THE UNIVERSITY *of* EDINBURGH

A function-first approach to *doubt*

Lilith Mace

A doctoral thesis in Philosophy, submitted to the University of Edinburgh in 2022

Abstract

Doubt is a much-maligned state. We are racked by doubts, tormented by doubts, plagued by them, paralysed. Doubts can be troubling, consuming, agonising. But however ill-regarded is doubt, anxiety is more so. We recognise the significance of doubting in certain contexts, and allow ourselves to be guided by our doubts. For example, the criminal standard of proof operative in the U.K., U.S., as well as in most other anglophone countries, Germany, Italy, Sweden and Israel, requires for conviction to be permissible that the defendant's guilt is proved beyond reasonable doubt; to feel a doubt about a defendant's guilt, so long as it is reasonable, is reason to refrain from convicting. But our folk understanding of anxiety ascribes no value to that state. Anxiety is inherently unpleasant and irrational; it prevents us from being able to perform well when it is most important to us that we do; it is an emotion that, if we could, we'd eliminate from our emotional toolbox. Yet in this thesis, I offer a vindication of doubt – a defence of doubt in terms of what it does for us – on which it ultimately turns out to be a kind of anxiety. The basic idea is that the concept *doubt* serves a function for us that we couldn't do without: it signals when we should begin inquiry. I will argue that *doubt* is able to serve this function because the state it picks out, the state of doubt, is a kind of anxiety: epistemic anxiety. I develop a picture of epistemic anxiety as an emotional response to epistemic risk: potential disvalue in the epistemic realm. Because doubt is a kind of anxiety, it has the right kind of representational and motivational profile to track epistemic risk in our environments, and motivate us to reduce or avoid that risk. This makes it hugely valuable for us, as knowledge-seeking creatures, given the incompatibility of knowledge with high levels of epistemic risk.

My central concern in this thesis is to develop this picture of the concept *doubt* and its object, the state of doubt. A secondary concern is metaphilosophical. In recent years, a number of philosophers have presented themselves as taking 'function-first' approaches to their inquiries. However, there is disagreement over what, exactly, the function-first approach consists in. Is it a distinct philosophical method, or a methodology: a collection of methods? What is the relationship between the function-first approach and the methods of conceptual engineering and conceptual reverse-engineering, which are often discussed together? I will argue that the function-first approach is a methodology that encompasses many philosophical methods, including

conceptual engineering and conceptual reverse-engineering. In my thesis, I hope both to clarify what the methodology of function-first philosophy and the methods of conceptual engineering and conceptual reverse-engineering consist in, as well as to provide a case study for the fruitfulness of the function-first approach and these two methods within philosophy, namely my own function-first approach to the concept *doubt*.

The thesis proceeds as follows. In Chapter 1, I develop a novel account of the function-first approach as a philosophical methodology, a collection of methods that includes both conceptual engineering and conceptual reverse-engineering. I argue that the function-first approach is distinguished from more familiar methodologies in that it treats facts about a concept's function as in some way prior to facts about the concept's intension or extension. Exactly what this priority consists in depends on the particular philosophical project in which the function-first philosopher is engaged. In Chapter 2, I give accounts of two methods that are particularly useful for philosophers taking function-first approaches to concepts: conceptual reverse-engineering and conceptual engineering. I argue that conceptual reverse-engineering is a method for confirming or disconfirming hypotheses about the functions of concepts by looking to what the concept does for some group of agents in a typical case: a case that is representative of how those agents use the concept. I argue that conceptual engineering is a method for improving our conceptual world by revising, replacing or abandoning defective concepts, improving non-defective concepts, or creating new concepts to serve legitimate purposes currently going unmet. In Chapter 3, I evaluate four conceptual reverse-engineering projects on *knowledge*, two that test the hypothesis that *knowledge* functions to flag good informants, and two that test the hypothesis that *knowledge* functions to signal the legitimate end of inquiry. I argue that the hypothesis that *knowledge* functions to signal the legitimate end of inquiry is to be preferred to its alternative. In Chapter 4, I reverse-engineer the concept *doubt*. I argue that similar considerations as those motivating the inquiry-stopper picture of *knowledge* suggest a distinct conceptual need for an inquiry-starter: a concept to signal when inquiry should begin. I test the hypothesis that meeting this need is the function of *doubt*. I then explore what *doubt* must be like to serve this function. A picture emerges on which *doubt* typically applies to some subject S with respect to a question Q when S has a questioning attitude to Q; S does not believe any complete answer to Q; S's situation with respect to Q is, or is represented to S as, epistemically risky; and

S is immediately motivated to inquire into Q. In Chapter 5, I argue that this concept *doubt* picks out an epistemic emotion, epistemic anxiety. I offer an account of epistemic anxiety as an emotional response to epistemic risk: potential disvalue in the epistemic realm. I then embark on a conceptual reverse-engineering project on *doubt*, replacing the inexact concept articulated in the previous chapter with the more exact concept *epistemic anxiety*, and I demonstrate the fruitfulness of this conceptual engineering project by applying the engineered concept to a number of debates within epistemology.

Lay summary

This thesis develops an account of the concept *doubt* in terms of what it does for us. It is argued that *doubt* serves a very important need for us: it signals when we ought to inquire into some question, with the aim of finding out its answer. This is important, because there are, at any given time, any number of questions into which we can inquire. Some of these questions are more important than others. But we are finite creatures, with limited time and mental resources for answering questions, so we cannot inquire into just any question. As such, we have a need for a concept that signals when a question requires an answer, so we should inquire into it. It is argued that meeting this need is the function, point, or purpose of the concept *doubt*. This is what *doubt* does for us that explains why we have the concept in the first place.

Given that *doubt* serves this function, the thesis then explores what *doubt* must be like. It is argued that *doubt* must pick out some state that represents a question as in need of an answer, and motivates a person to inquire into the question. Anxiety has this kind of profile. Anxiety represents something in a person's environment to her as risky, and because anxiety feels unpleasant, it motivates the person to do something to reduce or avoid this risk. It is argued that *doubt* picks out a kind of anxiety: anxiety about risks that have to do with our thinking about how things are, with the aim of coming to have knowledge. This kind of anxiety, called 'epistemic anxiety', can take many forms. For example, a person might be anxious about coming to have a false belief about how things are, or about failing to have a true belief, or she might be anxious about misunderstanding something, or failing to understand it. Because epistemic anxiety is unpleasant, one who experiences epistemic anxiety is motivated to take steps to reduce or avoid the relevant risk, by inquiring into some relevant question. So *doubt* is able to serve its function of signalling when we should inquire into some question because it picks out this emotion, epistemic anxiety, which is triggered by certain kinds of risk and which motivates people to inquire in order to reduce or avoid those risks.

Acknowledgements

I had a great time doing my PhD at Edinburgh, largely because I was in such good company throughout. I am grateful to all my friends and colleagues in PPLS for making that the case, as well as to my philosophy friends across Scotland. I was particularly lucky to know Angie O'Sullivan, John Heron, Dario Mortini, Ross Patrizio, Claire Field, Matt Jope and Josh Cox: thank you for all the philosophy you've chatted with me over the years.

I would like to thank my examiners, Christoph Kelp and Matthew Chrisman. It was a thrill to have such brilliant philosophers engage so closely with my work. I don't expect that to happen again any time soon.

I am grateful for the love and support of my family, especially my mother Jane, who never doubts for a second that I can do whatever I put my mind to, yet still gets excited at even my tiniest achievements.

I am grateful to my partner Iain Campbell, and to Mira and Elsie, for making my home life maximally pleasant while I completed my PhD.

I am grateful to every supervisor and mentor I have had. In chronological order: Adam Rieger, Fraser MacBride, John Hyman, Tim Williamson, Ofra Magidor, Aidan McGlynn and Mona Simion. Each has deeply shaped how I think about knowledge, doubt, possibility, and everything else that preoccupies me in philosophy, sometimes in ways that are difficult to square with each other. I'll spend the rest of my career trying to make that square.

Above all, I would like to thank my primary PhD supervisor Martin Smith. I was incredibly lucky to have such a generous, thoughtful, patient and kind supervisor, who is also the smartest person in the world. Without him, writing this thesis would not have been possible, let alone normal.

Table of contents

Introduction

Doubt is a much-maligned state. We are racked by doubts, tormented by doubts, plagued by them, paralysed. Doubts can be troubling, consuming, agonising. William Shakespeare wrote that "[o]ur doubts are traitors" (1991, Act 1, Scene 1, line 77); Gustave Flaubert that "[d]oubt is the death of the soul" (2001: 65). But however ill-regarded is doubt, anxiety is more so. We recognise the significance of doubting in certain contexts, and allow ourselves to be guided by our doubts. For example, the criminal standard of proof operative in the U.K., U.S., as well as in most other anglophone countries, Germany, Italy, Sweden and Israel, requires for conviction to be permissible that the defendant's guilt is proved beyond reasonable doubt; to feel a doubt about a defendant's guilt, so long as it is reasonable, is reason to refrain from convicting. But our ordinary understanding of anxiety ascribes no value to that state. As Charlie Kurth writes, "folk wisdom tells us that anxiety is an inherently unpleasant, pernicious emotion" (2018: 2). Anxiety brings "havoc and disaster", "taking our attention away from what matters or, worse, paralyzing us when we need to act" (2). Further, this folk conception of anxiety as unpleasant and unhelpful is borne out by psychological research, with a recent review of research investigating the effects of anxiety in evaluative settings concluding that anxiety is "predominantly harmful to task performance" (Zeidner and Matthews 2005: 147).

Yet in this thesis, I offer a vindication of doubt, a defence of doubt in terms of what it does for us, on which it ultimately turns out to be a kind of anxiety. The basic idea is that the concept *doubt* serves a function for us that we couldn't do without: it signals when we should begin an inquiry. I argue that the concept *doubt* is able to serve this function because the state it picks out, namely the state of doubt, is a kind of anxiety: epistemic anxiety. I develop a picture of epistemic anxiety as an emotional response to epistemic risk: potential disvalue in the epistemic realm. Because doubt is a kind of anxiety, it has the right kind of representational and motivational profile to track epistemic risk in our environments, and motivate us to reduce or avoid that risk. This makes it hugely valuable for us, as knowledge-seeking creatures, given the incompatibility of knowledge with high levels of epistemic risk (or so I will argue).

