by theoretical desiderata and explain in what sense interiptivists and refernalists think these desiderata support their theories and undermine their opponents'. Discussions over these desiderata can take many back-and-forths between the two camps, and it is certainly impossible to exhaust all of them, so I will only focus on some prominent points made by each side.

3.1 | Cognitive significance, categorisation and reference

Cognitive significance is often taken to be on the side of interiptivists. Interiptivists argue that concepts must be descriptive, or "informationally rich", because we need concepts to explain various psychological phenomena, which arguably cannot be achieved if concepts only have referential content. I have mentioned Frege's Cases in Section 1.2, namely the identity statements between co-referring terms like "Mark Twain" and "Samuel Clemens" and the substitutions of such terms in propositional attitude reports. Interiptivists' point is that if we do not introduce another layer of content besides reference, be it Fregean sense or something more liberal like "cognitive content" (as distinct from "referential content", in Prinz's (2002) and Weiskopf's (2009) terms), we cannot explain where cognitive significance arises. And this point accordingly can be generalised to the explanations of inferences and

behaviours using belief-desire psychology. For example, we can infer that apples have cores even when we cannot see the cores. This is often explained by our concept of apple having CORE as part of its content. Similarly, Oedipus' marrying his mother despite desiring not to is often explained by his concept of Jocasta not having MOTHER in its content. By contrast, referential content seems to lack the resources to offer similar explanations.

Interiptivists who are psychologists often find refernalism unacceptable for a similar reason: referential content is simply impoverished in explaining psychological phenomena regarding categorisation. For example, it is shown that objects falling under the same category will not be treated equally by the subject -- some of them will be thought of as more typical than others. This is often called the "typicality effects" (e.g. Rosch and Mervis 1975). A hypothetical toy example is that we may judge that apples are "fruitier" than tomatoes. Some psychologists hypothesise that the typicality effects are to be explained by concepts' "weighted" way of encoding categories' features: some features will have a higher weight in deciding an object's membership, and the more weighty features an object possess, the more typical it will be judged to be. Another important phenomenon is the so-called "psychological essentialism" (Medin and Ortony 1989; Newman and Knobe 2019; Neufeld 2022): people's judgments about objects' "inside" or "essence" can override the perceived surface similarity when categorising. An example is that however a toy bear resembles a real bear, children will still not count it as an animal because it cannot move by itself (Gelman 2009). To account for these phenomena, psychologists thus hypothesise that concepts are prototypes, exemplars, or mental theories. But regardless of their differences in detail, psychologists' typical default assumption is that concepts must encode informationally rich, descriptive content to account for categorisation.

In short, interiptivists take descriptive content to be inevitable for concepts in explaining psychological phenomena and thus accuse refernalism of being explanatorily "idle" (Prinz 2005).

Refernalists' responses can be divided into mainly two parts: 1) the individuation of co-referring concepts and 2) the appeal to the semantically "detached" model. 1)

refers to refernalists' efforts in explaining how cognitive significance can arise for co-referring concepts by appealing to resources other than descriptive content. Some prominent approaches like Fodor's (1998) and Recanati's (2012) rely on the syntactic differences in co-referring concepts: For example, Fodor would say though TWAIN and CLEMENS are two co-referring concepts, they still differ in the "shape" of their vehicles, which thus makes the subject treat them as if they are two different concepts. Recanati would say TWAIN and CLEMENS are two distinct "mental files" information in which usually encodes different properties[15] and does not interact, so the subject will not realise that the two concepts are about the same person. 2) refers to refernalists' strategy to explain away the "idleness", which is to emphasise that refernalism can still appeal to the same resources as what is used by interiptivists in explaining psychological phenomena, namely descriptions or conceptions, however they are characterised (Margolis and Laurence 1999; Edwards, 2009, 2013). For example, APPLE, even if it only has referential content, can still explain the inference about cores by its *association* with the belief that "apples have cores". It is only that this belief does not determine the extension of APPLE, and so do other descriptions associated with it. In other words, despite acknowledging description's role in explaining psychological phenomena, refernalism still distances itself from interiptivism in its unnegotiable rejection of the internalist thesis.

As I mentioned, refernalists' main reason for the rejection comes from Kripke's (1980) and Putnam's (1975) criticisms of the descriptive and internalist theses in the philosophy of language. Let us call them the Kripke-Putnam-style arguments. But they accordingly apply to interiptivism about concepts as well (Rey, 1983; Margolis and Laurence 1999; Edwards, 2013). The Kripke-Putnam-style arguments can be divided into mainly three kinds: the problem of error, the problem of ignorance, and the modal or Twin Earth argument. I shall focus on the former two and skip the modality issue here.

The problem of error is roughly that people often associate with a concept descriptions that do not pick out the right referent. I here use an example discussed

[15] According to Recanati, even if the information is the same, as long as the files containing it are distinct, the subject still will not treat them as the same concept.

by Margolis and Laurence (1999) -- the concept HUMAN BEING -- to illustrate the problem. Some people, perhaps the majority of laypersons, associate with HUMAN BEING a key description “the creatures who have souls”. But if we agree, as science has told us, that souls do not exist, it seems that those people’s HUMAN BEING would end up referring to nothing. Or hypothetically, suppose some creatures, unbeknown to us, do possess souls (while human beings still do not). In that case, those people’s HUMAN BEING will end up referring to those mysterious creatures, not human beings. But this looks wrong. Even if some people associate the wrong description with their HUMAN BEING, still it should not affect the fact that those concepts refer to human beings.

The problem of ignorance is roughly that people can be ignorant about the key descriptions that will pick out the right referents for concepts. And this is particularly true of many scientific concepts. For example, if you ask me what distinguishes those inert gases, like Helium and Neon, I have to say I do not know. But this does not prevent me from having two concepts HELIUM and NEON which refer to Helium and Neon respectively. Yet interiptivism will have trouble explaining this fact because the poor descriptions associated with my HELIUM and NEON may in the end pick out nothing more specific than all the chemical elements. There is some empirical evidence arguably verifying the existence of the problem of error (Wellman and Gelman, 1992; Inagaki and Hatano, 2002; Keil and Kominsky, 2014, 2015). For example, in their experiments, Keil and Kominsky (2014) asked the subjects to first estimate how many features they can list that would distinguish the so-called “unknowns” (pairs of closely related categories which common adults are proved to know little about, such as ferret/weasel). Even though the subjects typically estimated that they can list a few, yet when actually required to spell out the features, they listed almost none. This evidence is particularly troublesome for interiptivists who think the nature of descriptions is people’s naïve “theories” of things because, as summarised by Keil and Kominsky, “In short, naïve theories are strikingly empty in the minds of most individuals, who are often unaware of their own theoretical shortcomings. In addition, when they do have fragments of theories, they are often full of unrecognized contradictions” (2015, p. 682).

In short, the problems of error and ignorance together suggest that descriptions have their inherent limitations in determining the referents of concepts due to the epistemic limitations of the subject. By contrast, refernalism's appealing to some causal-historical relation in doing the work allows a subject's referential ability to in some sense go beyond her epistemic limits, which is arguably more plausible in most cases. I will suggest later in Section 5.2 that this is partly what makes the learning desideratum vindicate refernalism.

To summarise this section, there seems to be a trade-off between explanatory power and correct reference-determining results, but refernalists' semantically "detached" model of concepts and descriptions seems to achieve a balance. Of course, interiptivists can dispute that it is unclear why reference-fixing matters at all in psychological explanations. I will come back to this issue in Section 3.4 in discussing meaning publicity and psychological generalisations (Schneider 2005; Edwards 2009) and sketch my own answer in Section 5.2.

3.2 | Acquisition

Interiptivism has a relatively straightforward model of concept acquisition: since concepts are descriptions or bodies of information, then acquiring a concept is just a matter of selectively putting together some descriptions to form a single concept. By contrast, concept acquisition has been traditionally assumed to be a problem for refernalism. I take it that the problem is not inherent in refernalism itself, but rather arises from Fodor's (1975, 1998) provocative argument, often called "Fodor's Puzzle of concept acquisition", that on (at least his) refernalist account, virtually all primitive (i.e. uncomposed) concepts are innate -- even including our concepts of cell phones, laptops, quantum physics and so on. Fodor's argument is roughly like this: (1) Concepts are innate if they are unlearned. (2) To successfully learn a (primitive) concept of x is to first have a dummy mental word, form a hypothesis that this word is about x, and finally confirm this hypothesis. (3) Forming a hypothesis that has x as part of its content presupposes that the subject has a (primitive) concept of x. (4) So learning a primitive concept is impossible because it leads to vicious circularity. (5) Given (1) and (4), then virtually all the primitive concepts we have are innate. As

Laurence and Margolis (2002) point out, this absurd conclusion has made many theorists take it to be a *reductio* against Fodor's LOT and refernalism. There has been a huge amount of reactions to Fodor's argument made by other theorists. I shall devote Chapter 6 to discussing this issue and my own solution. For now, it suffices for me to just flag that concept acquisition has been a problem for refernalism.

3.3 | Compositionality

The compositionality of concepts is a huge and vigorous research project, which has drawn attention from many disciplines (Frankland and Greene 2019). It is beyond my ability to give a review of the development of this project. So I will focus on some classic discussions that bear directly on this thesis' topic.

Compositionality has traditionally been thought of as a deep problem with interiptivism. And this should be attributed to Fodor's famous criticism of the compositionality of prototypes (presumably can be generalised to exemplars and theories), which is often known as the "pet fish" problem (Fodor 1998; Fodor and Lepore 2002): Roughly, the principle of compositionality requires that the content of a complex concept or thought be a function of its constituents. However, it certainly does not look like PET FISH, whose prototype is probably goldfish-like, is a function of PET and FISH, whose prototypes are probably dog/cat-like and carp-like respectively. Fodor takes this to be a critical argument against the psychological notion of concepts because it shows such a notion will have trouble satisfying the compositionality desideratum.

Many theorists have responded to Fodor by sketching models that attempt to show how prototypes can indeed be composed (e.g. Prinz 2002, 2012). These models may or may not be successful, but a crucial thing is that, as observed by Margolis and Laurence (1999), compositionality is not easily handled by refernalism, either: given that for refernalism, content is reduced to a certain kind of causal-historical relation, it is unclear in what sense these relations can be composed to generate another. For example, there is no direct explanation of what makes a causal-historical relation to the pet category and a causal-historical relation to the fish kind together form a

causal-historical relation to the pet fish category. Refernalists may again appeal to the associated descriptions to do some explanations. But this already shows that compositionality is a problem for everyone -- both sides need to construct satisfactory models to explain how compositionality is realised.

3.4 | Meaning stability

Meaning stability has been traditionally used by refernalists to criticise interiptivism: even regarding the same thing, two people can hold radically different beliefs (descriptions) about it due to their differences in experiences, mental pathways and so on. What is more, an individual's beliefs about the same thing can also radically vary over time. Therefore, it is argued that if concepts are descriptions, then we cannot explain how concepts can be interpersonally and intrapersonally stable, and that refernalism avoids this problem because causal-historically fixed reference will not be altered despite differences in the subject's beliefs.

But why does the meaning stability of concepts matter? An often-mentioned reason is that it serves as the basis for people to agree and disagree with each other (Margolis and Laurence 1999). Another reason is to do with the generalisability of psychological explanations -- people's uniformity in interacting with the same thing can be explained by their possessing the same concept (Schneider 2005).

Even though some interiptivists are convinced by the above reasons, they still challenge that refernalist implicitly assume that the stability of concepts is underwritten by content identity, which is unnecessary and unreasonable -- content similarity can suffice (e.g. Smith and Rips 1984; Barsalou 1999; Connell and Lynott 2014). A way of characterising content similarity is to appeal to the "overlapping" of descriptions. For example, two people can share their concepts of birds if their BIRDs both encode, among other features, BEAK, FLY, FEATHER and so on. But as Fodor and Lepore (1992) and Margolis and Laurence (1998) point out, this kind of strategy would have to presuppose the content identity of those individual descriptions, otherwise it is unclear in what sense descriptions can overlap. So as long as interiptivists grant the meaning stability desideratum, they will need to find a solution

that clearly captures what content similarity is while thoroughly explaining away content identity.

3.5 | Beyond the classic

So far, I have gone through some major points made by the two camps based on the five common theoretical desiderata. I advertised that I will be arguing for an extra desideratum on theories of concept, namely learning, for it is a primary function of concepts. It might be helpful to "forecast", in this transitional section, how this learning desideratum will bear on the discussions above. I shall argue that it supports refernalism for mainly two reasons. First, as we have seen in Section 3.1, refernalism has been accused of being idle in explaining various psychological phenomena. Refernalists' move is to appeal to descriptions with the emphasis that they do not determine concepts' extension. However, we still lack an explicit, positive account of what externally-fixed, referential content (of concepts) itself *does* explain. The learning desideratum, as I will show in Chapter 5, will illuminate that such content can positively explain the successes of the operations of the memory system. Second, Section 3.2 pointed out that refernalism has trouble accounting for concept acquisition due to Fodor's Puzzle. I will show in Chapter 6 that the learning desideratum can help resolve the puzzle by illuminating that concept acquisition is a form of "meta-learning", which does not lead to the vicious circularity argued by Fodor. Before I properly engage with these tasks, let me first make explicit the learning desideratum and what it demands from a theory of concepts.

Part 3 Concepts as learning devices

CHAPTER 4 | AN (EVOLUTIONARY) FUNCTIONAL ANALYSIS OF CONCEPTS

4.1 | Introduction

In this chapter, I attempt to answer two questions: 1) What exact explanatory goals does the learning desideratum impose on a theory of concepts? 2) What justifies the ascription of these goals to concepts? To put my cards on the table, in answering question 1), I submit that the learning desideratum poses two functional roles to concepts, which requires a theory of concepts to be able to explain: a) Concepts function to enable the reidentification of their referents and thereby retain “learning channels” for the learner to acquire information about the referents. b) Concepts function as signals instructing the memory system to unify information acquired through the same learning channel into a memory “chunk”, within which different pieces of information can be easily retrieved together. In answering question 2), I shall argue that the justification comes from mainly the metaphysical structure of the world we are in -- it is made of units that tend to retain their identity and properties

across scenes and over time, which Millikan (2017) dubs the name “natural clumps”. It is the nature of these clumps that have made concepts, via the design of natural selection, come to possess the two functional roles. More specifically, it is because these clumps and their properties tend to remain stable in the environment that we evolved concepts as “sameness-marker” of these units so that we can better learn about them and thereby exploit their stability.