I am not the first person to put forward an account of doubt as epistemic anxiety. Christopher Hookway (1998, 2008), to whom the expression "epistemic anxiety" is to

be credited (1998: 222), argues that C. S. Peirce's (1877) notion of real doubt can be understood as "a kind of anxiety about any inquiry that relies upon a doubted proposition" (Hookway 1998: 221). Hookway, too, sees doubt as having an important role in motivating inquiry, writing that the anxiety involved in real doubt "can motivate us to inquire further, seeking the source of the anxiety, evaluating its appropriateness, or acting to revise our opinions so that it does not arise anymore" (2008: 62).

My project has been in part inspired by Hookway, and in particular by his work on Peirce. But there are substantive differences between the accounts of doubt that each of us offer. For one thing, I develop an account of doubt as a questioning attitude: an attitude we take to questions (Chapter 4, 'The point of *doubt*'). Hookway, in contrast, develops a picture of doubt as a propositional attitude (see for example 1998: 204-6; 2008: 61-2). I will argue that it is crucially important for the concept *doubt* to be able to function as I argue it does that this concept picks out a questioning attitude (though I argue, in §1.1. of Chapter 4, that the word 'doubt' is polysemous, picking out both a propositional attitude and a questioning attitude). Further, for Hookway, epistemic anxiety is anxiety about some proposition in which one already has a belief turning out to be false. On my picture, epistemic anxiety is much broader than this. I argue (in Chapter 5, 'From *doubt* to *epistemic anxiety*') that the formal object of epistemic anxiety is epistemic risk, broadly understood, such that any potentially obtaining epistemically disvaluable event is an epistemic risk event. Holding a false belief is just one kind of epistemic risk event, alongside, for example, failing to form a valuable true or knowledge-constituting belief, failing to understand something, or misunderstanding it. I have developed my account of epistemic anxiety with an eye to the value that the phenomenon has from the perspective of anti-risk epistemology (Pritchard 2015, 2016; Navarro 2019, 2021). It is because epistemic anxiety, on my account, is an emotional response to epistemic risk that it can helpfully guide our inquiries with the goal of achieving knowledge, on the assumption from anti-risk epistemology that knowledge is incompatible with high levels of veritic epistemic risk (see §3.2 of Chapter 5).

My central concern in this thesis is to develop this picture of the concept *doubt* and its object, the state of doubt. A secondary concern is metaphilosophical. In recent years, a number of philosophers have presented themselves as taking "function-first" approaches to their inquiries (see Hannon 2019, Simion and Kelp 2020, Queloz 2021). However, there is disagreement over what, exactly, taking a function-first approach

consists in. Is the function-first approach a distinct philosophical method, as Michael Hannon (2019) claims, or a methodology, as Matthieu Queloz (2021) suggests? And what is the relationship between the function-first approach and the method of conceptual engineering, popularised by Herman Cappelen (2018), though in which philosophers have purportedly been participating for decades (Cappelen claims as conceptual engineers Carnap (1950), Quine (1960), Railton (1989), Clark and Chalmers (1998), Haslanger (2000), Joyce (2001), Scharp (2013))? Further, what is the relationship between function-first philosophy, conceptual engineering, and the method of conceptual reverse-engineering (Dogramaci 2012, Queloz 2021)? Hannon holds that the method of conceptual engineering might be identical to his function-first method (2019: 25), but that the function-first method and conceptual reverse-engineering are distinct (22-23). I will argue (Chapter 1, 'The function-first approach') that a better way of understanding the function-first approach sees it as a methodology that encompasses a number of different methods, including conceptual engineering and conceptual reverse-engineering. In my thesis, I hope both to clarify what the methodology of function-first philosophy and the methods of conceptual engineering and conceptual reverse-engineering consist in, as well as to provide a case study for the fruitfulness of the function-first approach and these two methods within philosophy, namely my own function-first approach to the concept *doubt*.

The thesis proceeds as follows. In Chapter 1, I offer a novel account of the function-first approach as a philosophical methodology for theorising about concepts, which is distinguished from more familiar methodologies in that it treats the function of a concept as in some way prior to its intension or extension. Exactly what this priority consists in depends on the philosophical project in which the function-first philosopher is engaged. I raise and respond to some objections to my account of the function-first approach. Most significant for what follows is an objection raised by Cappelen (2018) against philosophers who speak of concepts having 'functions'. Cappelen argues that the only function that concepts have are to denote their objects; but the kinds of functions these philosophers talk about go beyond these basic denoting functions. For example, I argue that *doubt* functions to signal that inquiry should begin. I respond to Cappelen that, first, it is unclear whether all concepts have denoting functions; and second, that we can understand at least some concepts as having proper functions in Ruth Millikan's (1984) sense, which go beyond the denoting function that Cappelen identifies.

In Chapter 2, I give accounts of two methods that are particularly useful for philosophers taking function-first approaches to concepts: conceptual reverse-engineering and conceptual engineering. I argue that conceptual reverse-engineering is a method for confirming or disconfirming hypotheses about the functions of concepts by looking to what the concept does for some group of agents in a typical case: a case that is representative of how the concept is used by those agents. I outline a taxonomy of conceptual reverse-engineering, on which different sub-methods of conceptual reverse-engineering are distinguished in terms of how they represent the typical case: via a model or an abstract direct representation; and in virtue of whether this model or abstract direct representation includes a time-axis. I argue that conceptual engineering is a method for improving our conceptual world by revising, replacing or abandoning defective concepts, improving non-defective concepts, or creating new concepts to serve legitimate purposes currently going unmet. I offer a taxonomy of conceptual engineering, on which different sub-methods of conceptual engineering are distinguished in terms of whether they revise a concept or create a new one; and whether there was already some concept in use for some purpose that is being revised or replaced, or whether the newly created concept had no relevant predecessor. I raise and respond to two objections to conceptual engineering that have been articulated under the name of the 'Strawsonian challenge' to conceptual engineering. The first says that to engineer a concept in order to solve a philosophical problem, advance a debate, or otherwise engage in some philosophical project is to change the subject in such a way that one can no longer be solving the same problem, advancing the same debate, engaging in the same project. I call this the 'topic-preservation challenge'. The second says that concepts have their intensions and extensions essentially, so any change to a concept's intension or extension means abandoning that concept and creating a new one in its place. As such, the idea that we can revise concepts without numerical change of concept is incoherent. I call this the 'incoherence challenge'. I note that the incoherence challenge only applies to the sub-method of conceptual engineering on which concepts are understood as being revised without being replaced. As I don't engage in this sub-method, it is not important to me whether this challenge can be met. I survey several responses to the topic-preservation challenge, and argue that different responses are appropriate for different projects.

In Chapter 3, I evaluate four conceptual reverse-engineering projects on *knowledge*, undertaken by Edward Craig (1990), Hannon (2019), Christoph Kelp

(2011) and Klemens Kappel (2010). Craig and Hannon test the hypothesis that *knowledge* functions to flag good informants, while Kelp and Kappel test the hypothesis that *knowledge* functions to signal when inquiry should come to an end. I defend the inquiry-stopper picture of *knowledge* endorsed by Kelp and Kappel over the informant-flagging picture endorsed by Craig and Hannon, arguing that the inquiry-stopper picture is more robustly supported than the informant-flagging picture; that the inquiry-stopper picture explains why knowledge should have a modal structure (that is to say, whether a subject knows will depend in some way on what goes on in other possible worlds); and that the inquiry-stopper function is explanatorily prior to the informant-flagging function, in that *knowledge* having the former function would explain how it could also serve the latter, but not *vice versa*. I explain away some recalcitrant data for the inquiry-stopper picture: some cases in which a subject intuitively knows some proposition P, yet is epistemically permitted to continue inquiring into the question whether P.

In Chapter 4, I embark on my own conceptual reverse-engineering projects on *doubt*. I argue that similar considerations as those that motivate the inquiry-stopper picture of *knowledge* suggest a distinct need for a concept to flag when inquiry should begin. I test the hypothesis that meeting this need is the function of the concept *doubt*, or at least one concept that we call by the name '*doubt*', which picks out the questioning attitude of doubt. I reverse-engineer this concept using both a synchronic model and a diachronic model. Both conceptual reverse-engineering projects confirm my hypothesis. The picture of *doubt* that arises has it that the concept typically applies to some subject S with respect to a question Q when:

1. S has a questioning attitude to Q,
2. S does not believe any complete answer to Q,
3. S's situation with respect to Q is, or is represented to S as, epistemically risky,
4. S is immediately motivated to inquire into Q.

I raise and respond to some objections to this characterisation of *doubt*: first, that Jane Friedman's (2017, 2019) concept of *suspended judgement* could better play the role of the inquiry-starter than *doubt*; second, that not all inquiring creatures are capable of doubt, making my account of *doubt* as the (rather than a) inquiry-starter untenable; and third, that there is a disanalogy between the roles of *knowledge* and *doubt* in inquiry that undermines my claim that *doubt* is to starting inquiry as *knowledge* is to

stopping it. I end by considering whether my account of *doubt* commits me to any norms governing the start of inquiry.

In Chapter 5, I argue that the concept *doubt* articulated in the previous chapter picks out epistemic anxiety. Jennifer Nagel posits epistemic anxiety as a "force" (2010a: 408) that triggers subjects to gather information and reason more carefully in high-stakes contexts. But she doesn't have much to say about the nature of epistemic anxiety. I offer an account of epistemic anxiety as an emotional response to epistemic risk, and explain how epistemic anxiety, so understood, can do the epistemological work to which Nagel puts it. I argue that this account improves on extant accounts of epistemic anxiety in the literature: it is more fleshed-out than Nagel's, and it is more plausible than Juliette Vazard's (2018, 2021), on which epistemic anxiety is an emotional response to potential threat to one's practical interests. Vazard's account fails to distinguish epistemic anxiety from anxiety in general, and fails to capture all intuitive instances of epistemic anxiety as such. My account does better on both counts. I then argue that we should understand the questioning attitude of doubt as identical to epistemic anxiety, as doing so enables us to advance a number of debates within and outwith philosophy. Finally, I embark on my own conceptual engineering project: using Rudolf Carnap's (1947, 1950) method of explication, I replace the concept *doubt* articulated in Chapter 4 with the more exact concept *epistemic anxiety*, and demonstrate the adequacy of this explication on Carnap's own criteria. Much of the content of this chapter has been published in my article 'Epistemic anxiety and epistemic risk' (Newton 2022).