Note that I do not mean that natural selection is like an intelligent being who is capable of mentalistic designing as we humans are. Rather, by “design”, I mean the process in which a trait, like an organ or mechanism, gets preserved within a species due to its fitness-enhancing effect under a certain environmental condition. My motivation for adopting such an evolutionary approach is my assumption that concepts form a functional kind (Section 1.3) like tools. When understanding tools, we need to know what kind of problem they are designed to solve, like hammering nails and opening jars. To study concepts, similarly, we need to know what kind of problem they are designed to solve by their designer, namely natural selection (Barret, 2015). And on my view, the problem specifically faced by concepts is exactly how to enable us to effectively learn about natural clumps.

Section 4.2 shall unpack the ideas behind the two functional roles by critically employing the so-called “mental-file framework”. Section 4.3 aims to justify the ascription of the roles by first introducing Millikan’s (2017) idea of natural clumps (Section 4.3.1), and arguing that without concepts playing the two functional roles, we would fail to effectively learn about these clumps (Section 4.3.2).

4.2 | Unpacking concepts’ functional roles

Having presented functional roles a) and b) in the introduction, I shall in this section make explicit the ideas behind them by critically using the mental-file framework. My reason for adopting this framework is mainly that it provides a useful model of how people can “trade on identity” (Goodman and Gray 2022) of things in their thoughts and perception, which is highly analogous to the roles played by concepts I have in mind. More on this will be explicated in Section 4.3.

In this section, Section 4.2.1 briefly introduces the mental-file framework. Section 4.2.2 clarifies functional roles a) and b) by comparing them with the roles played by some elements in the mental-file framework, namely “mental files”, “indices” and “ER relations”.

4.2.1 | Introducing the mental-file framework

I take the mental-file framework to include both the mental-file theory in philosophy (Recanati 2012; Goodman and Genone 2020) and its counterpart, the “object-file theory”, in visual science (Kahneman et al. 1992; Green and Quilty-Dunn 2021). My central claim will be that my proposed concepts’ functional roles, if using file-theoretic terms, are analogous to, but with some crucial differences from, the roles played by the so-called “visual indices” and “ER relations”. But before that, let me first briefly introduce the mental-file framework.

The mental-file theory aims at offering a solution to Frege’s Puzzle (Section 1.2) without introducing a further layer of meaning, like Fregean senses. The mental-file theory proposes that two co-referential terms can still differ in their cognitive significance if they are each grounded in a different “mental file”. A mental file is often characterised as a mental repository or store that contains pieces of information, often in the form of beliefs, about its referent. Yet the contents of these beliefs bear no influence on the file’s reference. Instead, a file’s reference is governed by the so-called Epistemically Rewarded (ER) relation held between the subject S and an object O in the world. An ER relation, roughly, is a relation through which S can acquire information about O. Paradigmatic types of ER relations, according to Recanati (2012), include perceptual relation, recognition, memory, and so on. On the mental-file theory, mental singular terms like “Superman” and “Clark Kent” are just two mental files containing presumably different information about the superhero, and their differences in cognitive significance are explained by their being two syntactically different thought vehicles. More slowly, on the mental-file theory, the cognitive significance of a mental file is explained by the *syntactic* or *vehicular property* of the file, not information contained in the file. So two files can still be of

different cognitive significance even if they contain exactly the same information. Such syntactic properties are what Recanati (2012) means by “non-descriptive modes of presentation”.

The mental-file theory has its counterpart in visual science, namely the object-file theory (e.g. Kahneman et al. 1992; Pylyshyn 2000). Object files are a type of visual representation posited to bridge the gap between early visual processing and conceptual thoughts, and also to explain how the visual system achieves coherent scene segmentation and object correspondence. The former requires explanations because neurobiological studies show that the processing of different types of visual information (e.g. motion vs. shape) is done by multiple specialised visual subsystems located in different brain areas -- it is a substantive question how the information they each process is integrated and synchronised to form coherent visual experiences (Treisman 1998). The latter requires explanation because due to the movement of both the objects and the eyes, visual inputs are not always continuous and stable, which requires the visual system to constantly figure out whether an earlier object at one location is the same as a later one at a different location (Richard et al. 2008). How does the object-file theory offer an answer to both problems? Accordingly, an object file is composed of a so-called “visual index” (Pylyshyn 2000) and an information store (i.e. the “file”). The index functions to track an object while it is moving or changing features. The information store functions to record the properties of the tracked object. Different types of visual information can thus be integrated if they are in the same information store, which in turn contributes to the segmenting of visual scenes into individual objects. Similarly, discontinuous or disrupted visual inputs will be treated as being about the same object if they are in the same information store, which thus solves the object correspondence problem.

A key feature of the object-file theory is that objects are also individuated “non-descriptively”, which makes it analogous to the mental-file theory: whether an object will be taken to be numerically the same depends on whether it is causally tracked by the same visual index, not on whether it satisfies some fixed descriptions in the information store (for some complication, see the next section). This feature is vividly captured by Kahneman et al.’s (1992) imaginary example, in which people watching an object flying toward them shout, “It’s a bird. It’s a plane. It’s Superman!”. As we

can see in this example, people can perceive the flying object as retaining its numerical identity while keeping assigning different categorical features to it. Such a non-descriptive way of individuation is key to our visual experience in that it allows us to (in good conditions) faithfully register an object's *changing of features*, as opposed to mistakenly perceiving it as *becoming another object* due to the feature changes.

Though the mental-file theory and the object-file theory are designed to address different phenomena, they posit a common structure (call it the "file structure") to explain our crucial ability to "trade on identity" (Murez and Recanati 2016; Goodman and Gray 2022), whether in thoughts or in vision. One trades on the identity between an object a and an object b if one directly infers from "a is F" and "b is G" to "something is F and G". I shall use the file structure to help clarify my proposed concepts' functional roles exactly because, as I will show in Section 4.3, learning about natural clumps requires the learner to have the ability to reliably trade on their identity. Before that, let me try to identify the core components in the file structure that enables trading on identity and explain how they relate to but differ from my proposed concepts' functional roles.

4.2.2 | Clarifying concepts' functional roles by comparing with the file structure

How does the file structure explain one's ability to trade on identity? In the case of object files, it is clear that the visual index's tracking plays an important role. Features of an object will be put into the same information store only if they are marked by the same visual index, which in turn serves as the basis for feature integration. Feature integration, as I see it, can be assimilated into trading on identity in the sense that the visual system is doing something like "inferring" from, say, "a is square" and "b is red" to "something is a red square". In short, in the case of object files, one's ability to trade on the identity between a and b most primarily depends on a and b being marked by the same visual index.

In the case of mental files, the job is presumably done by ER relations: two beliefs will be put into the same mental file and thus available for trading on identity only if they are acquired through the same ER relation. The standard account of ER relations, namely Recanati's (2012) so-called "indexical model", however, is controversial (I will explain why below): On this account, ER relations are individuated by both their type and spatiotemporal continuity. A visual tracking relation between a subject S and an object O is not the same ER relation as a recognitional relation between S and O (distinguished by type); also, even if both relations are visual tracking ones between the same subject and object, if the tracking occurs at different places and time (i.e. spatiotemporally discontinuous), they will not count as the same (distinguished by spatiotemporal continuity). As a result, the beliefs one gains about the same object via different ER relations will still be put into different files and thus do not allow for immediate trading on identity -- some further file merging/information transfer process is needed. In short, in the case of mental files, one's ability to trade on the identity between a and b most primarily depends on a and b being acquired through the same ER relation.

As we can see, the core components of the file structure that allow for trading on identity are (the tracking events of) visual indices and ER relations. Recall my proposed functional roles of concepts: a) Concepts function to enable the reidentification of their referents and thereby retain "learning channels" for the learner to acquire information about the referents. b) Concepts function as signals instructing the memory system to unify information acquired through the same learning channel into a memory "chunk", within which different pieces of information can be easily retrieved together. It is tempting to say that on this proposal, concepts are doing something similar to what visual indices/ER relations do. I agree they are similar, but not identical -- there are some crucial differences between them. And I plan to use these differences to make my proposal clearer and more explicit:

First, my learning channels are not identical to ER relations. As I mentioned, on Recanati's (2012) account, ER relations are individuated by their type and spatiotemporal continuity. By contrast, my learning channels will be individuated by their targets. For example, my perceptual tracking of a motorbike, my recognition of it, and even my reading descriptions about it, all count as the same learning channel

because they are about the same thing. Yet they count as three different ER relations. And this distinction automatically leads to the next:

Second, my memory chunk is not identical to a mental file. This difference follows naturally from the first: since ER relations' individuation determines mental files' individuation, then it follows that one can have multiple mental files of a single object. By contrast, on my proposal, memory chunks will be individuated by the targets of the learning channels coupled with them. Therefore, there will often only be one memory chunk for a single object. My main motivation for drawing such differences is that, as Papineau (2013) points out, Recanati's (2012) way of individuating ER relations, and thus mental files, leads to the unnecessary multiplication of mental files and also information processing steps. It is much tidier if we only posit one permanent file for one object. I will say more about this in Section 4.3.2. A minor motivation for me to draw the difference is that Recanati's mental files are mainly used to explain occurrent rational transitions (Goodman and Gray 2022), whereas I want to emphasise that concepts have a role to play in constructing our long-term memory. But note that it does not mean on my proposal, our memory structure will be perfectly tidy, i.e. be strictly "one-target-one-chunk". After all, Frege's Cases can show up, where we have more than one chunk for a target, if the concept *failed* to reidentify its target and thereby mistakenly led to the creation of new learning channels and memory chunks.

Third, my reidentification is not identical to the tracking event of a visual index. A few more words on how visual indices track: it is traditionally assumed that visual indices rely solely on spatiotemporal information like the continuity of a trajectory and moving speed to track objects, without using features recorded in the information store (Kahneman et al. 1992; Pylyshyn 2000). Yet recent studies (Richard et al. 2008; Hollingworth and Franconeri 2009; Moore et al. 2010; Hein and Moore 2012; Quilty-Dunn and Green forthcoming) suggest that the object-file system can flexibly employ various kinds of information including colour, shape and even category for visual indices' tracking, depending on the complexity of the scene and the availability of different types of information. I think reidentification is similar to the tracking event of a visual index in the sense that the former is also flexible and can rely on various types of information to do the job (Millikan 2000, 2017). What distinguishes the two

resides on two points: First, the timescale is different. The tracking event of a visual index is episodic, whereas the reidentifying of a target is long-term and discontinuous. We can reidentify a thing even if we have not seen it for years. Second, the tracking ability of visual indices presumably is unlearned. By contrast, though we can also reidentify some categories like *face* without learning, many of our reidentifying abilities are learned, such as the ability to reidentify a new friend.

To summarise, on my view, despite the differences in the standard of individuation, timescale and so on, concepts indeed function in a way similar to visual indices/ER relations. This should not be a coincidence -- though for different purposes, they all aim at faithfully registering the identity of entities in the external world and thereby guiding the internal encoding of information to reflect the identity in our judgments and reasoning. In other words, they all aim to allow the subject to accurately trade on identity. I now show that concepts need to possess the two functional roles exactly because learning about natural clumps requires the learner to be able to trade on their identity.

4.3 | Justifying the proposed functional roles of concepts

This section attempts to justify my ascription of functional roles a) and b) to concepts. My overall argument is like this: Concepts as a functional kind have been bestowed by natural selection with the function of enabling effective learning about natural clumps. And because of the stability of these clumps and their properties, our conceptual architecture has been designed to directly trade on the identity of members within a clump. Therefore, this very design of concepts poses functional roles a) and b) to concepts. Let me unfold my argument by first introducing the idea of natural clumps.

4.3.1 | Introducing the Clumpy World Thesis

Millikan (2017) proposes an account of the general metaphysical structure of the world, which she calls the “Clumpy World Thesis”. To illustrate the big picture, Millikan invites us to think of the world as a multidimensional graph, with each

dimension representing a property, and every physical object in the world as a dot in the graph, with its position specifying what exact properties it possesses. The Clumpy World Thesis is thus that these dots will form clearly distinct (though not perfectly demarcated) clusters. As Millikan puts it, “Much of the natural world is self-organized into discrete individuals and closely knit real kinds with reasonably wide gaps between them” (2017, p. 11).

More specifically, the hypothesis says that our world is made of naturally-occurred “clumps” -- clusters members within which all share a cluster of highly correlated properties *for a reason*. Many theorists have been trying to give an account of this reason when discussing the basis of so-called “natural kinds”. Among them, Boyd’s (1991) “homeostatic property cluster account” might be an influential one. But here I shall endorse Godman, Mallozzi and Papineau’s (2020) account, which claims that as opposed to homeostatic mechanisms, natural clump’s properties are clustered together because of a simpler causal structure: those properties share an underlying common cause, which Godman, Mallozzi and Papineau (2020) call “super-explanatory property”. For example, all samples of *gold* will share their values of density, melting point, electric conductivity and so on, due to their all possessing the same atomic structure. It is this underlying super-explanatory property that explains why all samples of *gold* share those surface-level properties.

According to Millikan (2017), natural clumps include both individuals and what she calls “real kinds”: An individual is a physical object, like a person, but also a property cluster if we think of the “time slices” or “time stages” of a physical object as the members of a kind. More importantly, a physical object in each of its time stages will share a number of enduring properties with that in other time stages for a reason: the time stages bear spatiotemporal continuity to each other and all share a common origin stage. In the case of persons, these enduring properties may include faces, fingerprints, the sound waves of voices and so on. Real kinds, depending on the reason their properties get clustered, can be further divided into eternal kinds, historical kinds, and functional kinds. In Godman, Mallozzi and Papineau’s (2020) terms, it can be said that real kinds are divided by the type of super-explanatory property they possess. An eternal kind, like a chemical element, has an *intrinsic* physical property, namely an atomic structure in the case of chemical elements, as

its super-explanatory property. A historical kind, like an animal taxon, has a common origin, namely a common ancestor in the case of animal taxa, as its super-explanatory property. A functional kind, like a tool, has the same selective pressure, namely a certain design in the case of tools, as its super-explanatory property.

There are two lines of empirical evidence suggesting that the clumpy nature of the world has shaped the human mind. I now briefly go through them.

One line of evidence is the finding of so-called “basic-level categories” (Rosch et al. 1976; Rosch and Mervis 1981). Basic-level categories are “... the level at which categories maximize within-category similarity relative to between-category similarity” (Rosch and Mervis 1981, p. 92), such as *dog* and *chair*. Basic-level categories are often characterised as the “middle level of abstraction” relative to subordinate categories like *Golden Retriever* and superordinate categories like *animal*. In Rosch et al’s (1976) series of experiments, it is shown that basic-level categories are the most fundamental environmental structures to the human perceptual-cognitive system: for example, these categories, compared to sub- and superordinate ones, are more quickly recognised and more spontaneously assigned with names by people. Studies also demonstrate that these effects are cross-cultural (Malt, 1995), so they are reasonably an evolutionary endowment. Therefore, it is reasonable to believe that natural clumps, through natural selection, have “tuned” the human perceptual-cognitive system in a way that the latter can easily capture the former even in the early days.

The other line of evidence is about psychological essentialism as mentioned in Section 3.1 (Gelman and Wellman 1991; Newman and Knobe 2019; Neufeld 2022), which is roughly the disposition to categorise things based on information about their “insides”, or “essences”, rather than mere perceptual similarities. To demonstrate it, in an experiment done by Gelman and Brenneman (2004), they asked the subjects, a number of 3-to-4-year-old children, to sort some pictures into two groups, the “zoo” group and the “store” group. The pictures were either of a real animal or its fabricated replication, so a pair of them would look highly similar to each other. The experiment result shows that the children were quite good at the task, with an accuracy rate higher than 67% (the correct way to sort the pictures is to put pictures of real animals

into the zoo group and pictures of fabricated replications into the store group). This suggests that even uneducated children are capable of categorising things based on information “deeper” than mere surface similarities. Like basic-level categories’ perceptual-cognitive effects, studies show that psychological essentialism is also a cross-cultural phenomenon (Neufeld 2022). According to Newman and Knobe (2019), psychological essentialism is had by both children and adults towards a large number of things, including individual human beings, natural kinds, race, gender and other social categories, like *scientist*. Newman and Knobe (ibid.) argue that there is a common element underlying all these phenomena, which is “the tendency to try to explain observable features in terms of a further unifying principle” (2019, p. 586). I think the existence of psychological essentialism, combined with the special effects of basic-level categories, together suggests humans are not only tuned to be sensitive to natural clumps, but also their common underlying structure -- a super-explanatory property causally explaining the clustering of a cluster of highly correlated properties.

It is reasonable to believe that humans have evolved the sensitivity to natural clumps and their underlying structure for the purpose of learning about them (i.e. adapting to them). According to Millikan (2000, 2017), natural clumps are the central targets of learning not only because they are the basic units of the world, but also because they grant high learning payoffs due to their possessing the so-called “rich inductive potential” (Gelman and Coley 1991), which, if exploited properly, can significantly aid one’s adaptation to the environment. For example, if we discover that a sample of gold melts at approximately 1000 ℃, then we can relatively safely infer that all samples of gold do due to *gold* being a natural clump, namely an eternal kind. Conversely, by identifying something as a sample of gold, we can relatively safely predict that it will possess a cluster of properties which other samples of gold have. What is more, the clumpy nature of the world allows people to form what Millikan (2000, 2017) calls “substance templates”. Substance templates are roughly people’s meta-knowledge about a clump, constituted by determinables that meaningfully characterise it. We can think of a substance template as a “multi-slot file” (Papineau 2006) with each slot specifying a question that needs to be asked about the clump in order to effectively learn about it. For example, a substance template of animals will probably contain slots that ask about the way of giving birth, whether having a spine,

the food type and so on. In short, the world's clumpy nature provides a "shortcut" for learning in the sense that a learner does not have to learn each object as if it is unique. Rather, by identifying clumps in the environment, we can quickly and effectively learn about things by just studying a few instances and then generalising the results to others. And these results can further be used to build and refine those substance templates, which in turn can facilitate our future studying of instances and thus eventually become a beneficial loop.

Given the significance and payoffs of learning about natural clumps, it is reasonable to hypothesise that concepts as a functional kind have thus been bestowed by natural selection with the function of locking onto those natural clumps and facilitating the accumulation of knowledge about them for future use. Indeed, Millikan explicitly states that the function of concepts "is to enable us to reidentify substances [natural clumps] through diverse media and under diverse conditions, and to enable us over time to accumulate practical skills and theoretical knowledge about these substances and to use what we have learned" (2000, p. 2).

Now I can drive the point home: Our visual system has evolved the object-file mechanism to keep track of physical objects in the environment because they and their properties are relatively "enduring" -- physical objects usually will not suddenly lose or change all of their properties and become something else. This makes trading on their identity through object files relatively reliable and beneficial. Real kinds are just like individual physical objects -- they together form natural clumps -- in the sense that real kinds are also relatively enduring in the environment and that their properties tend to form stable clusters due to their possessing underlying super-explanatory properties. Therefore, trading on real kinds' identity would be reliable and beneficial as well. In fact, I think to effectively learn about natural clumps, to exploit their rich inductive potential, exactly requires the learner to be able to trade on their identity efficiently. And this justifies the ascription of functional a) and b) to concepts because as devices designed to enable us to learn about natural clumps, they need to help us trade on their identity in an efficient way. Let me now elaborate on this point.

4.3.2 | Learning about natural clumps and trading on identity efficiently

Let us begin by considering a toy example: Suppose I live in the mountains where wolves are a threat. How am I supposed to learn about them? I assume we will have in mind scenarios like this: at t_1 I made an observation that wolves howled and formed the corresponding belief "wolves howl"; at t_2 I made an observation that wolves ate small animals and formed the corresponding belief "wolves are dangerous"; later at t_3 I integrated the two beliefs and came to have the knowledge "wolves howl and are dangerous", which made me decide to run away upon hearing howling (assume that wolves are the only things I know that howl). Call this the "naive scenario".

What job does my concept of wolves do in the above learning scenario? A natural answer is that it was tokened in both of my beliefs at t_1 and t_2 so that it could lead to my integration of them at t_3. In Millikan's terms, my concept of wolves functions as a "middle term" in my learning process, which she calls "mediate inference". Millikan goes to length to emphasise the importance of such middle terms in our mental life:

> Every mediate inference, every recognition of a contradiction, everything learned either from perception or inference and applied in action, every belief or behavior issuing from coordination among sensory modalities, for example, eye-hand coordination, even such subpersonal activities as the use of images from two eyes in depth perception, depends upon recognition of content sameness [by middle terms]. (2000, p. 143)

Someone might wonder at this point why I am belabouring these issues. Is not it just obvious that concepts are mental terms we use to think? To answer this question is to consider how things could work if concepts didn't have functional roles a) and b):

Suppose my concept of wolves failed to satisfy functional role a) -- it did not enable me to reidentify wolves over time. What would happen this time? Since I did not identify the wolves at t_2 with those at t_1, then very likely I would use two separate mental terms, "wolves$_1$" and "wolves$_2$" respectively, in the two beliefs, i.e. "wolves$_1$

howl" and "wolves$_2$ are dangerous". As a result, at t_3, I would have to make some efforts, like recalling the observations at t_1 and t_2 vividly (i.e. using my episodic memory), to come to recognise that wolves$_1$ and wolves$_2$ form a kind that allows for the generalisation of properties among its members. And only after that, I could integrate the two beliefs to acquire the eventual knowledge "wolves howl and are dangerous" which would then inform my practical decision. More slowly, my reasoning at t_3 is presumably a process like this:

(1) Heard something howling;
(2) (1) activated the belief "wolves$_1$ howl" and thereby led to my judging that very likely some wolves$_1$ are nearby (again assume that wolves$_1$ are the only things I know that howl);
(3) I searched my memory for related information and by recalling my observations at t_1 (when I formed the belief "wolves$_1$ howl") and t_2 (when I formed the belief "wolves$_2$ are dangerous"), I came to realise that wolves$_1$ and wolves$_2$ form a kind that allows for the generalisation of properties among its members;
(4) I formed an extra belief "wolves$_1$ equal wolves$_2$";
(5) By putting "wolves$_1$ howl", "wolves$_2$ are dangerous" and "wolves$_1$ equal wolves$_2$" together, I traded on the identity between wolves$_1$ and wolves$_2$ and thus eventually acquired the knowledge "wolves (=wolves$_1$=wolves$_2$) that are howling nearby are also dangerous;
(6) I decided to run away.

As we can see, steps (1) to (6) are much more "clumsy" than the learning process described in the naive scenario, where I had just one permanent mental term for wolves (namely my concept of wolves). And these steps are very much representative of the kind of picture Recanati's (2012) account implies: even regarding the same target, we can still have multiple mental files governed by different ER relations. When doing rational inferences about the target, we need to first merge those files so that the information in them can then be integrated for reaching further conclusions/decisions. It is true that steps (1) to (6) also show we can *in principle* learn and live in this way, yet it is very unlikely they represent how our mind in fact works. Such clumsy steps not only will slow us down when we need

to constantly learn about and adapt to our changing environment, but also violate the general principle by which our mind selects its computational methods -- if two methods can deliver similarly accurate results, it will prefer the one that costs fewer cognitive resources (Leider and Griffiths forthcoming). And it is exactly this point that brings out the core of my argument: why would we think that the result of the learning process in the naive scenario is as accurate as what is delivered by steps (1) to (6)? After all, the learning process in the naive scenario involves an *approximation* (i.e. the direct trading on the identity between wolves observed in different contexts) without any rational examination of it (i.e. the recalling and comparing of the two observations). The answer resides exactly in the clumpy nature of the world. It is because natural clumps are enduring and stable property clusters that we can relatively safely do approximations over them. If everything in this world is unique and will radically change its properties every second, then the kind of learning process represented by steps (1) to (6) will probably be much more reliable -- every approximation/equivalisation and generalisation requires careful examinations.

To put it in another way, why should the learning channels retained by concepts, and thus the memory chunks coupled with them, be individuated by their target *simpliciter* as opposed to finer-grained entities like target-plus-contexts (again, this idea seems to be implicit in Recanati's 2012 account)? Is not the information segmented by the latter standard more accurate and reliable in most cases? My answers are that first, given the clumpy nature of the world, having a permanent, single concept for one natural clump is often as reliable as having multiple concepts for each of the clump's "contextual slices". Natural clumps' being enduring and stable property clusters allows us to safely directly trade on their members' identity without worrying that the equivalence will suddenly collapse. More importantly, having a permanent, single concept for one natural clump proves to be much more efficient and resource-saving -- the learner can trade on the clump's members' identity without using further premises explicitly stating that the members are equivalent and that their properties are generalisable. The efficiency and economy of this conceptual architecture presumably can enhance a learner's ability to adapt to the environment and is thus favoured by natural selection. In short, concepts are not mental files in Recanati's (2012) sense. A concept is dedicated to reidentifying a

natural clump simpliciter, not its contextual slices, because the clumpy nature of the world has granted the former more advantages and thus “picked it out” via natural selection.

It might be questioned what significance functional role b) has, then. Functional role b) says that concepts function as signals instructing the memory system to unify information acquired through the same learning channel into a memory “chunk”, within which different pieces of information can be easily retrieved together. What is the importance of emphasising how concepts guide knowledge storage and retrieval? Is not this a redundant role given that a) seems to have satisfied all the needs? Let me devote the rest passages to explaining why I think this role still needs to be mentioned.

My reason is mainly that knowledge gained about the same natural clump being stored in a way that allows for co-retrieval is also crucial to the efficiency of our learning and decision-making. To illustrate the idea, let us think about how online search engines work. Search engines use the keywords we give them to look for fitted websites stored in their databases and then return the results to us. And this is often done within just a few milliseconds. The quickness depends (partly) on the databases being pre-sorted into sub-regions by, say, their genres, so the search engines do not have to visit every corner of their databases to look for fitted results but only the sub-region whose genre matches the keywords. I think our mind also needs a similar method to enhance our reasoning efficiency. As Radulescu et al. claim, to enhance our decision making, the memory system needs to “[organize] past experience in long-term memory in a way that facilitates retrieval of their summary statistics in the relevant circumstances”, and for this purpose, “rather than encoding each observation into memory, say, by order of appearance, it is useful to infer which future situations current information might be relevant to—that is, to categorize observations by the states to which they pertain—and update the summary statistics of that state in long-term memory with the current observation” (2021, p. 259). To illustrate, consider again our toy example: if my “wolves howl” and “wolves are dangerous” are not stored in the same chunk which allows easy co-retrieval, after I entertain one of them, I would have to effortfully search my memory for other related and potentially useful beliefs to decide what to do next. This lengthy process would

presumably slow down my reasoning and put me in danger because I could not reach the decision to run away in time. There is also some empirical evidence suggesting that our memory system indeed uses a strategy similar to the search engines'.

For example, Gershman et al.'s (2017) study suggests that our memory system sorts incoming sensory inputs by their so-called "latent causes", which are the inferred common causes behind those inputs. A toy example is that we store the representations of lightning and thunder together because the memory system infers that there is a latent cause in charge of both of them (though we do not necessarily need to know what this latent cause *is*). Storing memories by latent causes also means upon detecting the re-occurrence of a latent cause, memories stored "around" it will thus be retrieved together. As Gershman et al. summarise it, "Conditions that promote the retrieval of a memory are, according to this account, precisely the conditions that promote the inference that the same previously inferred latent cause is once again active. If no previously inferred latent cause adequately predicts the current sensory data, then a new memory is formed" (2017, p. 1). I think it is reasonable to hypothesise that concepts function as signals standing for some latent causes. Recall that concepts are used by us to lock onto natural clumps. And a common structure underlying natural clumps is that they each possess a super-explanatory property which is the common cause of those clusters of highly correlated properties they exhibit. Therefore, tracking a natural clump is indeed tracking a latent cause, which can be used to reliably predict the cluster of properties caused by it. To tidy up my points here: I just showed that to enhance our reasoning and decision-making efficiency, our memory system adopts a specific policy of organising memories, which is to sort them into different chunks by their inferred latent causes. Therefore, to enhance our reasoning and decision-making efficiency specifically regarding knowledge about the properties of natural clumps, the memory system needs signals standing for these properties' latent causes, namely the clumps themselves. So concepts as what have been designed to lock onto natural clumps naturally play this role. These together should justify my ascription of functional role b) to concepts.