Chapter 1. The function-first approach

1. Introduction

In this chapter, I introduce the methodology that underpins my thesis: the function-first approach to concepts. In §1.1, I distinguish methods from methodologies, both within and outwith philosophy. In §2, I develop my account of the function-first approach to concepts as a methodology, in contrast to Michael Hannon's characterisation of the function-first approach as a singular method. I argue that two other methods, Edward Craig's hypothetical genealogy and Sally Haslanger's ameliorative method, are motivated by the same insight as Hannon's: that we can better understand, evaluate, and generally theorise about our concepts through an exploration of their functions. I argue that we should for this reason treat all three methods as 'function-first approaches', and as such we should not identify the function-first approach with any one method. I then offer a more general argument for thinking of the function-first approach as a methodology: this insight motivates a number of different inquiries, which require the use of different methods; we should think of the function-first approach as the collection of all these methods; thus, we should think of the function-first approach as a methodology. I illuminate this methodology by contrasting it with two competing methodological approaches: intension-first and extension-first approaches to concepts. On my account, what is distinctive about the function-first methodology is that it includes only methods that, in some sense to be unpacked (§2.3), give priority to the functions of concepts over their intensions and extensions in theorising about the concept. In §3, I consider and respond to some objections to my account of the function-first approach: first, that it cannot retain the benefits of Hannon's account of the function-first approach as a method; second, that concepts cannot have functions of the kind that the function-first approach requires.

1.1. Methods and methodologies

'Philosophical methodology' is an ambiguous phrase: it has two meanings. On the first, it means the collection of methods that philosophers use. Example: "philosophers use methods of various kinds: they philosophize in various ways. A philosophical community's methodology is its repertoire of such methods" (Williamson 2007: 3). On the second, it means the study of philosophy itself: how philosophers ought to do

philosophy, what makes for good philosophy, what methods philosophers ought to use, and so on. Example: "Philosophical Methodology is the study of philosophical method: how to do philosophy well" (Dever 2016: 20). Timothy Williamson uses "the philosophy of philosophy" to name the study of philosophy (2007: 5-6); others use "metaphilosophy" for the same purpose (Lazerowitz 1970; Overgaard, Gilbert and Burwood 2013; Moser 2015). I am not interested in the second sense of 'philosophical methodology' – I am not concerned in this thesis with the philosophical study of philosophy itself. When I use 'philosophical methodology' I intend it in the first sense: as a collection of methods that philosophers use when doing philosophy.

By 'method', I mean a way of doing something; by 'philosophical method', I mean a way of doing philosophy. Examples of philosophical methods are conceptual analysis, whereby a philosopher attempts to construct necessary and sufficient conditions for an object's satisfying a concept; conceptual engineering, whereby a philosopher revises or replaces a concept that is defective or which could be improved, or constructs a new concept to serve a particular purpose (see §3 of Chapter 2); and conceptual reverse-engineering, whereby a philosopher attempts to reconstruct the needs, both practical and theoretical, which a concept serves for a community in order to understand and evaluate the concept (see §2 of Chapter 2; Queloz 2021).

Methods can involve multiple steps. This is true both within and outwith philosophy. For an example within philosophy, consider conceptual analysis, as traditionally understood (see Craig 1990: 1; Williamson 2000: 3; Hannon 2019: 16; Queloz 2021: 43-4). This method, sometimes called "reductive conceptual analysis" (Hannon 2019: 16), involves three steps: first, the philosopher identifies the concept of interest, for example the concept *knowledge*; second, she constructs necessary and sufficient conditions for an object's satisfying the concept; third, she checks whether intuitive instances of (for example) knowledge are captured as such, and whether intuitive instances of non-knowledge are ruled out, by her analysis; and if sufficiently many, or sufficiently important, intuitive instances of knowledge are not captured, or intuitive instances of non-knowledge are captured, she revises her proposed necessary and sufficient conditions. This third and final step will involve the theorist seeking reflective equilibrium between extensional adequacy – capturing all and only intuitive instances of X under her concept *X* – and preserving other theoretical virtues of her account. For an example outside philosophy, consider the 'curly girl method' of haircare: first, the hair-haver washes her hair with conditioner or a sulphate-free

shampoo; second, she conditions and combs her hair; third, she gels her hair; fourth, she air dries her hair (Metzger 2021).

Methods have aims and outputs. The aim of conceptual analysis, traditionally understood, is to accurately describe one's concept of interest in a way that is both non-circular and illuminating. That is, to describe the concept in such a way that does not invoke the concept, and in terms of other notions that are better understood than the concept of interest. The output of conceptual analysis, at least when the method is successful, is a set of individually necessary and jointly sufficient conditions for an object's satisfying the concept. The aim of the curly girl method is to take care of curly hair; the intended output is a head of well-cared-for curls.

(I should note that this way of understanding conceptual analysis is not universally endorsed. Peter Strawson, for example, endorses instead an alternative conception of the kind of analysis that philosophers do of concepts, which takes the model of "an elaborate network, a system, of connected items, concepts, such that the function of each item, each concept, could from the philosophical point of view be properly understood only by grasping its connection with others, its place in the system" (1992: 19). On this understanding, the aim of the method is not to break down the concept of interest into simpler parts, nor is circularity an objection: "there would be no reason to be worried if, in the process of tracing connections from one point to another of the network, we find ourselves returning to, or passing through, our starting point" (19). But even though Strawson wants to retain the name 'conceptual analysis' for this method – "since it is consecrated by usage" (19) – the method Strawson here describes is quite different from the way that 'conceptual analysis' is standardly understood, such that it is not plausible that he is talking about the same method. Indeed, other philosophers use different names for Strawson's method, for example Christoph Kelp calls it "network analysis" (2021a: 5). Henceforth, when I talk about 'conceptual analysis', I mean conceptual analysis as traditionally understood; that is, I mean reductive conceptual analysis.)

I have said that a philosophical methodology is a collection of philosophical methods. However, it might sometimes be difficult to distinguish a method from a methodology. The hard cases will be ones where a method can take different forms. For example, I argue in the next chapter that the method of conceptual genealogy, which Edward Craig (1990) applies to *knowledge* and Bernard Williams (2002) to *truthfulness*, is a kind of conceptual reverse-engineering, as is Miranda Fricker's

(2016, 2019) method of paradigm-based explanation; but also that conceptual reverse-engineering is a method in itself. On this picture, conceptual reverse-engineering is a broad method, and different ways of doing conceptual reverse-engineering can be thought of as its sub-methods. But why not say instead that conceptual reverse-engineering is a methodology, and genealogy and paradigm-based explanation two methods under the methodology? Similarly, I argue that conceptual engineering is a method, but that other methods, such as explication, are ways of doing conceptual engineering. Again, why not say that conceptual engineering is a methodology, and explication one method under the methodology?

I want to retain 'methodology' for broader collections of methods than these. Perhaps the boundary between methodologies and methods will be vague, but I will attempt to uphold at least a rough distinction between the two. Let's say that a philosophical method is a way of going about some particular philosophical inquiry, for example, answering the question 'What is knowledge?' or explaining the value of the concept *justice*; while a philosophical methodology is the collection of all the ways that (some group of) philosophers might go about various different philosophical inquiries. A method may have sub-methods, or it may not; we might say that some methods are determinable, while some are determinate. For example, I will argue in §3.2 of the next chapter that explication is a sub-method of conceptual engineering; then conceptual engineering is determinable, while explication is determinate. But a method's being determinable, or having sub-methods, isn't enough to make a method a methodology: so long as the relevant sub-methods are all ways of going about the same kind of philosophical inquiry (answering 'What is X?' questions, or explaining the value of X, for example), then this collection of (sub-)methods constitutes a method, not a methodology.

In what follows, I talk about concepts having 'extensions' and 'intensions'. The extension of a concept is the set of all and only the objects to which that concept applies in a possible world. For example, the concept *cat*, relative to the actual world, applies to the set of all and only cats that exist in the actual world. The intension of a concept is a function from a world to an extension. The intension of *cat* is a function that takes us, for any possible world, to the set of all and only cats in that world. Less precisely, but perhaps easier to grasp, we can characterise the intension of a concept as something like a recipe for determining what things the concept would apply to for any possible way that the world could be. To grasp the intension of a concept is to grasp

what it would take for some object to fall under the concept. To grasp its (actual) extension is to grasp what objects actually do fall under the concept.

2. What are we doing when we put function first?

In this section, I introduce the terminology of *function-first approaches* to concepts. In §2.1, I credit this terminology to Michael Hannon and argue that, as Hannon understands the function-first approach, it is a method for generating a set of conditions that are satisfied in typical cases in which the concept applies. I compare Hannon's method to Craig's method of hypothetical genealogy, and argue that the two are distinct methods with important similarities. I argue that this gives us reason to broaden our conception of the function-first approach so that it is not identical to Hannon's method. I introduce another method with a similar motivation, Sally Haslanger's ameliorative method, which is dissimilar to Hannon and Craig's methods in aim and output. I argue that Haslanger's method should also be considered a function-first approach, thus that our conception of the function-first approach should be broadened even further. I suggest that what unites these methods under the banner of 'function-first approaches' is that they are all useful for a philosopher who thinks that we can better understand, evaluate, and otherwise theorise about concepts through an exploration of their functions. I call this the 'function-first insight'. In §2.2, I provide a more general argument for thinking of the function-first approach as a methodology: that the function-first insight motivates many different inquiries, which require different methods. I suggest that the function-first approach should be thought of as the collection of methods that a function-first philosopher could make use of for these different inquiries. I illuminate the function-first methodology by comparing it to two more familiar methodologies: intension-first and extension-first approaches to concepts. It will emerge that what is distinctive about the function-first methodology is that each of the methods within the methodology prioritise the function of the concept of interest over its intension and extension in theorising about that concept.

2.1. Function-first as a method?

Hannon is to be credited with introducing the terminology of 'function-first' approaches to concepts in his book *What is the point of knowledge? A function-first epistemology*. He writes that the aim of his book is to "shed light on the nature and importance of

knowledge by investigating what our epistemic words, concepts, norms, and practices are *for*", and calls this approach to the study of our epistemic life "function-first epistemology" (2019: 2). A function-first epistemologist "seeks to explain the nature and value of an epistemic concept, norm or practice by reflecting on its functions or purposes" (12). She will ask questions like, "Why do humans speak and think in terms of 'knowing,' 'understanding,' and 'rationality'?" and "What might life be like without our current practices of epistemic evaluation?" (12).

Other philosophers use different names for the same approach. John Greco and David Henderson use "purposeful epistemology" to name a "methodology in epistemology", whereby an epistemologist "consider[s] the point(s) or purpose(s) of our epistemic evaluation, and … pursue[s] epistemological theory in light of what can be ascertained about such matters" (2015: 1). But the approach is not available only to epistemologists. A philosopher could take a function-first approach to a non-epistemic concept, norm, and so on, for example a moral concept like *good* or a semantic concept like *truth*. Georgi Gardiner calls this approach to concepts in general the "teleological approach". She writes that a philosopher who takes a teleological approach to a concept asks questions like: "What is the purpose of the concept? What role has it played in the past? If we imagine a society without the concept, why would they feel the need to invent it?" (2015: 31). The idea underpinning the approach is that we can better understand a concept by thinking in terms of what it does for us: "examining the function of the concept illuminates the contours of the concept itself" (31). We can take the 'teleological approach' and the 'function-first approach' to pick out the same general approach to concepts, and 'purposeful epistemology' and 'function-first epistemology' to name the application of this approach to epistemic concepts.

Though Hannon is responsible for introducing the terminology, it is not entirely clear which he takes the function-first approach to be: a methodology or a method. He sometimes calls it a methodology, for example writing that "the main goal of this book is to make use of a fairly new methodology in epistemology that I call *function-first epistemology*" (2019: 12). But in other places he calls it a method, for example: "[y]ou might endorse the method of function-first epistemology while rejecting my hypothesis about the function of the concept of knowledge" (3); "people often find it unclear what this method really amounts to" (3); "[m]y view is deeply indebted to Craig, but the method of function-first goes beyond his proposal" (12). This brief sampling of

Hannon's writing suggests that he doesn't put much stock in the method/methodology distinction.

In practice, I argue, Hannon treats the function-first approach as a method – a way of undertaking a philosophical inquiry – rather than a methodology. In his most explicit characterisation of the approach, he writes that it involves "three broad steps" (4). The first step is to offer a *prima facie* plausible hypothesis about the functional role that one's concept of interest plays in human life. The second step is sketching what a concept that plays this role must be like: what will be required to satisfy this concept. The third step is to compare the sketch of the concept from the second step with our ordinary judgements and intuitions about the concept of interest (4-5). For example, will a concept whose intension is specified as in step two pick out all and only things that we intuitively take to be in the extension of the concept? Insofar as these two concepts are relevantly similar, one can take the sketch of the concept from the second step to constitute the "core" of the concept of interest (18), a notion that Hannon takes from Craig. The core of the concept is a "description of a prototypical instance" of the concept's application (Craig 1990: 33): a set of conditions for the satisfaction of the concept that hold in typical cases, but not necessarily all cases. The core of the concept shows the concept at its most functional, which in turn explains why this community has the concept in the first place: they have the concept because it fulfils this function. As I have characterised the distinction between methods and methodologies, on Hannon's characterisation of the function-first approach it is a method: a way of going about some specific philosophical inquiry; specifically, identifying the functional core of some philosophically interesting concept.

Understanding Hannon's characterisation of the function-first approach as a method, rather than a methodology, makes sense of his lengthy comparison of the function-first approach with conceptual analysis, which is also a method (see Hannon 2019: 15-21). Hannon argues that his function-first method has an important advantage over conceptual analysis: it is relatively invulnerable to counterexamples. Conceptual analysis aims to provide necessary and sufficient conditions for the satisfaction of a concept. Then any genuine counterexample to these conditions – any instance of an item which ought to fall under the extension of the concept but which does not satisfy the proposed necessary conditions, or which satisfies the conditions but ought not fall under the extension – shows the analysis to be incorrect, and thus constitutes sufficient reason to reject the analysis. This is the case even if the

counterexample is "freakish", to use Craig's term (1990: 14): if it is a counterexample that will arise only rarely (or even never in the actual world), or only under bizarre circumstances. As the function-first method does not aim to generate necessary and sufficient conditions for a concept's satisfaction, it is not so vulnerable to counterexample. If a counterexample, even though genuine, is sufficiently freakish, it needn't undermine one's claim to having correctly described the core of the concept.

We can break down this claimed advantage into two aspects. The first concerns the improbability of finding satisfactory analyses of many philosophically interesting concepts, including *knowledge*, which is Hannon and Craig's concept of interest. Hannon writes that "all attempts to analyze knowledge" in terms of necessary and sufficient conditions "have succumbed to a pattern of counterexamples", which inspires a pessimistic meta-induction on his part: "we have good reason to think that whatever the next proposed analysis might be, sooner or later a counterexample will emerge", thus the analysis will be shown to be false (2019: 17). So if we want a satisfactory account of *knowledge* (or whatever philosophically interesting concept has so far eluded conceptual analysis), we have reason to try other methods. In particular, we should try the function-first method, which won't face the same problem. The second aspect is that the function-first method captures features of concepts that aren't necessary for a concept's satisfaction, as the concept can be satisfied in sufficiently unusual circumstances without these features being present, but which nevertheless are centrally important for understanding the concept. Hannon, following Craig (1990: 14), gives the example of belief to the concept *knowledge*. Any case in which a subject intuitively has knowledge without having belief, such as the case of the unconfident history quiz participant offered by Colin Radford (1966), will mean that belief drops out of a conceptual analysis of *knowledge*, "despite the fact that people generally believe what they know" (Hannon 2019: 18). But on the function-first approach to *knowledge*, a feature that is "extremely important, but not obviously necessary," such as belief, can still "play a crucial role in our epistemological theorizing" (18): though belief won't be necessary for *knowledge*, it will be present in all typical cases of *knowledge*, and thus can partly constitute the core of the concept.

A second advantage Hannon claims for his method over conceptual analysis is that his method not only illuminates the nature of the concept under consideration, but explains the value of having the concept. Even a satisfactory conceptual analysis of, for example, *knowledge* "would leave unanswered some significant questions in our

epistemological theorizing" (20). Conceptual analysis cannot tell us why some conditions are necessary and jointly sufficient for satisfying *knowledge*. Nor can it explain why *knowledge* "plays a crucial role in our life" (21), as evidenced by 'know' being one of the ten most commonly used verbs in English (Davies and Gardner 2010) and the first cognitive verb that children learn (Shatz, Wellman and Silber 1983), as well as finding a meaning-equivalent in all known languages (Goddard 2010). In contrast, the function-first method "tackles [these questions] head on" (Hannon 2019: 21): it explains why *knowledge* has the features it does and why it plays a crucial role in our lives by appeal to the function that the concept serves, which is to meet a need that all creatures like us, in environments like ours, will have. (The content of Hannon's function-first approach to *knowledge*, including his identification of its function, will be a topic of Chapter 3.)

Hannon's function-first method is inspired by the method developed by Craig in his book *Knowledge and the state of nature*. Craig is standardly credited with popularising the function-first approach (Greco and Henderson 2015: 2; Gardiner: 2015: 31; Kusch and McKenna 2020: 1057). But Craig's and Hannon's methods are distinct. Craig's method begins by positing "some prima facie hypothesis about what the concept … does for us, what conditions would govern its application" (1990: 2). This resembles the first two steps of Hannon's method. But Craig's method then diverges from Hannon's, as it involves constructing a hypothetical genealogy of one's concept of interest: an imagined narrative about why a concept similar to the concept of interest would develop in a state of nature, consisting of creatures with the same basic needs as us but who lack the concept, and why that concept would change, in response to various practical pressures these creatures could be predicted to face, to look more like the concept that we recognise as ours. The idea is that, if the concept that emerges from this process closely resembles our concept in relevant ways, we have good reason to think that our concept serves the same need for us as the concept that arises in the state of nature serves for the creatures therein. In contrast, Hannon's function-first method "is not a genealogy and makes no reference to a fictional state of nature" (2019: 2).

Though Hannon and Craig's methods are distinct, they resemble each other in important ways. First, the motivation for both methods is the same. Craig, like Hannon, notes that traditional conceptual analysis has so far failed to deliver a satisfactory account of *knowledge*, so argues that it is "worthwhile to try to think of another

approach" (1990: 1). And Craig, again like Hannon, argues that his method is able to answer questions that conceptual analysis cannot get a grip on, such as "why has a concept demarcated by those conditions enjoyed such widespread use?" (2). Craig's hypothetical genealogy can make plausible that the concept of interest "answers to some very general needs of human life and thought" (2). But conceptual analysis is silent on the question why the concept should have the necessary and sufficient conditions that application of the method generates. Second, the output of Craig and Hannon's methods is the same: both methods generate a set of conditions that characterise the "core of the concept" (1990: 33): the conditions that obtain in the typical case, but which need not hold in all instances in which the concept applies. Consequently, both methods have the same advantage over conceptual analysis of being relatively invulnerable to counterexample.

That Hannon and Craig develop different methods for describing and explaining concepts suggests that we should broaden our understanding of the function-first approach so that it is not identical to any one method. Their methods are similar in motivation, aim and output. Crucially, both put function 'first' in that the methods begin by hypothesising about the function of a concept, and explicit accounts of the intension and extension of the concept come later; and the adequacy of an account of the concept's intension and extension is determined by how well a concept with that proposed intension and extension would serve the hypothesised function. In the next chapter, I argue that Hannon and Craig's methods are both sub-methods of one broad method: conceptual reverse-engineering. Then broadening the function-first approach to cover both methods is not much of a broadening at all. In particular, the function-first approach might be identical to conceptual reverse-engineering. However, I argue now that this is not the case, for another method with a different aim and output should count as function-first, a method that is not a kind of conceptual reverse-engineering: Sally Haslanger's ameliorative method.

Haslanger takes what she calls an "analytical approach" to the concept *knowledge* (1999: 467), as well as to gender and race concepts like *woman*, *man*, *white*, *black* (2000: 34). In more recent work, she calls this an "ameliorative approach" (2012: 386), noting that the earlier terminology is confusing, or at least insufficiently illuminating, given that "'analytical' is commonly used to characterise Anglo-American philosophy in general" (385, fn. 5). Haslanger describes her ameliorative approach as a two-step method for inquiring into the question 'What is X?' for some concept X (386).