In this chapter, I submitted that concepts should have the function of locking onto natural clumps and facilitating the accumulation of knowledge about them for future use and that this function constitutes a learning desideratum on theories of concepts. I then made explicit what this desideratum demands a theory of concepts to explain, which are the two functional roles of concepts: 1) Concepts function to enable the reidentification of their referents and thereby retain "learning channels" for the learner to acquire information about the referents. 2) Concepts function as signals instructing the memory system to unify information acquired through the same learning channel into a memory "chunk", within which different pieces of information can be easily retrieved together. I in the end justified my ascription of the two functional roles by showing that the clumpy nature of the world has made concepts playing the two roles more preferable in evolution because of their efficiency in guiding learning and thus decision-making. What I would like to do now is to show that the two functional roles vindicate refernalism about concepts.

CHAPTER 5 | LEARNING AND THE CONTENT OF CONCEPTS

This chapter takes on one of the tasks I flagged in Section 3.5, which is to answer the following question: granting that refernalism can respond to the explanatory idleness challenge by appealing to semantically-"detached" conceptions, what does externally-fixed referential content ("refernalist content") itself positively explain? I show that the answer will emerge if we think about concepts' content through the lens of learning.

Section 5.1 lays out my two assumptions about mental content in general -- the explanatory goal of positing mental content and the origin (metasemantics) of mental content -- and then explains how they apply to concepts' content. Section 5.2 responds to a potential challenge that my assumptions are ill-suited for concepts. I show that it can be resolved by drawing on concepts' status as learning devices. Section 5.3 uses memory storage as a case study to demonstrate the positive explanatory value of concepts' refernalist content.

5.1 | From mental content to concepts' content

How to think about mental content, or the intentionality of mind, has been one of the most central issues in the philosophy of mind. It is beyond the scope of this thesis to give a review of this big issue. But nonetheless, I would like to state in the beginning how I think of mental content in general and then apply it to concepts' content in particular. I will do so by answering two related questions: 1) What do we want to explain by positing mental content? 2) How does a contentful mental item get the content it has?

In responding to question 1), I shall assume that the main explanatory value of mental content resides in the explanation of behavioural *successes* (Papineau 1987, 1993; Shea 2018). And behavioural successes are to be understood as achieving certain distal effects that meet the agent's goal/desire through interactions with the environment. In other words, we assign contents to mental items for the purpose of explaining how using these items in psychological processing contributes to eventual behavioural successes. I think concepts' content is no exception. We assign concepts with contents for the purpose of explaining how using them contributes to our behavioural successes in interacting with their referents. Before I consider a critical complication, let me first answer question 2) quickly.

In responding to question 2), I shall assume that a so-called "teleosemantic" approach to (the metasemantics of) mental content (Millikan, 1984; Papineau, 1987; Shea, 2018). It is beyond the scope of this thesis to give a thorough review and defence of this approach. So here I will just state its main idea and briefly explain how it can be applied to concepts' content. The core idea of teleosemantics is that a contentful mental item ("mental representation") represents its "Normal condition" (Millikan, 1984), which is the type of condition it must correspond to in order for its downstream effects to result in behavioural successes. According to standard teleosemantics, a mental representation's Normal condition is fixed historically -- it is the type of condition under which the mental representation's downstream effects led to behavioural successes and thus got selected (i.e. preserved, either through natural selection or learning) in the past. In short, Normal conditions are

approximately previous success conditions. As we can see, my choice of teleosemantics is directly influenced by my assumption about the explanatory goal of mental content: we posit mental content to explain behavioural successes, so naturally, mental content arises from the condition in which behavioural successes obtain. As I have shown in Section 2.3, this general idea applies to concepts as well (Millikan 2000, 2017). A concept's content is also the condition under which the tokening of it led to behavioural successes. More specifically, as we have seen in the last chapter, a key role played by concepts is to allow us to reidentify their referents so that we can more efficiently integrate information about them and thus make good decisions. So we can say a concept refers to the entity the reidentification of which through the concept 1) led to behavioural successes like making good decisions and thereby 2) kept the concept in one's conceptual repertoire (as opposed to getting eliminated). For our convenience, from now on I will just say a concept refers to *what it has been selected to reidentify*.

Now we must deal with a complication: it seems that, unlike full-fledged beliefs, concepts *do not directly figure in the production of behavioural successes*. For example, we can say my belief "there is a glass of milk in the fridge" directly contributes to my success in getting that glass of milk for drinking, given that the belief is true. But apparently simply tokening my MILK concept would not have such an effect. We are not even sure what kind of behaviour is supposed to be the effects of MILK. It is even worse given my contention that concepts are learning devices: not everything we learn will have behavioural effects. To continue with our previous example of wolves, it could be the case that I simply gained the knowledge "wolves howl and are dangerous" but never used it in any decision-making. Artiga expresses a similar concern in assessing the plausibility of applying teleosemantics to cognitive representations (so presumably including concepts), which he calls the "problem of isolation": "In a nutshell, the worry is that most cognitive representations are *isolated* from the original input as well as from potential behaviors" (2016, p. 490; emphasis original). Does this mean that my assumptions about the explanatory goal of mental content and choice of teleosemantics are ill-suited for concepts? I now deal with this worry.

5.2 | Giving concepts their successes

To resolve the worry above, I think we must re-examine the scope of the notions of behaviour and success. To put my cards on the table, I suggest that concepts are in charge of producing the “mental behaviour” (as distinct from “bodily behaviour”) of building useful internal knowledge structures, or “models”, of external entities, and that the successes of such mental behaviour can come in various forms, such as bringing in new information to the knowledge structures and reducing their prediction errors.

My inspiration comes from Barto’s (2013) proposed way of modifying the standard reinforcement learning modelling to accommodate the so-called “intrinsic motivational system”. Here we need a quick characterisation of the notion of intrinsic motivation. The best way to do so might be to contrast it with “extrinsic motivation”, which we should be more familiar with: we can think of extrinsic motivations as corresponding to organisms’ fundamental biological needs, such as survival and reproduction. In this sense, behaviours like feeding, drinking and avoiding danger are all *extrinsically* motivated. By contrast, intrinsically motivated behaviours are not directly to do with the satisfaction of such biological needs. Rather, they are considered as being performed “for their own sake”. Examples include playing, exploration and so on (Baldassarre, 2011). Tooby and Cosmides (2001) hypothesise that intrinsically motivated behaviours evolved exactly for the purpose of building useful knowledge structures (namely learning) for future use. As they put it, “Thus, these [intrinsic] motivational guidance systems are vital components of developmental adaptations designed to help construct adaptive brain circuitry, and to furnish it with the information, procedures, and representations it needs to behave adaptively when called upon to do so” (Tooby and Cosmides, 2001, p. 16). In short, we can think of intrinsic motivation as the motivation to learn about the world around us to prepare ourselves for future tasks.

Now let us come back to Barto (2013). According to him, the standard reinforcement learning framework must include an environment as an essential element, which serves as the so-called “critic” that evaluates the behaviours performed by the agent

via its motivational system's signals. When speaking of extrinsically motivated behaviour, the environment is very evaluative because the satisfaction of biological needs clearly depends on whether the environment is cooperative or not, i.e. whether it offers the thing the agent needs. So an extrinsic motivational system can send signals in accordance with clear yes-or-no feedback given by the environment. However, when it comes to intrinsically motivated behaviour, it is unclear in what sense the environment is still evaluative: there is no yes-or-no feedback about, say, my merely playing with a toy. How is an intrinsic motivational system supposed to work? Barto's (2013) proposal is that we broaden our conception of the notion of environment and make it also cover what we may call an organism's "informational internal environment". This term is meant to refer to an organism's overall cognitive state, including its memories, knowledge and beliefs. As put by Barto,

> Novelty, surprise, incongruity, and other features that have been hypothesized to underlie intrinsic motivation all depend on *what the agent has already learned and experienced, that is, on its memories, beliefs, and internal knowledge state, all of which are components of the state of the organism's internal environment*. (2013, p. 36; emphasis mine)

So I take the idea to be that when it comes to intrinsically motivated behaviour, what plays the role of the evaluative critic is not an environment in the traditional sense -- it is the agent's informational internal environment, its memories, knowledge and beliefs, that do the job.

Following this line of thought, Barto (2013) also proposes to broaden the notion of "behaviour" or "action" in a similar way. As he puts it, "Similarly, an RL [reinforcement-learning] agent's "actions" are not necessarily like an animal or robot's overt motor actions; they can also be actions that affect the agent's internal environment… such as entertaining a thought or recalling a memory" (Barto, 2013, p. 22, p. 36). In this sense, all mental events that have an impact on one's memories, knowledge, beliefs and so on can in principle count as "mental behaviour".

Therefore, I think intrinsically motivated behaviour can come in two forms: "intrinsically motivated bodily behaviours" like playing and exploration, and

“intrinsically motivated mental behaviours” like building knowledge structures through memorising and reasoning (namely learning). This also naturally leads to a new way of thinking about behavioural successes: intrinsically motivated mental behaviours can also be success-apt because they can be assessed by whether they contribute to the building of knowledge structures. And the intrinsic motivational system has presumably several standards of measuring such contribution, such as whether a mental behaviour brings in novel/incongruent information to the knowledge structures or whether it reduces the latter’s prediction errors (Kaplan and Oudeyer 2007). According to Millikan (2017), the confirmation and non-contradiction of information can also count as such standards.

To summarise my point, in responding to worry that concepts’ content cannot be accommodated by my assumptions about mental content because concepts do not directly figure in producing behaviour (and thus behavioural successes), I show that the worry can be resolved by broadening the conceptions of behaviour and behavioural success to cover mental behaviour and its corresponding standards of successes. Following this approach, I submit that concepts participate in the (producing of) mental behaviour of building knowledge structures and that concepts result in mental behavioural successes when they meet one of the standards mentioned above. I shall now take this way of thinking about concepts and behavioural successes as given and proceed to discuss how refernalist content allows concepts to explain mental behavioural successes. I shall focus on a type of mental behaviour we have seen in the last chapter: the memory system’s creation of memory chunks.

5.3 | Explaining mental behavioural successes: memory storage as a case study

To reiterate, concepts participate in the mental behaviour of building useful internal knowledge structures or models. I take it that one of the main systems that is in charge of such behaviour is the memory system. Therefore, to see how refernalist content allows concepts to explain mental behavioural successes, we need to have an idea of what our memory system does and how concepts participate in it.

Aronowitz (2019) indeed recently argues that our memory system is a modelling system in the sense that it does not just store representations (like beliefs) passively or conservatively, it also actively forms these representations into patterns or networks that encode relations between representations. According to Aronowitz (2019), the memory system stores representations in this way for the purpose of making memory retrieval more useful (i.e. more relevant, more complete and so on) for problem-solving, given that "[t]he symmetry between how the memory is stored… and then consequently retrieved is striking [and that] effective retrieval involves tight cooperation between how information is stored, and how it is searched" (2019, p. 486). Therefore, I think the memory system stores information in a predictive, or forward-looking, and dynamic manner: it needs to consider the condition under which the information it stores will likely be retrieved (the predictive aspect), as well as adjust the way previous information is stored according to the results of each retrieval (the dynamic aspect). In short, by monitoring the results of memory retrieval, the memory system seeks to minimise the prediction errors produced by the memories (knowledge structures) it stores.

Here is my toy example to illustrate the idea: upon perceiving a bird, my memory system begins to store representations of the bird's properties, say, having a beak, a pair of wings and flying in the sky, extracted from the perception. My memory system predicts that the three properties will co-occur again, so puts them into the same memory chunk that boosts co-retrieval. Some days later I see another bird, it is flying in the sky spreading its wings. But it is at a distance so I cannot clearly observe its detailed looking. Due to the representations of the previous bird's properties being stored together, I then infer that the bird also has a beak, which proves to be correct when I see it more closely. We can imagine that the intrinsic motivational system takes the provement to be indicating that the mental behaviour of storing the three property representations together has met the standard of, say, contributing to reducing prediction errors, and thus releases reward signals to strengthen the connections between the representations.

A key issue I have not addressed is, how does the memory system make those predictions? Based on what does it predict that, say, some properties will co-occur

again in the future? I have shown in Section 4.3 that according to Gershman et al.'s (2017) study, the memory system sorts sensory data into memory chunks by their inferred *latent causes*. And a memory chunk will be retrieved when its corresponding latent cause is detected again. As Radulescu et al. put it, "… signals that disambiguate latent causes will be useful in organizing memory encoding and retrieval" (2021, p. 261). In Section 4.3 I hypothesised that concepts serve as such signals because they function to lock onto natural clumps, each of which is a stable latent cause of a cluster of highly correlated surface-level properties.

Now we can return to the main issue. I think concepts should be assigned with refernalist content exactly because this allows them to explain the successes of the memory system's operations in which they participate. To continue with our example of birds, why is my memory system's putting the representations of having a beak, a pair of wings and flying in the sky together successful (since it does lead to the reduction of prediction errors)? What explains this success? I think the answer resides in the content of the signal, namely the concept, my memory system uses to infer the latent cause. In this case, it is presumably my concept BIRD that serves as the signal. Assigning BIRD with a refernalist content -- a reference to the kind/category *bird* fixed by selection history (i.e. previous successes) -- can allow us to explain the success by showing that it is neither a coincidence nor an accident.

To elaborate, we can take it that the non-coincidency of the success is explained by the referential part of BIRD's refernalist content, while the non-accidency is explained by the teleosemantic part of it. The success is not a coincidence in the sense that there exists a correspondence between the knowledge structure and the environmental structure. More specifically, the knowledge structure is the memory chunk formed by the three representations of having a beak, a pair of wings and flying in the sky. The environmental structure is the kind/category *bird* as a property cluster, which indeed includes the properties of having a beak, a pair of wings and flying in the sky. Therefore, it is not just a coincidence that the co-retrieval of the three property representations proves to be useful in inferring about birds -- they are useful because they "map onto" (namely correspond to) the property-cluster structure of the kind/category *bird*. And BIRD's reference to *bird* explains this non-coincidental success in the sense that the reference is exactly what gives rise to the

correspondence/mapping. It is through using BIRD as a signal for the latent cause that my memory system comes to build up the correspondence/mapping between the knowledge structure and the environmental structure.