In the first step, one asks "what is the point" of the concept in question: "what work does it, or (better) could it, do for us?" (1999: 466). In the second step, one "consider[s] what concept would best accomplish this work" (467). So far, this sounds very much like Hannon's method. However, Haslanger's ameliorative approach is very different to both Hannon and Craig's methods in its aim and output. Where it is found that one's concept of interest is not serving its legitimate function(s) as well as it should, Haslanger's method aims to *improve* the concept to better serve its functions. In contrast, Hannon and Craig's methods aim to accurately *describe* the core of the concept of interest, and *explain* why it is as it is. The output of Haslanger's method is a potentially improved concept to that with which one began, given the work to which one wants to put the concept. In contrast, the output of Craig and Hannon's methods is a description of the core of the concept of interest: the set of conditions that are satisfied in typical instances where the concept applies.

To take Haslanger's ameliorative approach to a concept is to investigate what functions that concept serves, or could serve; to ask how well it serves those functions; and if the concept does not serve those functions as well as it could, to revise the concept. She writes:

> On an analytical approach the task is not simply to explicate our ordinary concept of X … instead we ask what our purpose is in having the concept of X, whether this purpose is well-conceived, and what concept (or concepts) would serve our well-conceived purpose(s) – assuming there to be at least one – best. … [T]his approach is quite comfortable with the result that we must revise – perhaps even radically – our ordinary concepts and classifications of things. (1999: 467-8)

Using Haslanger's method, if it is discovered that a concept like *knowledge* does not serve its legitimate purpose(s) as well as it could, then it should be revised – its intension and extension changed – so that it better serves its purpose(s). Then an inquiry into *knowledge* undertaken using Haslanger's ameliorative method might well end up an instance of what Davide Fassio and Robin McKenna call "revisionary epistemology": an approach to an epistemological concept that claims that the concept is not as it ought to be, and because of this it is in need of revision (2015: 755-6).

In contrast, Hannon and Craig's methods both permit only minor concept revision. The primary aims of their methods are, first, to accurately describe the core

of the concept of interest, and second, to explain why the concept is as it is by appeal to its function. They permit that their descriptions of the core of the concept might differ slightly from the "intuitive concept" that we find in pre-theoretical, "ordinary usage" (Craig 1990: 2). But where there are such differences, there must be "some special and especially plausible explanation of the mismatch". Without this explanation, "our original hypothesis about the role that the concept plays in everyday life would of course be the first casualty" (2; see also Hannon 2019: 14). Inquiries into *knowledge* or other epistemic concepts undertaken using Hannon's or Craig's methods thus will only be instances of revisionary epistemology under special circumstances. Their methods are, in any case, much less revisionary than is Haslanger's.

Although Haslanger's method differs in aims and outputs to Craig and Hannon's, it is motivated by the same insight: that we can better understand and evaluate or concepts, and more effectively improve our conceptual world, by thinking about what our concepts do, or could do, for us. Call this the 'function-first insight'. I suggest that we should think of any philosopher whose project is guided by this insight as a function-first philosopher, and any method that such a philosopher might make use of in her inquiries as a method falling under the function-first approach. Haslanger, Hannon and Craig are all guided by this insight, thus their methods are function-first methods. This is already reason to reject the identification of the function-first approach with any one method. But we can now make a more general argument: taking seriously the function-first insight motivates a number of different inquiries, which require the use of different methods; we should think of the function-first approach as the collection of all these methods; therefore, we should think of the function-first approach as a methodology.

2.2. Function-first as a methodology

The function-first insight tells us that we can better understand, evaluate, ameliorate, and generally theorise about concepts through explorations of their functions. This insight does not suggest that we should embark on any single philosophical inquiry using any particular method. Rather, taking seriously this insight motivates a number of different inquiries, and these different inquiries call out for the use of different methods. For example, a philosopher who is attempting to accurately describe the intension and extension of a concept who takes seriously the function-first insight might need to begin her project by finding out the function of the concept. For this

project, the method of conceptual reverse-engineering, whereby one reconstructs the practical problems to which the concept offers a salient solution, might be useful. (I have more to say about conceptual reverse-engineering in Chapter 2, §2.) But as Matthieu Queloz notes, a function-first philosopher might not want to find out the function of her concept of interest, either because she already knows its function, or because she is interested in what function the concept *should* serve (2021: 45). This philosopher would embark on different inquiries, which would require different methods. She might be interested in whether the function she knows the concept to have is one that users of the concept endorse as valuable; in that case, she might borrow methods from experimental philosophy, such as surveying ordinary language-users, to answer this question. Or she might be interested in changing the concept to better serve the function it should have; in this case, the method of conceptual engineering will be useful.

As a function-first philosopher might embark on a number of different inquiries, for which she must make use of different methods, it is too limiting to identify the function-first approach with just one method. Rather, I suggest we should think of the function-first approach as a methodology, consisting of all the methods that a function-first philosopher might use. The function-first philosopher works on concepts; call philosophers who work on concepts 'conceptual philosophers', and philosophy whose subject matter is conceptual 'conceptual philosophy'.[1] What is distinctive about the collection of methods that makes up the function-first approach? In this section, I will answer this question by contrasting the function-first approach with two other methodologies: intension-first and extension-first approaches to concepts.

I borrow the terminology of 'intension-first' and 'extension-first' approaches from Queloz and Gardiner. Queloz's way of distinguishing between the two kinds of

1 Williamson (2007) uses "conceptual philosophy" to name the work of philosophers who accept the first two of Dummett's tenets characterising analytic philosophy: "first, that the goal of philosophy is the analysis of the structure of thought; secondly, that the study of thought is to be sharply distinguished from the psychological process of thinking" (1978: 458). This is not what I mean by 'conceptual philosophy'. I am not trying to capture an evaluative position in my use of this term. I intend it to be a purely descriptive label: a conceptual philosopher is a philosopher who works on concepts. She needn't have any view on the proper subject matter of philosophy, such as that philosophy's proper subject matter is limited to concepts (what Williamson calls "absolute idealism about the subject matter of philosophy" (2007: 14)), nor about the goal of philosophy.

approaches appeals to direction of explanation: intension-first approaches to concepts take facts about a concept's intension to explain its extension; while extension-first approaches do the opposite, using the concept's extension to specify its intension. Gardiner, on the other hand, makes the distinction in terms of intuitive access: intension-first approaches to concepts assume that we have intuitive access to the intensions of our concepts, that is, to what it takes for an object to fall under a concept, in any given case; while extension-first approaches assume that we have intuitive access to the extensions of our concepts, that is, to what objects fall under our concepts. I will outline Queloz and Gardiner's taxonomies in turn, before offering my own account, according to which what is distinctive about the function-first methodology is that it includes only methods that, in some sense to be unpacked (§2.3), give priority to the functions of concepts over their intensions and extensions in theorising about the concept.

But first, I want to note that Queloz and Gardiner are not the first philosophers to distinguish philosophical views along roughly these lines. Rather, their taxonomizing here is reminiscent of Roderick Chisholm's writing on the problem of the criterion. Chisholm introduces the problem by distinguishing two questions:

a) What do we know? i.e. What is the *extent* of our knowledge?
b) How are we to decide, in any particular case, whether we know? i.e. What are our *criteria* for knowledge? (1977: 6)

If we know the answer to either one of these questions, then we could use this to answer the other: by specifying the criteria of knowledge, we would have a way of deciding how far our knowledge extends; or by knowing how far our knowledge extends, we may be able to formulate criteria that demarcate those things we know from those that we don't. But do we have an answer to either? This is the problem of the criterion.

Chisholm calls those philosophers who think that we have an answer to (a), which we can use to answer (b), "particularists"; and those philosophers who think that we have an answer to (b), which we can use to answer (a), "generalists" or "methodists" (1977: 7). This is similar to the extension-first/intension-first distinction that Queloz and Gardiner make. Particularists try to derive our criteria for knowledge from the extent of our knowledge (i.e., the extension of *knowledge*), thus are akin to extension-first philosophers. Methodists try to determine what we know via general criteria for how to decide, for any case, whether we know, thus are akin to intension-first philosophers. So although the terminology of 'extension-first' and 'intension-first'

approaches to concepts is novel, the claim that some philosophers think that we have special epistemic access to either intensions or extensions of our concepts, or that we can explain intensions in terms of extensions or *vice versa*, is not new.

For Queloz, the difference between intension-first and extension-first approaches to concepts is a difference in direction of explanation. An intension-first approach "focuses first on identifying the concept's intension in order to then take it as a basis for explaining the concept's extension", while extension-first approaches "seek to make sense of why the intension is as it is on the basis of a prior grasp of the nature of the extension" (2021: 44). The extension of the concept is, as I defined it in §1.1, "the set of all cases to which the concept applies" (25). Queloz understands the intension of a concept as that in virtue of which "we can say, for most situations and with some assurance, whether the concept applies in it or not." This may take the form of a "strict definition of the concept of X – a definition in terms of individually necessary and jointly sufficient conditions for something to be X" (25). In both intension-first and extension-first approaches to concepts, the guiding question is the same: What is X? (For example, what is knowledge? What is justice? What is freedom?) But there is a "reversal of explanatory direction" (44) from intension-first to extension-first approaches: in intension-first approaches, that all and only these objects fall under the concept X is explained in terms of what it takes for something to be X; in extension-first approaches, what it takes for something to be X is explained in terms of the set of objects that falls under the concept X.

To make this clearer, Queloz offers some examples. For Queloz, intension-first approaches are "paradigmatically exemplified" by conceptual analysis, "which aims to provide an explicit intension to be measured against the intuitive extension in the hope that it will explain why the extension has the boundaries it has" (43-4). The aim of conceptual analysis (at least as traditionally understood; see §1.1) is to articulate individually necessary and jointly sufficient conditions for an object's falling under the concept. This is a way of specifying the intension of the concept. That the intension is as it is – that is, that these conditions are individually necessary and jointly sufficient – is then taken to explain, for any token instantiation of the concept, why it falls under the concept. For example, a conceptual analysis of *knowledge* might articulate the following as individually necessary and jointly sufficient conditions for satisfying the concept: a subject S's relationship to a proposition P is a token of the concept *knowledge* iff S believes that P, P is true, and S is justified in believing that P. Then for

any given case in which the concept applies, this will be because these conditions are met. Thus conceptual analysis is, on Queloz's way of marking the distinction, an intension-first approach to concepts.

Note that Queloz doesn't necessarily understand the method of conceptual analysis as *beginning* with the articulation of necessary and sufficient conditions for an object to fall under the concept X. Temporally, a theorist could begin her project by sketching an intuitive extension for the concept, or at least by gathering up a good few cases. Once this intuitive extension has been sketched, she can "provide an explicit intension to be measured against the intuitive extension" (43-4). For example, in taking this approach to *knowledge*, a theorist might first build up a stock of cases that are part of the concept's intuitive extension: that they know their own name, that someone who sees that it's raining knows that it's raining, that someone can come to know that Calamity Jane was born on May 1st 1852 by watching a BBC documentary about her, and so on. But the theorist will then attempt to articulate individually necessary and jointly sufficient conditions that apply in all these cases, and she will take the application of these conditions to be what explains why all these subjects have knowledge. What matters is that the intension explains why the extension is as it is. In traditional conceptual analysis, Queloz's thought goes, the concept X applies to all and only the objects to which it does apply because those objects satisfy some set of necessary and sufficient conditions.

Queloz holds that philosophers take extension-first approaches to natural kind concepts. The extension of a natural kind concept, Queloz argues, explains why it has the intension that it does. All the items that fall under the concept's extension are "objectively unified … as a result of homeostatic mechanisms, shared chemical structures or reproductive chains" (2021: 44). That these objects are unified in this way is sufficient for their falling under the extension of the natural kind concept. This, in turn, explains why the concept's intension is as it is: namely, because this allows the concept to track the natural kind. For example, all and only the water in the world is objectively unified by virtue of having the chemical composition H_2O. This is sufficient for all and only the water in the world to fall under the extension of the concept *water*. The intension of *water* is specified as all and only stuff with the chemical composition H_2O, and it is so specified as this is what enables us to be "suitably sensitive" to the objectively unified stuff, water, that precedes the concept *water* (44). Thus the

extension of the concept explains its intension, and this is an extension-first approach to a concept, on Queloz's account.

Function-first approaches are likewise distinguished, for Queloz, by what explains what. Function-first approaches are guided by what Queloz calls the "Pragmatic Question": "Why do we have the concept of X? What does it *do* for us? What is the value of living by a concept that delineates just *this* extension by means of just *that* intension?" (44). The concept's function is to serve as the basis for an explanation of why the concept has the intension and extension that it has. As an example of a function-first approach, Queloz gives conceptual reverse-engineering: the method by which a theorist reconstructs the needs, both practical and theoretical, which a concept serves for a community in order to understand and evaluate the concept. This is a kind of function-first approach "because it primarily seeks to identify the function of a concept in order thereby to explain why we have that concept" (45).

Queloz argues that the three approaches are not mutually exclusive in a strong sense, on which they cannot complement or inform each other. One can take different approaches to a concept at different times in one's broad pursuit of understanding the concept. For example, before embarking on an extension-first approach, one could "profitably consult extant attempts to analyse the concept at issue" – recall that conceptual analysis for Queloz is an intension-first approach – as these analyses "may helpfully broaden or sharpen one's sense of the various properties that the concept might be tracking" (45). Similarly, one's function-first approach might be informed by findings from extension-first approaches to the concept: "understanding that a concept's extension is unified by certain natural principles might yield a clue as to what function the concept performs by giving one an independent grip on the projectability of its extension's properties from one sighting to the next." But one cannot pursue any combination of the approaches simultaneously, "since they bestow explanatory priority on the intension, extension and function of concepts respectively, and only one corner of that triangle can form the apex at any given time" (45).

Unlike Queloz, Gardiner does not draw the distinction between intension-, extension- and function-first ('teleological') approaches to concepts in terms of explanation. Rather, these three approaches are distinguished by what they assume we have "intuitive access" to – which, for Gardiner, means what we are able to make "reliable pre-theoretical judgements" about – and as such can take as a starting point in our philosophical theorising (2015: 32). The extension-first approach assumes that

we have intuitive access to the extensions of our concepts, for example to the extension of *knowledge*, where this means that we make "reliable pre-theoretical judgments about when the proposition … *S knows that p*, is correct" (32). Proposed analyses of concepts are tested by how well they align with this intuitive extension.

Gardiner offers experimental philosophy as an example of an extension-first approach. One method of experimental philosophy is to survey concept-users to collect data on their judgements about whether the concept applies in test cases. This data can then be compared to proposed intensions of the concept. Insofar as those proposed intensions demarcate extensions that exclude cases from the extension that the concept-users judged to be in the extension, or include cases that the concept-users judged to be outside its extension, this is taken to undermine the proposed intension; insofar as the demarcated extension and the surveyed concept-users' intuitive extension align, this is reason to accept the proposed intension. For example, Adam Feltz and Chris Zarpentine (2010) conducted surveys on users of the concept *knowledge* to test whether they applied *knowledge* in high-stakes cases, such as Keith DeRose's (2002) bank case (though they use Jason Stanley's (2005: 5) formulation of the bank case in their survey). They found that concept-users were largely inclined to attribute knowledge to the subject in the bank case. This finding is taken to undermine stakes-sensitive theories of knowledge, according to which whether S knows P is determined partly by what is at stake for S in the context (2010: 696-7).

Gardiner also classes more traditional philosophical methods as extension-first, for example, the method of testing proposed theories of concepts by constructing thought experiments, and seeing whether our intuitions about whether the concept applies in the thought experiment align with the extension of the concept that the theory demarcates. Where there is a mismatch between our judgement in the thought experiment and the extension that the theory demarcates, we must either reject the intuition and offer an error-theory to explain this mistaken intuition, or sanction the intuition and amend the target theory accordingly. Whether an intuition about the concept's extension can be explained away or whether it must be sanctioned will depend on whether and to what extent it is "a confident and unambiguous judgment about a relatively central case", versus a "less confident" judgement about a "more peripheral or obscure" case (32). For these latter cases, it will be "easier and less costly" to offer an error-theory for the judgement, while for the former kind of cases, the theory must be amended (32). In this way, the theorist attempts to achieve

"reflective equilibrium … between intuitive judgments about particular cases and general principles" (33); but this counts as extension-first, for Gardiner, as it is the extension of the concept – which cases fall under the concept – to which we have reliable, pre-theoretical (i.e., intuitive) access, while the epistemic access we have to general principles is not pre-theoretical.

The intension-first approach, on the other hand, assumes that we have intuitive access to the intensions of our concepts, where this means that we make reliable pre-theoretic judgements about what it would take for a concept to apply in some instance, rather than about what are the particular instances in which the concept does or does not apply. This intuitive access to the intension of a concept then informs the account one gives of the concept. For example, an intension-first approach to a concept might begin by offering a number of "platitudes" about what it takes for the concept to apply to an object, which a theorist accesses "directly from [her] understanding of the concept, without proceeding via the intuitive extension" (33). The theorist can then build these platitudes into an explicit account of the concept of interest.

Function-first ('teleological') approaches, in contrast, do not assume that we have intuitive access to either the intension or the extension of our concepts. These approaches aim to illuminate our concepts "by asking what the *point* of that concept is; what *purpose* it fulfills, what *need* it meets, what *function* it has, or what *role* it characteristically plays", with the guiding idea being that "by focusing on the needs the … concept fulfills we illuminate the nature of the concept" (35). This approach does not rely on our intuitions about what it would take for the concept to apply in any given case, or whether the concept does or does not apply in particular cases. Rather, the function-first approach "begins by looking at our socio-cognitive economy, and determining what concepts might be useful, rather than starting with any particular claims about a concept's contours, such as particular instantiations or intensions" (35). Gardiner gives Craig's hypothetical genealogy as an example of a teleological approach to a concept (39-40).

Queloz writes that he and Gardiner propose "similar organization[s]" of these three approaches (2021: 44, fn. 40). However, I argue that we can see, just from the preceding discussion, that Queloz and Gardiner do not mean the same thing by 'intension-first' and 'extension-first' approaches to concepts. Recall that I said that, for Queloz, an intension-first approach to a concept can begin by building up an intuitive extension of that concept. Not so for Gardiner. For Gardiner, whether an approach to a

concept is intension-first or extension-first is precisely a matter of what it is assumed we have intuitive access to: intension or extension. On Queloz's picture, extension-first approaches can assume that we have intuitive access, in Gardiner's sense (that is, that we make reliable pre-theoretic judgements), to the intensions of our concepts, and *vice versa*. What matters is that facts about a concept's extension explain its intension, in extension-first approaches; while facts about a concept's intension explain its extension, in intension-first approaches.

In this way, Gardiner's intension-first/extension-first distinction is closer to Chisholm's methodist/particularist distinction. Intension-firsters, on Gardiner's picture, build their accounts of their concept of interest out of "intuitive judgments about the intension of a concept" – what it would take for the concept to apply in any given case – "without proceeding via the intuitive extension" (2015: 33); while for Chisholm, methodists think that we can "formulate a criterion of knowledge without appeal to *any* instances of knowledge" (1977: 6-7). Extension-firsters think that we can "mine our judgments about the intuitive extension", then use these judgements to "provide an account of the concept" (32-3); in the same way that particularlists, for Chisholm, think that we can "identify instances of knowing without applying any criteria of knowing" and from these instances can extract "criteria enabling us to mark off the things we do know from those that we do not" (1977: 6-7).

We can see the difference between Queloz and Gardiner's ways of marking the distinction between intension-first and extension-first approaches through their conflicting treatment of a method that offers platitudes about a concept's intension as the starting point for theorising about a concept. Gardiner gives an example of this method applied to the concept *knowledge*. A theorist begins by listing a number of "platitudes" about, by which Gardiner means "intuitive judgements about the intension" of, *knowledge*. For example, "the value platitude (knowledge has value), the anti-skeptical platitude (we have some knowledge), the ability platitude (knowledge is, at least in part, to the credit of the knower), and the anti-luck platitude (knowledge enjoys some kind of modal stability)" (2015: 33). These intuitive judgements about the intension of *knowledge* then "significantly inform [her] theory" of *knowledge* (33): an intension for *knowledge* can be specified after reflection on these judgements, which can then be used to demarcate the bounds of its extension. For Gardiner, this is an intension-first approach. Indeed, it is for her "*the* intension-first approach" (33, my emphasis).

But for Queloz, it is not at all clear that this method would count as intension-first. Queloz writes of a method that underpins "attempts to understand our concepts via metaphysical inquiries into the nature of the things we speak or think about" (2021: 44). This method "begin[s] with … statements we take to be true" about the concept of interest, and then "ask[s] what *makes* them true, or what their terms *refer to*" (44). For example, we might take this approach to a moral concept, like *wrongness*, by listing moral statements that involve the concept and that we take to be true, then asking what makes them true. "Mysterious truth-makers such as … moral facts then come under metaphysical scrutiny aiming to discover what … moral facts really are" (44). That is, we investigate whether there is, for example, any property in the world that *wrongness* picks out, which makes true those statements invoking *wrongness* that are true, and what this property is like. For Queloz, this is a quintessential extension-first method, as facts about the extension of the concept – that all objects therein have the property of wrongness, say – which explain the concept's intension: why the intension picks out those objects in this world.