The success is not an accident in the sense that my memory system "trusts" BIRD as a signal for *bird for a reason*: BIRD has been selected to reidentify *bird*. Given this selection history, the reliability of my memory system's use of BIRD is secured because were BIRD not good at reidentifying *bird*, it would have been preserved within my conceptual repertoire. This will be clearer if we think of selection history also as an "improving history": BIRD has learned, from each of its previous successful reidentifications, how to better reidentify *bird* via a certain reinforcement-learning mechanism. I will come back to this issue in Chapter 6, where I argue that the acquisition of a concept is indeed itself a special form of learning.

To summarise, concepts' refernalist content serves to connect the operations of the memory system to the metaphysical structure of the world. And concepts' refernalist content can explain the successes of the memory system exactly because it shows that the memory system does not succeed just by luck. There is a *good reason* for it. Consistent with Aronowitz's (2019) argument, concepts' refernalist content makes it clearer that the memory system does not just function as a passive and conservative information store. Rather, it actively seeks to build knowledge structures that reflect or mimic the underlying metaphysical structures of the world -- that is why it uses concepts as its signals, like using a periscope, to monitor the world. By contrast, if we assign concepts with internally-fixed, descriptive content ("interiptivist content"), we arguably will lose the sense of how the memory can succeed, if not by luck. My reason is that as I have shown in Section 3.1, we could often be wrong about how the world is (e.g. the case of HUMAN BEING) or sometimes simply lack an idea of the joints of nature (e.g. the cases of inert gas and ferret/weasel). If using what we *know* as concepts' content, then we cannot explain why the memory system will use them as signals since such signals would presumably do poorly in guiding the memory system. In turn, we cannot explain why the memory system can succeed given the poor guidance it receives.

To end this chapter, I think there is a takeaway message delivered by the above discussions: Refernalist content is beyond one's epistemic limitations in the sense that it is not restricted by what one knows. Yet exactly because of this, assigning refernalist content to mental items like concepts, in many cases, is more "faithful" to how the mind works because such content can reveal to us that many brain systems have a future-orienting, forward-looking aspect -- they often seek to achieve a long-term goal through learning from feedback signals and thereby improving their operations. And refernalist content can make us appreciate this aspect by telling us what their goal is, namely what environmental structure they seek to better model.

CHAPTER 6 | ACQUIRING CONCEPTS AS META-LEARNING

This chapter attempts to take on the other task I flagged in Section 3.5: to respond to the criticism that refernalism has trouble accounting for concept acquisition, which has been presumably caused by Fodor's Puzzle (Fodor 1975, 1998). I shall argue that conceptualising concept acquisition as a form of "meta-learning" can help us resolve Fodor's Puzzle and thereby show that the criticism is implausible. Section 6.1 presents Fodor's Puzzle in more detail and shows how it can be generalised into a wider problem raised by (Cummins 1997). Section 6.2 starts by introducing Rupert's (2001) illuminating response to Cummins (ibid.) and then proceeds to explain how the features of innate concepts can become a breakthrough point. Section 6.3 uses innate concepts to introduce the notion of meta-learning and then shows how this notion can provide a unified conceptualisation of both innate and ontogenetically acquired concepts (I call the latter "novel concepts" hereafter). I also explain, in this section, how this conceptualisation can help us resolve Fodor's Puzzle. Section 6.4 responds to a pre-emptive challenge made by Fodor that concept acquisition other than hypothesis testing is unintelligible.

6.1 | A stimulating puzzle

Let us start by looking at Fodor's Puzzle more closely. Here is my reconstruction of the puzzle:

(P1) Concepts are either learned, brutally acquired (such as being hit in the head or through neural surgery) or innate;
(P2) Concepts can only be learned through hypothesis testing (and confirmation);
(P3) If a concept is learned through hypothesis testing, then it must be a structured/compositional one, otherwise it will run into vicious circularity;
(P4) So primitive, unstructured concepts are not learned;
(P5) Few, if not none, of our concepts are brutally acquired;
(C) Therefore, virtually all primitive concepts are innate.

(P3) might need some further elaboration here: what Fodor has in mind when talking about hypothesis testing learning are roughly scenarios in which a subject is first given a "dummy" word, say, "XYZ", which has no specific meaning to her. The subject is then asked to put some given objects or pictures into a group if they belong to, or instantiate, XYZ. After each trial, the subject will be told whether her grouping is correct or not. So for Fodor, learning a concept is like finding out what "XYZ" means through a number of trials. In each trial, the subject will group the objects or pictures following a "hypothesis" about the meaning of "XYZ", e.g. "'XYZ' means red". If she is told to have been wrong, in the next trial, she will attempt another hypothesis, and eventually find out the correct, confirmed one. At that time, the subject counts as having learned the concept of XYZ.

According to Fodor, the problem with this concept learning model is that to form hypotheses like "'XYZ' means red", the subject needs to possess the concept of redness so that it can constitute the hypothesis. This should not be surprising given Fodor's commitment to LOT. However, if so, then the subject does not really learn a new concept. More importantly, as Fodor (1975, 1998) argues, this also implies that a primitive concept can never be learned because to learn it, one must already possess it in one's conceptual repertoire so that one can use it to form the correct hypothesis. And this leads to vicious circularity. Structured/compositional concepts might avoid the problem because the subject only needs to find out the correct combination of two or more concepts. There is nothing wrong with already possessing the concepts in the first place.

Having explained the idea behind (P3), I think Fodor's Puzzle is relatively clearer now. Interestingly, this argument is called a "puzzle". The choice of wording suggests, which is in fact the case, that many theorists see this argument as urging them to figure out what is wrong with it, and to come up with the kind of concept acquisition account that does not fall prey to the puzzle. Among them, Margolis and Laurence's (1999, 2002, 2011; also Margolis 1998) series of responses are hard to ignore. I think a key contribution of their responses is that they point out that we cannot discuss concept acquisition in the void of a metasemantic theory of mental content. As they claim, "If possessing a concept means possessing a contentful representation, the issue of acquisition should be recast as the following question: Given the correct theory of mental content, how can one come to be in a state in which the conditions that the theory specifies obtain?" (Margolis and Laurence 2002, p. 35). This illuminating suggestion thus gives us a "handle" of the puzzle to initiate more detailed responses.

As I mentioned in Section 2.3, refernalists typically commit to causal-historical theories of mental content. Margolis and Laurence themselves, for example, take seriously Fodor's (1990) asymmetric dependence theory of mental content (again see Section 2.3). So they propose that to acquire a concept is thus to acquire the mechanism that can *sustain* the nomic, co-variational relation specified by Fodor's theory between a concept and its referent but does *not directly constitute* the concept. To give an example, one type of such sustaining mechanism proposed by them is what they call "syndrome-based mechanisms" (Margolis and Laurence, 2002), which are roughly one's knowledge about those typical surface-level properties of the referent, plus an essentialist disposition to override one's judgment when realising that the occurrence of these surface-level properties is not underlain by the right "essence" (Section 4.3.1). They think such syndrome-based mechanisms can sustain the required nomic, co-variational relations by "triggering" the corresponding concept upon detecting the presence of its associated typical surface-level properties. As we can see, acquiring a mechanism that can sustain the nomic, co-variational relation between a concept and its referent is in some sense similar to Millikan's (2000, 2017) idea of acquiring the ability to reidentify the referent (Section 2.3). There are certainly differences in Margolis and Laurence's and Millikan's thinking. But I take it that the differences will not matter to our discussion here. So I

will use “acquiring sustaining mechanisms” and “acquiring the ability to reidentify” interchangeably hereafter to avoid the disconnection between this chapter and the previous ones.

Margolis and Laurence’s proposal is surely quite illuminating. But I think it also invites a wider problem that refernalists must face: Fodor’s Puzzle, given Margolis and Laurence’s proposal, can arguably be generalised into the challenge that *how one can ever acquire the required sustaining mechanism given a causal-historical theory of mental content*. Let us now consider the problem.

Cummins (1997) argues that a causal-historical theory of mental content would have trouble explaining the acquisition of a contentful mental symbol. His argument is basically that on a causal-historical theory of mental content, it is impossible for the subject to establish the kind of relation required to be held between a mental symbol and its referent because doing so would lead to vicious circularity. Here is a re-construction of his argument for our present purposes:

> (P1) On a causal-historical theory of mental content, to acquire a mental symbol M with the content C is to put M in a certain co-variational relation to (instances of) C;
> (P2) To establish such a relation between M and (instances of) C requires the subject S to acquire an explicit mental theory of C which can instruct S to reliably detect (instances of) C;
> (P3) To acquire an explicit theory of C, S must possess a mental symbol with the content C so that she can construct the “axioms” that constitute the theory;
> (P4) However, this implies that for anyone to acquire a mental symbol, she must already possess that symbol in the first place, which is viciously circular.
> (C) So given a causal-historical theory of mental content, no mental symbol can be acquired.

(P2) may need some explanations. What does Cummins mean by “an explicit mental theory”? According to Cummins (1997), an explicit mental theory of C is composed

of mental sentences (like theoretical axioms) in the LOT sense that can "instruct" the subject to reliably detect (instances of) C in the environment. For example, for one to reliably detect, say, cats, one needs to have an explicit mental theory of cats looking like this: "(A) Cats have whiskers. (B) Cats have four legs. (C) Cats have fur" (Cummins, 1997, p. 537). Using this example, Cummins' argument is basically that one can never acquire a mental symbol for cats because to do so one must possess that symbol in the first place to compose (A), (B) and (C), which is viciously circular.

As we can see, Cummins' argument can be seen as a "follow-up" to Fodor's Puzzle, given Margolis and Laurence's proposal: granting that acquiring a concept is a matter of acquiring some mechanisms that can sustain the required causal-historical relation between the concept and its referent, still, *acquiring such mechanisms presupposes the concept itself*. For example, Cummins might argue that Margolis and Laurence's syndrome-based mechanisms are just mental theories in his sense so directly fall prey to his challenge. This follow-up requires further responses from refernalists.

I think many will agree that one of the key presumptions in Fodor's and Cummins' arguments[16] is that acquiring a concept or concept's sustaining mechanisms is a matter of manipulating LOT sentences about the concept's referent in mind, be they hypotheses or mental theories, which thus leads to the trouble-making result that the concept itself must pre-exist to constitute those LOT sentences that give birth to it. So to resolve Fodor's Puzzle, we have not done enough by pointing out that acquiring a concept is a matter of acquiring its sustaining mechanisms. More importantly, we need to show that acquiring a concept's sustaining mechanisms *does not necessarily involve manipulating LOT sentences*[17]. In the next section, I shall present Rupert's (2001) inspiring thoughts on this issue and explain how I think it can be extended.

[16] For convenience, from now on, I use "Fodor's Puzzle" to mean Fodor's own argument plus Cummins'.

[17] Another possible move is to argue that even if acquiring a concept's sustaining mechanisms requires manipulating LOT sentencesm, the concept itself need not be a constitutent. I will not consider this move in this thesis due to space limit.

6.2 | Innate concepts: a lesson to learn

In this section, I shall show how the features of innate concepts can illuminate a way out of Fodor's Puzzle. Section 6.2.1 presents Rupert's (2001) response to Cummins' (1997) challenge, which relies on the features of innate concepts. Section 6.2.2 explains how I understand innate concepts and how this understanding can lead to my solution to Fodor's Puzzle.

6.2.1 | Starting with Rupert's question

In his responses to Cummins' (1997) argument, Rupert (2001) first points out a point made by Cummins: if the co-variational relation between a symbol and its referent does not rely on explicit mental theories, then a causal-historical theory of mental content might still apply to it, and so-called "innate concepts" can count as such symbols because these concepts operate based on mental theories *implicitly* stored in one's inherited cognitive architecture (Cummins' own example is our geometrical concept SQUARE). It is crucial here how we should understand innateness and implicitness, especially given that both notions are controversial in the literature (e.g. Carruthers et al. 2005; Dienes and Perner, 1999). I will elaborate on my understanding of them below in Section 6.2.2. For now, we can take innateness to roughly mean being evolutionarily pre-wired and implicitness to mean not being formulated by LOT sentences.

To continue with Cummins' point, he, however, does not take the existence of innate concepts to undermine his argument because he supposes few of our concepts are innate, so causal-historical theories of mental content, even if they can account for innate concepts, would still be highly limited in their scope. What Rupert (2001) preciously questions in his response to Cummins is, therefore, that, "If innate, implicit theory can, in the case of 'square', play the detection-mediating role required [by a causal-historical theory of mental content], why should not implicit theory, innate or otherwise, play such a role in the case of 'cat'?" (2001, p. 503). I take Rupert to be questioning why implicit theories or knowledge cannot play the role of sustaining mechanisms (though Rupert does not use this notion himself) for *both* innate and

novel concepts. In other words, can there be such sustaining mechanisms that are both implicit and ontogenetically acquired?

Rupert thinks yes and proposes that such sustaining mechanisms are to be accounted for on the neural-physical level, as opposed to the LOT level. He raises two theoretical possibilities: neural selectionism (e.g. Edelman 1987) and neural constructivism (Quartz and Sejnowski 1997). Neural selectionism is roughly the idea that through our constant interactions with the world, those neural connections or "neuronal groups" proved to be useful will be selectively strengthened and thus preserved while the useless ones will be eliminated. Neural constructivism is roughly the idea that those useful brain regions or neural circuits will grow in terms of the number of their synapses. By raising these possibilities, Rupert means to show that we have the resources to account for sustaining mechanisms on this neural-physical level. For example, we can explain one's acquisition of the ability to stably reidentify cats by appealing to one's neural-architectural changes caused by neural selection or growth, such as the strengthening/growth of the receptors that are specialised for detecting the features of cats. Importantly, as Rupert emphasises it, an account on this level need not involve the LOT term CAT. Rather, Rupert proposes to think of these neural-architectural changes as the formation of the syntax or vehicle of CAT, which thus arguably avoids the kind of circularity raised by Cummins.

To summarise, Rupert, using innate concepts as a breakthrough point, challenges the plausibility of the presumption that acquiring a concept must consist in the manipulation of LOT sentences and shows that we can alternatively "locate" the account on a neural-physical level. I think Rupert's thoughts have provided us with an invaluable starting point to keep working on. What I shall do is connect his thoughts with my idea that concepts are learning devices and suggest that concept acquisition can be conceptualised as "meta-learning", which can give us the conceptual tool to resolve Fodor's Puzzle. Before that, let me first clarify my understanding of innate concepts to prepare us for the later discussion.

6.2.2 | Innate concepts as learning initiators

To begin with, I here endorse Margolis and Laurence's (2013) way of characterising innate concepts: they are the information or knowledge stored (through inheritance) in the domain-specific learning systems used for acquiring further concepts. To better explain their view, let me quickly introduce the big picture behind it. Margolis and Laurence's (ibid.) main focus is to clarify what they think is the core disagreement between so-called "Nativist" and "Empiricist". According to them, this disagreement is on "the character of the psychological systems that underlie the acquisition of psychological traits" (Margolis and Laurence, 2013, p. 695). As Margolis and Laurence see it, Empiricists hold that most of our psychological traits are acquired using a few domain-general learning systems with minimal information built in them, whereas Nativists hold that most of our psychological traits are acquired using a substantive number of domain-specific learning systems with relatively rich information built in them. In other words, Empiricists think we can just use one or a few universal learning mechanisms (e.g. associative learning) to effectively learn about all kinds of domains (namely, categories, like *artifact*, *animal* and so on) in the environment through our experiences, whereas Nativists think our learning of each domain is often guided by a specialised learning system containing information specifically about this domain. Margolis and Laurence (2013) call such specialised learning systems "Nativist Acquisition Base". And innate concepts are thus characterised by them as the components of this base.

I think Margolis and Laurence's (2013) way of characterising innate concepts makes it clearer what innate concepts are for -- they serve to initiate our learning about the world. And this idea is highly consistent with my contention that concepts are learning devices. The question is, then, *how* do they play the role of initiating learning?

My answer is that they do so by primarily being the "controllers" or "filters" of the perceptual system that select perceptual inputs specifically relevant for learning about different domains. Carey (2009) uses a similar notion called "dedicated input analysers". To explain, let us consider some concrete examples of innate concepts. Typical innate concepts include those of objecthood, number, animacy, agency and others' mental states, and so on (Cosmides and Tooby 1994; Carey, 2009; Gelman, 2009). To use animacy as an example, having an innate concept of animacy is to

have the perceptual system “pre-wired” to selectively extract relevant information about animacy from perceptual stimuli, such as whether an object “communicates with and responds in kind to like objects, moves by itself, and is made up of what we consider biological material” (Gelman, 2009, p. 227). Another example is infants’ innate concepts of human faces. According to Morton and Johnson (1991; found in Shea 2016), infants have their perceptual system pre-wired to be selectively attracted to objects in the environment exhibiting a three-dot-triangular configuration, which gives them the perceptual inputs they need to learn about human faces.

As we can see, innate concepts as input controllers/filters play the learning-initiating role in the sense that they save the learner from getting lost in the enormous amount of information presented in its experiences and help it focus on information that is most relevant for learning about a domain. To continue with the face example, by tracking the three-dot-triangular configurations in their environment, infants can resist the distraction of other environmental features and thus quickly start to extract information about *individual* faces, say, of their parents and thereby memorise them efficiently.

Innate concepts being input controllers/filters can also help resolve our previous unclarity about “implicitness”. Recall that in Rupert’s (2001) discussion of Cummins’ (1997) argument, it is taken that innate concepts have implicit sustaining mechanisms, which are in contrast with explicit mental theories in the form of LOT sentences. Now we can see in what sense innate concepts have *implicit* sustaining mechanisms: their co-variational relations to their specialised domains are sustained by the pre-wired architecture of the perceptual system. In other words, innate concepts’ sustaining mechanisms consist of those processing “parameters” or “assumptions” that the perceptual architecture adopts by default. These parameters, unlike LOT sentences, are not directly available to our thinking and are hard to report verbally, which make them implicit. Using Shea’s (2015) standard to distinguish implicit and explicit representations, these parameters are implicit representations (if they count as representations at all) in that except for making the perceptual system dispositionally respond to certain stimuli in certain ways, they cannot be directly used as inputs to other psychological processing. By contrast, mental theories in the form of LOT sentences are explicit representations because they can be used in an

indefinite number of, say, reasoning and inferencing processes. In short, we can roughly think of innate concepts' sustaining mechanisms as implicit *know-how* rather than explicit *know-what*.

Having clarified my understanding of innate concepts, it is time to return to our main issue: how can innate concepts help us resolve Fodor's Puzzle? To answer this question, we must consider how innate concepts are acquired. According to Cosmides and Tooby (1994), evolution history allows our ancestors' perceptual system to derive, through their interactions with different domains, the statistical regularities between the domains and some surface-level properties, such as between animacy and self-moving, between others' mental states and their gazes/facial expressions, and so on. These statistical regularities were encoded into genetic representations, passed on to us, and eventually become the processing parameters of our perceptual system today. In other words, evolution history is itself a learning history stretched over generations of a species. And acquiring (innate) concepts at this scale consists in the natural selection's "fine-tuning" of a species' perceptual system and "writing" the extracted statistical information into the system's processing parameters. Similar to what Rupert (2001) proposes, I think acquiring a novel concept of x can also be thought of as fine-tuning one's perceptual system within a lifetime (as opposed to across generations) to make it selectively sensitive to information specifically about x. I shall now show that the notion of "meta-learning" can provide us with a unified conceptualisation of the acquisition of both innate and novel concepts. I shall also argue that this notion can help resolve Fodor's Puzzle.

6.3 | Meta-learning as a unified conceptualisation of concept acquisition

Meta-learning is a notion that originated in psychology (Harlow 1949) and has recently got popular in reinforcement learning and neuroscientific research (Lake et al. 2017; Botvinick et al. 2019; Wang 2021). Meta-learning means "learning to learn", which can be characterised as deriving certain "biases" or "assumptions" from past "first-order" learning experiences to enhance the efficiency of future first-order learning.

The kind of meta-learning that is presumably most familiar to us is exactly the acquisition of innate concepts. To illustrate the idea, recall infants' tendency to track three-dot-triangular configurations in their environments. This tendency counts as meta-learned in that it is an assumption -- three-dot-triangular configurations indicate human faces -- derived from our ancestors' learning experiences (i.e. learning to recognise individual faces in their environments) through evolution and that this assumption can boost infants' first-order learning about individual faces. Not just so, all kinds of innate concepts, like so-called "intuitive physics" and "intuitive psychology" (Lake et al. 2017), are paradigmatic knowledge we meta-learned.

But meta-learning does not only occur at an evolutionary scale, it is also pervasive within one's lifetime. For example, we can quickly first-orderly learn to use a new operating system (say, Windows 11) on a computer because we have meta-learned many assumptions from our previous first-order learning experiences (e.g. learning how to use Windows XP, 7 and 10), which prevents us from learning everything from scratch. We can easily find the files we deleted in Recycle Bin the first time we use a new operating system because we have meta-learned the assumption that a bin icon very likely means it is a folder where we can find deleted files.

How does the notion of meta-learning bear on our discussion of concept acquisition? I suggest that we think of the acquisition of a concept, innate or novel, as a form of meta-learning. I have two reasons for my suggestion: 1) the notions of meta-learning and concept acquisition denote extremely similar processes which may even just be the two sides of the same coin; 2) introducing this notion can free us from the Fodorian sense of learning and thus give us a conceptual tool to resolve Fodor's Puzzle. Let me start by elaborating on reason 1).

6.3.1 | Concepts, meta-learning and attention-allocation policies

Consider how we learn to use a new cup. Cups can differ in many of their aspects, such as their materials (glass, porcelain or paper), their colours, their sizes, whether having handles, whether having words written on them and so on. However, we can effortlessly learn to use a new cup to contain liquid and drink, whatever its material,

colour and so on. Given the notion of meta-learning, it is easy to see this constitutes a simple case of meta-learning: we have meta-learned the assumption that a cup's material, colour and so on most of the time are irrelevant to the task of using it -- all we need to appreciate is the cup's structure, i.e. being a hollow cylinder with a bottom, which allows it to be used in the way we use other cups. How is such meta-learning realised?

According to Radulescu et al. (2021), we can think of such meta-learning as "learning to attend": we learn to simplify the representation of an object/scene by only allocating attention to some of its property dimensions and "ignoring" the rest. Learning to attend is crucial to our learning specific tasks because, as the example of cups demonstrates, an object/scene can exhibit an enormous amount of properties but not all of them are useful. If the learner needs to process all of them, fast learning would be impossible. Therefore, the learner must adopt what I call an "attention-allocation policy" specific to an object/scene to help it focus on the properties that really matter, and this policy can be derived, namely meta-learned, from the learner's past learning experiences with similar objects/scenes. In the case of learning to use a new cup, the attention-allocation policy is extremely simple: all we need to focus on is the cup's structure. For now, let me be vague about the nature of attention-allocation policies and continue to present the resemblance between learning to attend and concept acquisition.

How we can acquire a concept of cups? I have, following Margolis and Laurence (1999, 2002, 2011), assumed that acquiring a concept is a matter of acquiring a mechanism that can sustain the co-variational relation between the concept and its referent, and I have taken it to be the same thing as acquiring the ability to reidentify the referent. Therefore, to acquire a concept of cups is basically to acquire the ability to reidentify instances of the (functional) kind *cup*. As I showed above, cups can differ in many of their aspects. So the ability to reidentify instances of *cup* will necessarily involve doing *approximation* (I also mentioned a similar point in Section 4.3.2 when discussing trading on the identity between different members of wolves): different instances of *cup* will be reidentified (or more precisely "co-identified") because we simply only take into consideration their sole commonality, namely their structures (and thus the same function they can serve) and ignore their differences

regarding other property dimensions. Not just with cups, I take it that all cases of reidentification involve a certain amount of approximation. For example, even the reidentification of an individual, say, a friend at different times will require us to ignore her changes in her hairstyle, the clothes she wears and so on.

My point is, then, that *the approximation in reidentification is realised similarly by adopting a certain attention-allocation policy*. We reidentify different instances of *cup* because we only pay attention to the property dimension of structure. We reidentify different time-slices of a friend because we only pay attention to her (approximately) invariant properties like her face and sound. Therefore, I think acquiring the ability to reidentify x (i.e. acquiring a concept of x) is, to a large extent, a matter of acquiring an attention-allocation policy specific to x. Also, I think acquiring an innate concept, though on a different scale, also consists in acquiring a specific attention-allocation policy. This should not be a surprise given my characterising innate concepts as perceptual controllers/filters -- having an innate concept of a domain is for the perceptual system to possess certain processing parameters which select inputs relevant for learning about that domain. So naturally, acquiring an innate concept is also a matter of deriving a specific attention-allocation policy, namely the policy of directing attention to perceptual inputs in the environment relevant for learning about a domain. It is just that such acquisition is conducted on an evolutionary scale, not within a lifetime.

In short, there is an overlap between meta-learning and concept acquisition: they both require the learner to learn to attend. I think this is not a coincidence. As I have shown in Chapter 4, we acquire a concept of a target to more efficiently learn about it (via trading on identity and memory organisation). Also, in Section 6.2.2, I showed that we possess innate concepts because they can allow us to efficiently initiate our learning about the world. Similarly, we meta-learn from our past experiences also to more efficiently (first-orderly) learn in the future. As we can see, it is not a coincidence that meta-learning and concept acquisition are highly similar (if not the two sides of the same coin) because they both function to promote learning. This is why I think concept acquisition can be conceptualised just as a form of meta-learning. And a theoretical virtue of this conceptualisation is, as I mentioned in the title of Section 6.3, that it helps *unify* the acquisition of both innate and novel

concepts -- they both consist in the acquisition of attention-allocation policies, only at different scales. According to Prinz (2002; see Section 1.2), this unification meets a desideratum specific to concept acquisition.

My tasks in this section have not ended here. I have been leaving the nature of attention-allocation policies, and also how we can acquire them, vague. But recall that the central task we must take on in response to Fodor's Puzzle is to show that acquiring a concept does not necessarily involve manipulating LOT sentences. Therefore, it is my duty here to show that acquiring attention-allocation policies -- as the way I suggested concepts are acquired -- does not necessarily involve manipulating LOT sentences. I shall now proceed to this task. Note that I am not attempting to give a full-fledged account of how we learn to adjust our attention allocation, which presumably requires much more work. What I aim to do is only to show that, with some suggestive evidence, we can learn to attend without manipulating LOT sentences.

To begin with, let me first clarify what I mean by "attention" since it is notoriously ambiguous. I use this term to denote the kind of "processing resource" or "commodity" at the subpersonal level (Allport 2011, p. 25). This way of using "attention" suggests that attention is limited and can only be allocated to a limited number of stimuli. Consequently, we must possess some corresponding mechanism or system that is in control of the allocation of attentional resources/selection of stimuli to assign attention. Also, thinking of attention as processing resources suggests that the attended stimuli will receive further "modulation", such as being "sharpened" and entering downstream processing (Chun et al. 2011). With these clarifications, it should be easier to explain what I mean by "attention-allocation policy": it is again, a kind of "assumption" or "bias" that the attention-control mechanism/system follows to decide how to allocate attentional resources under a given condition. Now the question is, how such policies are acquired?

A potential worry here is that the control of attention must be guided by explicit LOT sentences like mental theories. For example, it might be argued, using Cummins' (1997) example, that only by having a mental theory of cats can I manage to appropriately allocate my attention to some of their properties when encountering

them. In other words, the worry here is that acquiring an attention-allocation policy presupposes the acquisition of a certain mental theory in the form of LOT sentences.

More generally, in studies of attention, it is standardly assumed that attention control is modulated in exhaustively two ways: top-down, goal-directed way and bottom-up, stimulus-driven way (Egeth and Yantis 1997). The former means attention is governed by the subject's beliefs, goals or intentions in a voluntary manner, while the latter means attention gets "captured" by physically salient (in terms of, say, luminance and motion) stimuli in an involuntary, automatic manner. It is unlikely that the kind of attention allocation I have been talking about is governed in a bottom-up, stimulus-driven manner because if attention allocation is fully, passively driven by saliency, then it does not make much sense to say we can somehow acquire or (meta-)learn a policy specific to an object/scene. An object/scene could have one of its properties being salient at this moment but another at the next moment due to environmental changes. We cannot always attend to the most useful/relevant property as we want.