Is this the same method that Gardiner discusses? It's not entirely clear. Queloz writes that this method "begin[s] with, say, mathematical, modal or moral statements we take to be true, and then asks what *makes* them true, or what their terms refer to" (44). He does not offer examples of what these statements might be. Moral statements (to continue with the example from the previous paragraph) can be more or less general. Some moral statements state general principles, for example, that it is wrong to treat others as mere means to an end (Kant 2012: 40). Others concern specific applications of moral concepts, for example, that eating meat is wrong. The kinds of intuitive judgements that make up Gardiner's "platitudes" are judgements about the truth of statements of the first kind, those that state general principles. Which kinds of statements is Queloz concerned with? Plausibly, both kinds. His idea might be that moral facts are posited to explain the totality of moral statements we take to be true, both those that state general principles and those that concern specific applications of moral concepts. That moral facts are as they are explains why these statements are true, when they are true. But this is enough to demonstrate conflict between Queloz and Gardiner's intension-first/extension-first distinctions. For insofar as a moral fact, say, is taken to explain a general moral statement, this is for Queloz an extension-first approach to a moral concept; but where the truth of that general statement is one we are supposed to have intuitive access to, this approach is, for Gardiner, intension-first,

regardless of the underlying metaphysics that are supposed to make true the statement (if it is true).

So Gardiner's and Queloz's accounts of the distinction between intension-first and extension-first approaches to concepts are not the same. Which is better? Both have issues. Gardiner's account is too limiting in its focus on intuitive access. But before making this point, a disambiguation: Gardiner understands having 'intuitive access' to something as being able to make "reliable, pre-theoretical judgements" about that thing (2015: 32). There are two ways of understanding 'pre-theoretic'. On the first, 'pre-theoretic' means prior to any theoretical (beyond folk theoretical) consideration of the matter at hand (for examples of 'pre-theoretic' used in this way, see Cappelen 2012: 62; Clarke 2010: 278; Comesaña 2020: 251). One's judgements about *knowledge*, for example, can be pre-theoretic in this sense so long as they are informed only by 'folk epistemology' and the conceptual resources that it supplies (Gerken 2017). On its second sense, 'pre-theoretic' means prior to endorsing some specific theory of the matter at hand (for examples of 'pre-theoretic' used in this way, see Ichikawa 2017: 155; McKenna 2017: 186; Strawson 1992: 10; Weinberg 2007: 338). 'Pre-theoretic' in this latter sense means something more akin to 'theory-independent' than 'temporally prior to theory' (Ichikawa explicitly paraphrases 'pre-theoretic' as 'theory-independent' (2017: 155)). Which sense does Gardiner intend when she writes, for example, that the judgement that *knowledge* "enjoys some modal stability" is pre-theoretic (2015: 33)?

It is not plausible that this judgement is 'pre-theoretic' in the first sense. Putting aside worries about whether epistemologists can have pre-theoretic judgements in this sense (perhaps if they try *really hard* they can get into a frame of mind wherein they draw only from the conceptual resources supplied to them by folk epistemology), insofar as it has become widely accepted within contemporary epistemology that *knowledge* involves modal stability, this is due to decades of debate, largely but not solely post-Gettier (Russell's stopped clock example was published in 1948), about why, exactly, justified true belief is insufficient for knowledge. Then charity demands that Gardiner be interpreted as meaning 'pre-theoretic' in the second sense. Then that we make "reliable, pre-theoretic judgements" about the intensions of our concepts, according to her intension-first approach, and about the extensions of our concepts, according to the extension-first approach, should be taken to mean that prior to (or

independently of) endorsing any theory of the concept of interest, we make reliable judgements about its intension or about its extension, respectively.

Because of its focus on pre-theoretic access, Gardiner's way of making the distinction between intension- and extension-first approaches to concepts does not leave room for certain important theoretical approaches we take to concepts. In particular, it leaves no room for our approach to natural kind concepts. Natural kind concepts are demarcated in terms of the natural kinds they track. It is not generally the case that we have reliable, pre-theoretic access (in the relevant sense) either to what objects fall under a natural kind concept, or to what it takes for something to fall under the concept; that is, to either natural kind concepts' extensions or intensions. For example, *gold* is a natural kind concept, which tracks the natural kind gold. Gold is a chemical element, so it is essential to something's being gold (and thereby falling under the natural kind concept *gold*) that its atoms have a particular atomic structure. But prior to (that is, independently of) endorsing this chemical theory of what gold is, and thus of what the intension and extension of *gold* are, one would not make reliable judgements about either its extension or intension. One who did not endorse the theory on which gold is a chemical element, to be identified by its atomic structure, would not chemically analyse objects to determine whether they are gold. Rather, they would try to identify gold by its sensible qualities: what it looks like, its hardness, and so on. This would lead to an account of the intension of *gold* as anything that looks a particular way in certain temperatures, that is hard to a particular degree, and so on; rather than as anything that has the correct atomic structure. As a consequence of this incorrect intension, one would assign objects to the extension of *gold* that are not part of its extension, such as objects composed of pyrite (i.e. 'fool's gold'). This is not how we approach natural kind concepts. Rather, our approach to natural kind concepts does not involve pre-theoretical access to intension or extension of those concepts. So our approach to natural kind concepts does not count as either intension- or extension-first, on Gardiner's account.

Gardiner may well not recognise this as an objection to her view. This is because she seems to think of the intension-first approach and the extension-first approach not as collections of methods – as methodologies – but as methods. She writes that "[t]he 'extension-first' method elucidates a concept using various cases to mine our judgments about the intuitive extension and to test the intuitive extension against a proposed claim or theory" (2015: 32), and thereby identifies the extension-first

approach with the broad method of testing our intuitions about the extension of a concept in various cases (this method includes both experimental philosophy and use of thought experiments as sub-methods). Similarly, in describing the method that begins by gathering platitudes about the intension of the concept, she states that she is discussing "*the* intension-first approach" (33, my emphasis), again suggesting that she identifies the intension-first approach with this singular method. As such, she might simply grant neither of these methods are ones that we can take use to generate accounts of natural kind concepts. But these are only two possible methods that philosophers can use, not two broad methodologies, and as such we needn't be able to classify all, or even many, philosophical methods under them.

If this is how Gardiner intends us to understand the intension-first and extension-first approaches – as methods, rather than as methodologies – then her account is not suitable for my purposes. In order to be relevantly comparable to the function-first approach, these approaches must be methodologies. But Queloz's taxonomy, with its focus on explanation, is also too limited to capture all relevantly similar methods under the banner of 'function-first approaches'. Recall that for Queloz, whether an approach is intension-, extension- or function-first is determined by what explains what: whether intension explains extension (intension-first), extension explains intension (extension-first), or function explains both. However not all philosophical methods, including those that should fall under the 'function-first' banner, aim to explain why a concept's intension or extension are as they are.

For example, Haslanger's ameliorative method applied to the concept *woman* does not aim to explain why *woman* has the intension 'is systematically subordinated along some dimension (economic, political, legal, social, etc.) and is marked as a target for this treatment by observed or imagined bodily features presumed to be evidence of a female's role in reproduction' (2000: 39) by appealing to this concept's function, which is, for Haslanger, to play a role in tackling injustice, for example by allowing us to identify and explain inequalities between males and females (36). Rather, Haslanger stipulates that this is the function she thinks *woman* ought to have, and argues that the concept could serve this function well if it had that intension. She is not explaining why the concept has this intension in terms of its function: she doesn't think that the concept *woman* does have this intension, nor that it serves this function. As Queloz fleshes out his taxonomy, Haslanger's ameliorative method falls through its cracks.

2.3. My characterisation of the function-first approach

The problem with Queloz's taxonomy is in his focus on explanatory priority. Not all projects that we want to call 'function-first' have explanatory aims: they do not all attempt to explain some fact or set of facts about a concept in terms of facts about the concept's function. Note that this is a problem for Queloz, too. He understands conceptual reverse-engineering as a "species within the genus of function-first approaches, because it primarily seeks to identify the function of a concept in order to thereby explain why we have the concept". But he argues that conceptual reverse-engineering is "only a species, however, because there are also function-first approaches that do not reverse-engineer at all, either because they already know the function or because they focus on the function a concept should serve" (2021: 45, my emphasis). Haslanger's ameliorative method is a species of the genus of function-first approaches that does not aim to explain anything in terms of the function of some concept. Rather, Haslanger has an ameliorative aim: to revise gender and race concepts so that they can serve functions that she thinks they ought to, because their serving those functions would be politically efficacious for feminists and anti-racists. Haslanger's method can be understood as treating the (potential) functions of concepts as in some sense having priority over their intensions and extensions, as any revision to intension and extension must serve the goal of enabling the concept to serve a particular function. But this is not explanatory priority. Indeed, we might think that, in Haslanger's ameliorative method, a concept's function still explanatorily depends on its intension. In order for the concept *woman*, for example, to have the function that Haslanger thinks it should, its intension must be revised in such-and-such a way. So in the counterfactual situation in which *woman* does have this function, this is because it has that intension. This is precisely why we are to engage in the ameliorative method.

There is an easy fix here. To capture Haslanger's method (and other relevantly similar methods) as function-first approaches, we should replace Queloz's talk of *explanatory* priority with talk of priority *of some kind or other*. Where a method has an explanatory aim, the kind of priority required to bring the method under the function-first methodology is explanatory priority. For example, the method of conceptual reverse-engineering aims to explain why we have a concept with a particular intension and extension in terms of what that concept, with that intension and extension, does for us. That is, conceptual reverse-engineering explains why some concept's intension

and extension is as it is in terms of its function. This is a function-first method with an explanatory aim: facts about the concept's function explain facts about its intension and extension. Thus, the latter kinds of facts *explanatorily depend* on the former kind. But for a function-first method with an ameliorative aim, the kind of priority involved is what we might call *teleological* priority. The guiding aim in revising the concept has to do with the function the theorist wants it to serve. In this way, the (desired) function guides the ameliorative project. For example, in Haslanger's method, it is because we have the aim of revising the concept to have such-and-such a function that we revise its intension, and perhaps its extension too, in such-and-such a way.