However, then it looks like the kind of attention allocation I have been talking about must be governed in a top-down, goal-directed manner. If so, the worry above would emerge: beliefs, goals or intentions required for the top-down, goal-directed control are often thought of as LOT sentences. So it could be argued that to acquire an attention-allocation policy specific to x, one must acquire the appropriate beliefs like a mental theory of x to exert the right sort of top-down, goal-directed control. And that will take us back to where we started.

My response to this worry relies heavily on the studies of Awh et al. (2012) and Failing and Theeuwes (2018). The core thesis of their studies is that the distinction between the top-down, goal-directed and the bottom-up, stimulus-driven control of attention is not exhaustive. There is a third category which they call the "reward-based selection history control": the attention-control mechanism/system can learn to allocate attentional resources to the properties or property dimensions attending to which historically led to reward.

This reward-based selection history control of attention shows a possible way in which attention-allocation policies can be acquired without manipulating LOT sentences. To give a toy example, let us continue with the case of learning to use a new cup. We can think of our interactions with different cups as a series of "trials". In our first trials, the attention-control system will launch some random[18] policies, which may make us waste time attending to properties that do not matter to how to use them, like their colours, and thus slow us down in figuring out how to use a new cup because we cannot efficiently identify its commonality with previous ones. But once the attention-control system launches a policy that preferentially allocates attentional resources to the cups' structures, the increased speed in learning to use a new cup will lead to reward, which will in turn reinforce the adoption of this policy in future trials. After a period of reinforcement, we will come to possess a good enough attention-allocation policy specific to cups. Importantly, I think the above process can be accomplished purely at a subpersonal level, with the dopaminergic system playing the role of producing reward signals (Failing and Theeuwes, 2018). It does not require the manipulation of LOT sentences to form explicit hypotheses or mental theories about how to allocate attention. It could be accomplished by the neural network underlying the attention-control system adjusting its hidden-layer nodes' weights through reinforcement, which is highly consistent with Rupert's (2001) suggestion that concept acquisition can be accounted for at a neural-physical level.

To summarise this long section, I started by demonstrating the resemblance between concept acquisition (both innate and novel) and meta-learning, which is that they both consist in acquiring a certain attention-allocation policy. I claimed that this is not a coincidence because both processes are for enhancing learning and suggested that we can just conceptualise concept acquisition as a form of meta-learning. I then showed that acquiring attention-allocation policies do not necessarily involve the manipulation of LOT sentences so can avoid the kind of circularity raised by Fodor and Cummins. To move on, I shall now drive the point home and explain how I think the notion of meta-learning can help us resolve Fodor's Puzzle.

[18] In fact I think even the initial policies would be much better than random since we presumably possess the innate concept of artifacts, which can guide our attention to relevant property dimensions. But for the sake of illustration, I shall assume that we start with random policies.

6.3.2 | Using meta-learning to buffer Fodor's attack

Recall my reason 2) for conceptualising concept acquisition as meta-learning: it can free us from the Fodorian sense of learning and thus give us a conceptual tool to resolve Fodor's Puzzle. To elaborate, as I see it, at the core of Fodor's Puzzle is the premise that if a concept is unlearned, then it is either brutally acquired or innate. But as we have seen, Fodor has a very narrow understanding of learning, which is hypothesis testing by manipulating LOT sentences. Therefore, for Fodor, any non-hypothesis-testing way of acquiring a novel concept will not count as *learning* the concept -- so it must be brutally acquired. In other words, for Fodor, any non-hypothesis-testing way of acquiring a novel concept will be assimilated into those non-intelligible ways of acquiring a concept, such as being hit in the head or magical neural surgery, and thus made absurd and unintelligible. Theorists like Margolis and Laurence (2011) go to length to argue that the extension of the notion of learning should be much richer than mere hypothesis testing. I definitely agree. But I think introducing the notion of meta-learning gives us another way to deal with the dispute here.

What the notion of meta-learning does is that it gives us a theoretical space between learning as hypothesis-testing and brute acquisition and thereby "buffers" Fodor's move from the negation of the former to the latter. More specifically, granting Fodor that learning must be hypothesis testing, still, it does not mean if we do not learn a concept, we acquire it brutally. Rather, we can acquire it through meta-learning attention-allocation policies. And I have shown that such meta-learning indeed need not be like learning in Fodor's sense -- it can be accomplished without manipulating LOT sentences -- but it clearly is not brute acquisition, either (see the full argument below). With this buffer, we can block Fodor's appeal to the absurdity and unintelligibility of non-hypothesis-testing ways of concept acquisition, which, in turn, allows us to eventually reject his infamous conclusion that virtually all concepts are innate.

However, it is very likely that Fodor will not be convinced simply by a new conceptualisation. I suppose he will challenge that calling my proposed way of concept acquisition meta-"learning" does not automatically make it as intelligible as learning in his sense -- the term could just be a guise. So let me now move on to deal with this potential challenge and argue for the intelligibility of meta-learning.

6.4 | Defending the intelligibility of meta-learning

Fodor (1998) gives a more concrete argument for the unintelligibility of non-hypothesis-testing concept acquisition: these concept acquisition accounts would very likely have difficulty explaining why it is the case that acquiring a concept often requires the subject to be exposed to, namely having experiences of, the referent. He dubs it "the doorknob/DOORKNOB problem": "why is it so often experiences of doorknobs, and so rarely experience with whipped cream or giraffes, that leads one to lock to doorknobhood?" (Fodor, 1998, p. 127). Fodor also argues that a "rationalist" account of concept acquisition, namely his hypothesis testing model, can deal with the problem easily, which makes it intelligible. His explanation is that in a hypothesis testing model, experiences of the referent are required because they need to serve as *evidence* for hypothesis confirmation. As Fodor puts it,

> According to the hypothesis-testing model, the relation between the content of the concepts one acquires and the content of the experiences that eventuate in one's acquiring them is *evidential*; in particular, it's mediated by content relations between a hypothesis and the experiences that serve to confirm it. You acquire DOORKNOB from experience with doorknobs because you use the experiences to confirm a hypothesis about the nature of *doorknobhood*; and doorknobs, unlike giraffes or whipped cream, are *ceteris paribus* a good source of *evidence* about the nature of doorknobs. Come to think of it, one typically gets DOORKNOB from experience with *good* or *typical* examples of doorknobs, and good or typical doorknobs are a *very good* source of evidence about doorknobs. (Fodor, 1998, pp. 127-128; emphasis original)

In short, Fodor thinks the doorknob/DOORKNOB problem is easily solved by the hypothesis-testing model of concept acquisition because it has a convincing epistemic answer to tell, whereas it is unclear what answer other concept acquisition accounts could afford. The doorknob/DOORKNOB problem then automatically challenges my contention that concept acquisition is a form of meta-learning: can meta-learning explain why acquiring a concept often requires experiences of its referent?

To show that the answer is affirmative, let us consider again the case of cups. What makes it the case that acquiring a concept of cups requires experiences with cups? I think the answer is quite clear from the meta-learning perspective: being able to quickly learn to use a new cup depends on our deriving (i.e. meta-learning), from previous experiences with cups (the "trials"), the attention-allocation policy that allows us to filter out those irrelevant property dimensions like material and colour, and simply focus on the cup's structure. And I have argued in the last section that acquiring the sustaining mechanism of a concept of cups -- acquiring the ability to reidentify cups -- is very likely the same process as meta-learning the attention-allocation policy specific to cups. Now my answer to the doorknob/DOORKNOB problem emerges: acquiring a concept of cups requires experiences with cups because the latter provides the resources to the former for it to derive the correct attention-allocation policy from the trials. Without such experiential resources for us to meta-learn, it is almost impossible to come to possess, just by luck, the correct attention-allocation policy that can allow us to quickly learn to use a new cup/to reliably reidentify cups. It might be wondered, then, what this relationship *is* between experiences with cups and meta-learning the correct attention-allocation policy. Fodor (1998) claims that experiences are *evidence* for confirming hypotheses. For me, experiences give *feedback*[19] to the meta-learning/concept-acquisition mechanism for it to improve its functioning. More specifically, we can think of the meta-learning/concept-acquisition mechanism as an information-exploiting process that attempts to exploit the natural information presented in the environment through experiences.

[19] A philosophical complication here is in what sense feedback differs from evidence. I think this is an intriguing question but beyond the scope of this thesis. I only attempt to show that concept acquisition as meta-learning can be made intelligible.

To elaborate, recall that tools are functional kinds (Section 4.3.1), which are property clusters unified by common selective pressures (namely human design in the case of tools). In other words, instances of the same tool kind will share a number of surface-level properties, though often much less than those shared by members of a biological taxon (Godman, Mallozzi and Papineau, 2020), caused by their common design. In the case of cups, the surface-level property they share is basically just their structures. This is naturally the case because designing a thing that functions as a cup probably just consists in giving it a hollow-cylinder-with-a-bottom structure. Therefore, we can think of the meta-learning/concept-acquisition mechanism as a process exploiting this connection: by deriving the correct attention-allocation policy and thereby focusing on the surface-level commonality all cups share, it indirectly "locks on" to the super-explanatory property that unifies all instances of cups into a functional kind. In other words, cups' structures carry natural information about their "essence" in a similar way smoke carries natural information about fire -- the formers are caused by the latters. And the meta-learning/concept-acquisition mechanism exploits this information by attempting to derive the attention-allocation policy that assigns attentional resources to those information-carrying surface-level properties. Some studies of so-called "categorical perception" (e.g. Feldman 2021) also suggest that our perceptual system can be tuned by experiences, in an unsupervised manner, to preferentially allocate attentional resources to features that carry more information about the underlying category structures like their boundaries. And I have shown in Chapter 4 and 5 that being able to reliably detect such hidden, essence-like properties (e.g. latent causes) can significantly enhance learning and give rise to mental behavioural successes. Therefore, it is crucial for the meta-learning/concept-acquisition mechanism to be "fed" with relevant experiences so that it can discover which surface-level properties are more informative.

To finish my point, let us reconsider the doorknob/DOORKNOB problem following my above line of thought: acquiring DOORKNOB often requires experiences of doorknobs but not experiences of, say, giraffes because the latter cannot give the meta-learning/concept-acquisition mechanism the feedback it needs to derive the correct attention-allocation policy regarding doorknobs. This is in turn because giraffes' surface-level properties do not carry natural information about doorknobs'

essence. The meta-learning/concept-acquisition mechanism cannot use the experiences of giraffes to find out how to properly interact with doorknobs. And this should constitute a proper answer to the doorknob/DOORKNOB problem.

I think I have demonstrated that conceptualising concept acquisition as meta-learning does give us an intelligible account of concept acquisition so cannot be assimilated into brute acquisition. If so, then we can, as I have advertised, resolve Fodor's Puzzle because its core premise -- if a concept is unlearned, then it is either brutally acquired or innate -- has been shown to be false: even if a concept is unlearned, it can still be meta-learned.

Conclusion

As I mentioned in Chapter 1, adding a new item to the desiderata on a theory of concepts can often change the field. So what I have done in this thesis can be seen as an “experiment”: I added learning to the desiderata to see what would happen. My very speculative, tentative conclusion drawn from this experiment is that the learning desideratum vindicates refernalism. Abstracting away from the details, my core argument for this conclusion is that learning is a dynamic adaption to the environment, which requires the learner to use devices to keeping “monitoring” the environment so that it can adjust its internal modelling in time. Concepts are such devices. More specifically, they are devices dedicated to monitoring natural clumps. Therefore, contra the mainstream view, concepts themselves should not be thought of as knowledge states *per se*. Rather, they are used for gaining or constructing such states by playing the roles of middle terms and signals for memory storage. It is misleading, then, to assign concepts with descriptive content because that will conflate concepts with their productions. Refernalism better reflects or captures the roles of concepts in that it, by presenting concepts as referential and externally-directed, highlights concepts’ status as the bridge between its downstream systems and environmental structures.

The main limitation of this thesis is perhaps its scope. For example, we do not have concepts just for individual and kinds, but also for abstract, fictional and social entities. I did not in my thesis address how thinking of concepts as learning devices would bear on these categories. So this may be one research question I or others could investigate in the future. Also, I did not in my thesis address how concepts as learning devices interact with their status as compositional atoms. So it might be interesting to explore how we gain information will affect how we freely make use of it in a productive manner. Finally, the convergence between concept acquisition and meta-learning might be worth investigating as well because it may give rise to some new theoretical integration we did not expect before.

Bibliography

Allen, C. (1999). Animal concepts revisited: The use of self-monitoring as an empirical approach. *Erkenntnis*, 51(1), 537-544.

Allport, A. (2011). Attention and integration. *Attention: Philosophical and psychological essays*, 24-59.

Aronowitz, S. (2019). Memory is a modeling system. In *Mind & Language* (Vol. 34, Issue 4, pp. 483–502). https://doi.org/10.1111/mila.12220

Artiga, M. (2016). Teleosemantic modeling of cognitive representations. *Biology & Philosophy*, 31(4), 483-505.

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in Cognitive Sciences*, 16(8), 437–443.

Baldassarre, G. (2011). What are intrinsic motivations? A biological perspective. In *2011 IEEE international conference on development and learning* (ICDL) (Vol. 2, pp. 1-8). IEEE.

Baldassarre, G., & Mirolli, M. (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer Science & Business Media.

Barrett, C. H. (2015). The evolution of conceptual design. In Margolis, E., & Laurence, S. (2015). *The Conceptual Mind: New Directions in the Study of Concepts*. MIT Press. 151-184.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences* 22 (4):577-660.

Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems* (pp. 17-47). Springer, Berlin, Heidelberg.

Block, N. (1986). Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy* 10 (1):615-678.

Bloom, P. (2005). *Descartes' Baby: How the Science of Child Development Explains what Makes Us Human*. Random House.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. In *Trends in Cognitive Sciences* (Vol. 23, Issue 5, pp. 408–422). https://doi.org/10.1016/j.tics.2019.02.006

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 61(1/2), 127-148.

Braun, D. (2000). Russellianism and Psychological Generalizations. In *Nous* (Vol. 34, Issue 2, pp. 203–236). https://doi.org/10.1111/0029-4624.00208

Camp, E. (2007). THINKING WITH MAPS*. In *Philosophical Perspectives* (Vol. 21, Issue 1, pp. 145–182). https://doi.org/10.1111/j.1520-8583.2007.00124.x

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Carruthers, P., Laurence, S., & Stich, S. (2005). *The Innate Mind: Structure and Contents*. Oxford University Press.

Chalmers, D. J. (2002). *Philosophy of Mind: Classical and Contemporary Readings*. OUP USA.

Chomsky, N. (1972). *Language and Mind*. New York: Harcourt Brace Jovanovich.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A Taxonomy of External and Internal Attention. In *Annual Review of Psychology* (Vol. 62, Issue 1, pp. 73–101). https://doi.org/10.1146/annurev.psych.093008.100427

Connell, L. & Lynott, D. (2014). Principles of Representation: Why You Can't Represent the Same Concept Twice. *Topics in Cognitive Science* 6 (3):390-406.

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In Hirschfeld, L. A., & Gelman, S. A. (1994). *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press.

Cummins, R. (1997). The Lot of the Casual Theory of Mental Content. *The Journal of philosophy*, 94(10), 535-542.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and brain sciences*, 22(5), 735-808.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. MIT Press.

Edelman, G. (1987). *Neural Darwinism: The Theory Of Neuronal Group Selection*.

Edwards, K. (2009). What concepts do. *Synthese* 170 (2):289 - 310.

Edwards, K. (2010). Unity amidst heterogeneity in theories of concepts. The *Behavioral and Brain Sciences*, 33(2-3), 210–211.

Edwards, K. (2013). Keeping (Direct) Reference in Mind. *Noûs* 47 (1):342-367.

Egeth, H. E., & Yantis, S. (1997). Visual attention: control, representation, and time course. *Annual Review of Psychology*, 48, 269–297.
Eliasmith, C. (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford University Press.
Failing, M., & Theeuwes, J. (2018). Selection history: How reward modulates selectivity of visual attention. *Psychonomic Bulletin & Review*, 25(2), 514–538.
Feldman, J. (2021). Mutual Information and Categorical Perception. *Psychological Science*, 32(8), 1298–1310.
Figdor, C. (2009). Semantic externalism and the mechanics of thought. *Minds and Machines*, 19(1), 1-24.
Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.
Fodor, J. A. (1979). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. MIT Press.
Fodor, J. A. (1990). *A Theory of Content and Other Essays*. MIT Press.
Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.
Fodor, J. A. (2008). *LOT 2: The Language of Thought Revisited*. OUP Oxford.
Fodor, J. A. & Lepore, E. (1992). *Holism: A Shopper's Guide*. Blackwell.
Fodor, J. A., & Lepore, E. (2002). *The compositionality papers*. Oxford University Press.
Frankland, S. M., & Greene, J. D. (2020). Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology*, 71, 273-303.
Frege, G. (1893). On Sense and Meaning. In Black, M., & Geach, P. (1953). *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Blackwell.
Gelman, R. (2009). Innate learning and beyond. In Chomsky, N. (2009). *Of minds and language: a dialogue with Noam Chomsky in the Basque country*. Oxford University Press. (pp. 223-35).
Gelman, R., & Brenneman, K. (2004). Science learning pathways for young children. *Early childhood research quarterly*, 19(1), 150-158.
Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: early understandings of the non-obvious. *Cognition*, 38(3), 213–244.
Gershman, S. J., Monfils, M.H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *eLife*, 6. https://doi.org/10.7554/eLife.23763

Gertler, B. (2012). UNDERSTANDING THE INTERNALISM-EXTERNALISM DEBATE: WHAT IS THE BOUNDARY OF THE THINKER? In *Philosophical Perspectives* (Vol. 26, Issue 1, pp. 51–75). https://doi.org/10.1111/phpe.12001

Godman, M., Mallozzi, A., & Papineau, D. (2020). Essential Properties Are Super-Explanatory: Taming Metaphysical Modality. In *Journal of the American Philosophical Association* (Vol. 6, Issue 3, pp. 316–334). https://doi.org/10.1017/apa.2019.48

Goodman, R., & Genone, J. (2020). Singular Thought and Mental Files. In *Singular Thought and Mental Files* (pp. 1–18). https://doi.org/10.1093/oso/9780198746881.003.0001

Goodman, R., & Gray, A. (2022). Mental filing. *Noûs*, 56(1), 204-226.

Green, E. J. & Quilty-Dunn, J. (2021). What Is an Object File? *British Journal for the Philosophy of Science* 72 (3):665-699.

Harlow, H. F. (1949). The formation of learning sets. In *Psychological Review* (Vol. 56, Issue 1, pp. 51–65). https://doi.org/10.1037/h0062474

Harman, G. (1999). (Nonsolipsistic) Conceptual Role Semantics. In *Reasoning, Meaning, and Mind* (pp. 206–232). https://doi.org/10.1093/0198238029.003.0013

Hein, E., & Moore, C. M. (2012). Spatio-temporal priority revisited: the role of feature identity and similarity for object correspondence in apparent motion. *Journal of Experimental Psychology. Human Perception and Performance*, 38(4), 975–988.

Hollingworth, A., & Franconeri, S. L. (2009). Object correspondence across brief occlusion is established on the basis of both spatiotemporal and surface feature cues. *Cognition*, 113(2), 150–166.

Inagaki, K., and G. Hatano. 2002. *Young Children's Naive Thinking about the Biological World*. New York: Psychological Press.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, 24(2), 175–219.

Kaplan, F., & Oudeyer, P.-Y. (2007). In search of the neural circuits of intrinsic motivation. *Frontiers in Neuroscience*, 1(1), 225–236.

Keil, F. C., & Kominsky, J. F. (2015). Grounding concepts. In *The conceptual mind: new directions in the study of concepts*, 677-692.

Kominsky, J. F., & Keil, F. C. (2014). Overestimation of knowledge about word meanings: The "misplaced meaning" effect. *Cognitive science*, 38(8), 1604-1633.

Kriegel, U. (2013). *Phenomenal Intentionality*. Oxford University Press.

Kripke, S. A. (1980). *Naming and Necessity*. Harvard University Press.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. In *Behavioral and Brain Sciences*, 40, e253.
Laurence, S. & Margolis, E. (1999). Concepts and Cognitive Science. In Margolis E. & Laurence S. (eds.), *Concepts: Core Readings*. MIT Press. pp. 3-81.
Laurence, S., & Margolis, E. (2002). Radical concept nativism. *Cognition*, 86(1), 25–55.
Lee, P. (2018). Mental files, concepts, and bodies of information. *Synthese* 195 (8): 3499-3518.
Lieder, Falk & Griffiths, Thomas L. (forthcoming). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*:1-85.
Löhr, G. (2020). Concepts and categorization: do philosophers and psychologists theorize about different things? *Synthese*, 197(5), 2171-2191.
Machery, E. (2009). *Doing without concepts*. Oxford University Press.
Machery, E. (2010). Précis of doing without concepts. *The Behavioral and Brain Sciences*, 33(2-3), 195–206; discussion 206–244.
Margolis, E. (1998). How to acquire a concept. *Mind & Language*, 13(3), 347-369.
Margolis, E., & Laurence, S. (1999). *Concepts: Core Readings*. MIT Press.
Margolis, E., & Laurence, S. (2011). Learning matters: The role of learning in concept acquisition. *Mind & Language*, 26(5), 507-539.
Margolis, E., & Laurence, S. (2013). In defense of nativism. In *Philosophical Studies* (Vol. 165, Issue 2, pp. 693–718). https://doi.org/10.1007/s11098-012-9972-x
Margolis, E., & Laurence, S. (2015). *The Conceptual Mind: New Directions in the Study of Concepts*. MIT Press.
McLaughlin, B. P., & Cohen, J. (2009). *Contemporary Debates in Philosophy of Mind*. John Wiley & Sons.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. In *Psychological Review* (Vol. 85, Issue 3, pp. 207–238). https://doi.org/10.1037/0033-295x.85.3.207
Medin, D. L., & Ortony, A. (1989). Psychological essentialism. *Similarity and analogical reasoning*, 179.

Mendelovici, A. (2018). *The Phenomenal Basis of Intentionality*. Oxford University Press.
Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual review of psychology*, 32(1), 89-115.
Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.
Millikan, R. G. (1998). A common structure for concepts of individuals, stuffs, and real kinds: More mama, more milk, and more mouse. *Behavioral and Brain Sciences*, 21(1), 55-65.
Millikan, R. G. (2000). *On Clear and Confused Ideas: An Essay about Substance Concepts*. Cambridge University Press.
Millikan, R. G. (2017). *Beyond Concepts: Unicepts, Language, and Natural Information*. Oxford University Press.
Mole, C., Smithies, D., & Wu, W. (2011). *Attention: Philosophical and Psychological Essays*. Oxford University Press.
Moore, C. M., Stephens, T., & Hein, E. (2010). Features, as well as space and time, guide object persistence. *Psychonomic Bulletin & Review*, 17(5), 731–736.
Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychological Review*, 98(2), 164–181.
Murez, M., & Recanati, F. (2016). Mental files: An introduction. *Review of Philosophy and Psychology*, 7(2), 265-281.
Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289–316.
Neufeld, E. (2022). Psychological essentialism and the structure of concepts. *Philosophy Compass*, 17(5).
Newman, G. E., & Knobe, J. (2019). The essence of essentialism. *Mind & Language*, 34(5), 585-605.
Papineau, D. (1987). *Reality and Representation*. Wiley-Blackwell.
Papineau, D. (1993). *Philosophical Naturalism*. Wiley-Blackwell.
Papineau, D. (2006). Phenomenal and perceptual concepts. In Alter, T., & Walter, S. (eds), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.
Papineau, D. (2013). Comments on François Recanati's Mental Files: Doubts about Indexicality. *Disputatio*, 5(36), 159-75.

Papineau, D. Swampman, Teleosemantics and Kind Essences. MS.
Prinz, J. J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. MIT Press.
Prinz, J. J. (2005). The return of concept empiricism. In H. Cohen & C. Leferbvre (eds.), *Categorization and Cognitive Science*. Elsevier.
Prinz, J. J. (2009). Is consciousness embodied. In *The Cambridge handbook of situated cognition*, 419-436.
Prinz, J. (2012). Regaining composure: A defense of prototype compositionality. *The Oxford handbook of compositionality*, 437-453.
Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7:131-193
Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Sciences*, 4(5), 197–207.
Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: a constructivist manifesto. *The Behavioral and Brain Sciences*, 20(4), 537–556; discussion 556–596.
Quilty-Dunn, J. (2020a). Concepts and predication from perception to cognition. In *Philosophical Issues* (Vol. 30, Issue 1, pp. 273–292). https://doi.org/10.1111/phis.12185
Quilty-Dunn, J. (2020b). Perceptual Pluralism. In *Noûs* (Vol. 54, Issue 4, pp. 807–838). https://doi.org/10.1111/nous.12285
Quilty-Dunn, J. (2021). Polysemy and thought: Toward a generative theory of concepts. *Mind and Language* 36 (1):158-185.
Quilty-Dunn, J. & Green, E. J. (forthcoming). Perceptual attribution and perceptual reference. *Philosophy and Phenomenological Research*.
Radulescu, A., Shin, Y. S., & Niv, Y. (2021). Human representation learning. *Annual Review of Neuroscience*, 44(1), 253-273.
Recanati, F. (2012). *Mental files*. Oxford University Press.
Rey, G. (1983). Concepts and stereotypes. *Cognition* 15 (1-3):237-62.
Richard, A. M., Luck, S. J., & Hollingworth, A. (2008). Establishing object correspondence across eye movements: Flexible use of spatiotemporal and surface feature information. *Cognition*, 109(1), 66–88.
Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4), 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. In *Cognitive Psychology* (Vol. 8, Issue 3, pp. 382–439). https://doi.org/10.1016/0010-0285(76)90013-x

Rupert, R. D. (2001). Coining terms in the language of thought: Innateness, emergence, and the lot of Cummins's argument against the causal theory of mental content. *The Journal of philosophy*, 98(10), 499-530.

Russell, B. (1905). On denoting. *Mind*, 14(56), 479-493.

Schneider, S. (2005). Direct Reference, Psychological Explanation, and Frege Cases. In *Mind and Language* (Vol. 20, Issue 4, pp. 423–447). https://doi.org/10.1111/j.0268-1064.2005.00294.x

Shea, N. (2015). Distinguishing Top-Down From Bottom-Up Effects. In D. Stokes, M. Matthen & S. Biggs (eds.), *Perception and Its Modalities*. Oxford University Press. pp. 73-91.

Shea, N. (2016). Representational development need not be explicable-by-content. In Müller, V. C. *Fundamental Issues of Artificial Intelligence* (pp. 223-240). Springer, Cham.

Shea, N. (2018). *Representation in Cognitive Science*. OUP Oxford.

Smith, E., Medin, D., and Rips, L. (1984). A Psychological Approach to Concepts: Comments on Rey's "Concepts and Stereotypes." *Cognition*, 17, 265—274.

Tooby, J., & Cosmides, L. (2001). Does beauty build adapted minds? Toward an evolutionary theory of aesthetics, fiction, and the arts. *SubStance*, 30(1), 6-27.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373), 1295-1306.

Vicente, A. (2018). Polysemy and word meaning: an account of lexical meaning for different kinds of content words. *Philosophical Studies*, 175(4), 947-968.

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. In *Current Opinion in Behavioral Sciences* (Vol. 38, pp. 90–95). https://doi.org/10.1016/j.cobeha.2021.01.002

Ward, T. B., & Becker, A. H. (1992). Learning Categories with and without Trying: Does it Make a Difference? In *Advances in psychology* (Vol. 93, pp. 451-491). North-Holland.

Weiskopf, D. A. (2009). Atomism, Pluralism, and Conceptual Content. In *Philosophy and Phenomenological Research* (Vol. 79, Issue 1, pp. 131–163). https://doi.org/10.1111/j.1933-1592.2009.00269.x

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual review of psychology*, 43(1), 337-375.