Philosophical methods can have other aims than these explanatory and ameliorative aims just described. Some methods for conceptual philosophy aim to evaluate a concept without revising it, should it be found lacking. For example, Alexis Burgess and David Plunkett use "conceptual ethics" to name a "field" in philosophy that is concerned with "how one ought (or would do well) to think and talk" (2013: 1091). In more recent work, Plunkett and Cappelen describe conceptual ethics as "concern[ing] a range of normative and evaluative issues about thought, talk, and representation. Those include issues about what concepts we should use, ways in which concepts can be defective, what we should mean by our words, and when we should refrain from using certain words", as well as "evaluative issues about which concepts are better than others (and why)" (2020: 4). Plunkett and Cappelen explicitly contrast conceptual ethics with conceptual engineering, distinguishing the two by holding that only the latter "involve[s] trying to actually change conceptual or linguistic practices" (5). Given that I am understanding conceptual engineering as a method (albeit a broad one, which includes sub-methods such as explication), I will likewise understand conceptual ethics as a (presumably similarly broad) method, one which aims to evaluate concepts. We can evaluate a concept in terms of how well it serves some function (without revising the concept to better serve that function, should it be found lacking), and would thereby be engaged in a function-first project of conceptual ethics. This function-first project, we might say, treats the function of a concept as having *normative priority* over its intension and extension.

Yet other methods of conceptual philosophy have purely descriptive aims. For example, the aim of traditional (reductive) conceptual analysis is to generate a set of individually necessary and jointly sufficient conditions for an object to satisfy the concept of interest. This set of necessary and sufficient conditions constitutes the

intension of the concept. As noted in §1.1, a philosopher engaged in conceptual analysis will attempt to reach reflective equilibrium between extensional adequacy and the preservation of other theoretical virtues in her intension. But in any case, the method of traditional conceptual analysis is extension-first, in that the raw data to which the philosopher has observational access in constructing its intension is the concept's (intuitive) extension. In this way, we might say that a concept's extension is the *observandum* in conceptual analysis, and the priority given to extension over intension is that of *observational priority*. A function-first method with a descriptive aim would thus be one that gives observational priority to a concept's function over its intension and extension in trying to describe that concept.

Thus I suggest that we characterise the function-first approach as a methodology for conceptual philosophy that includes only methods that give priority, in the relevant way, to the function of the concept of interest over its intension and extension. (The intension-first approach will be a methodology for conceptual philosophy that includes only methods that give priority, in the relevant way, to the intension of the concept over its extension and function; the extension-first approach a methodology for conceptual philosophy that includes only methods that give priority, in the relevant way, to the extension of the concept over its intension and function.) For methods with descriptive aims, the relevant kind of priority is observational priority; for methods with explanatory aims, it is explanatory priority; for methods with evaluative aims, it is normative priority; and for methods with ameliorative aims, it is teleological priority. The function-first methodology will thus be a fairly multifarious collection of methods, united by function being treated as prior in some sense, but without any deep similarity between the methods.

An interesting result of this characterisation is that a method can properly be called function-first relative to one way of giving priority, but intension- or extension-first relative to another. Haslanger's ameliorative method is an example. Haslanger's method is function-first in that it bestows teleological priority over a concept's function over its intension and extension. But as already noted, it plausibly bestows explanatory priority on intension over function: it is only if the concept has such-and-such an intension that it would be able to serve such-and-such a function. That is, in the counterfactual situation where the concept does serve this function, that is because it has that intension. I am happy with this result. What it suggests to me is that a method is not exclusively intension-, extension-, or function-first, but that it falls under one of

these categories insofar as it can be used for inquiries with certain aims. What I mean by this will become clear in the next section.

2.4. Descriptive, explanatory, evaluative and ameliorative projects

Throughout this chapter, I have used the phrase 'philosophical projects', which I have so far left as an intuitive notion. I now want to say more about how this should be understood. A philosophical project is a particular inquiry undertaken using a particular philosophical method. There are (at least) two kinds of inquiry: inquiry into questions, and inquiry into phenomena. Christoph Kelp argues that these two kinds of inquiry have different aims. The aim of inquiry into a question – call this 'Q-inquiry' – is to settle the question Q, where settling Q is "properly closing Q for oneself in the affirmative/negative" (Kelp 2021a: 2). What is it to *properly close* Q for oneself? It is to be in some particular relationship with the true, complete answer to Q.

On the dominant way of thinking about questions, a question is a partition on possibility space, creating jointly exhaustive and mutually incompatible cells which determine what are the possible answers to the question (Groenendijk and Stokhof 1984). For example, the question 'Who is the oldest Marx brother?' partitions possibility space thus:

Chico is the oldest Marx brother.	Harpo is the oldest Marx brother.	Groucho is the oldest Marx brother.	Gummo is the oldest Marx brother.	Zeppo is the oldest Marx brother.

A complete answer to a question fully settles the question, by ruling out all but one cell as that which contains the actual world. A partial answer rules out some cells as those which contain the actual world, but leaves open more than one. The true, complete answer to 'Who is the oldest Marx brother?' is 'Chico is the oldest Marx brother': this answer (correctly) rules out all but one cell as that which contains the actual world. A true, partial answer to this question is 'Groucho is not the oldest Marx brother': this rules out one cell as that which contains the actual world, but doesn't fully settle the question, as there remain multiple cells that could contain the actual world.

Different philosophers have different views about what relationship the inquirer must have to Q's true, complete answer in order to have properly closed Q. Some philosophers hold that inquirers must know the true, complete answer to Q (Millar 2011; Kelp 2014, 2021a); others that they must believe Q's true, complete answer

(Kvanvig 2003; Lynch 2005); yet others that they must justifiably believe Q's true, complete answer (Rorty 1995, 2000; Feldman 2002; Davidson 2005). But the aim of inquiry into a phenomenon – call this 'Ph-inquiry' – is not to settle some question Q. Kelp suggests instead that it is to *understand* the phenomenon of interest (2021a: 2).

Some philosophical inquiries are Q-inquiries: they are inquiries into specific questions. For example, epistemologists inquire into questions like 'What is knowledge?', 'Is justified true belief sufficient for knowledge?'; metaphysicians into questions like 'Is time real?', 'What is causation?'; moral philosophers into questions like 'What is goodness?', 'Are there moral facts?'; and so on. Philosophical methods are methods for these kinds of inquiries. For example, traditional (reductive) conceptual analysis is a method for answering 'What is X?' questions. But other philosophical inquiries are Ph-inquiries. For example, epistemologists inquire into knowledge, metaphysicians into the structure of the world, ethicists into good and evil. And of course, the two kinds of inquiries are not independent of each other. An epistemologist inquires into the question 'What is knowledge?' because she is interested in the phenomenon of knowledge. Indeed, philosophers may inquire into some questions without thinking they could ever properly close them, but in the hope that the process of Q-inquiring will lead to progress in a related Ph-inquiry. For example, an epistemologist might think that she will never settle the question 'What is knowledge?', but that her Q-inquiries into this question, undertaken using certain philosophical methods, will nevertheless shed light on the phenomenon of knowledge.

We can understand a broad Ph-inquiry as consisting of many Q-inquiries: a broad inquiry into the phenomenon of knowledge might consist in a number of Q-inquiries into questions like 'What is knowledge?', 'What does the concept *knowledge* do for us?', 'Is knowledge sensitive to stakes?', 'Is knowledge incompatible with luck?', and so on. This fits well with Kelp's account of the two different kinds of inquiries. For Kelp, the aim of Q-inquiry is knowledge, and that of Ph-inquiry is understanding; but understanding, for Kelp, is reducible to knowledge: to understand some phenomenon is to have "systematic knowledge" about it, to have "various pieces of knowledge [that are] hooked up in the right way" (2021a: 3). Then it is clear how a number of Q-inquiries can constitute a Ph-inquiry: each Q-inquiry, if successful, can generate a piece of knowledge such that, when all these pieces are 'hooked up in the right way', can constitute systematic knowledge – i.e., understanding – of the broad phenomenon.

I said at the beginning of this section that a philosophical project is an inquiry undertaken using a particular philosophical method. We can now say more precisely that a philosophical project is a Q-inquiry undertaken using a particular philosophical method, as philosophical methods are methods for answering Q-inquiries. Though all Q-inquiries have the same aim – namely, settling Q, whether this means the inquirer coming to know Q's true answer, believing it, or justifiably believing it – different philosophical projects have different aims. For example, the aim of inquiry into the question 'What is knowledge?' undertaken using traditional (reductive) conceptual analysis is to generate a set of individually necessary and jointly sufficient conditions for satisfying the concept *knowledge*. This aim is descriptive: the theorist aims to describe what it takes for an object to fall under the concept *knowledge*. In contrast, an inquiry into the question 'Why do we have the concept *knowledge*?' conducted using the method of conceptual reverse-engineering aims to reconstruct practical problems that we face to which the concept *knowledge* offers a salient solution (I will talk more about conceptual reverse-engineering in §2 of the next chapter). This aim is explanatory, rather than descriptive: the theorist aims to explain why it is that creatures like us have a need for this concept.

I propose to distinguish philosophical projects by their aims, and in particular, in terms of four following aims that philosophical projects can have: descriptive, explanatory, evaluative and ameliorative. Call a 'descriptive project' any project with a descriptive aim, an 'explanatory project' any project with an explanatory aim, an 'evaluative project' a project with an evaluative aim, and an 'ameliorative project' a project with an ameliorative aim. The same method can be used for different projects. For example, in the next chapter I argue that Craig makes use of the sub-method of conceptual reverse-engineering, hypothetical genealogy, to inquire into two questions: Why do we have the concept *knowledge*? And what is that concept like – what is its intension and extension? But Miranda Fricker likewise uses hypothetical genealogy to inquire into the question: is the function that *knowledge* serves legitimate? Fricker reconstructs the interactions of our concept *knowledge* with other features of our cultural and political context to draw out what we might call 'secondary functions' that knowledge will serve in our socio-political context if its primary function is, as Craig hypothesises, to flag good informants. One secondary function of *knowledge* that Fricker identifies is that of assigning credibility to speakers, which, given our socio-political context, leads to testimonial injustice, whereby speakers who are in some way

socially marginalised are "wrongly denied credibility" due to the hearer's prejudices about some social identity of the speaker (1998: 170; see also Fricker 2007). Craig and Fricker both use the same method, hypothetical genealogy (a sub-method of conceptual reverse-engineering), but their projects have different aims: Craig undertakes two projects, one with a descriptive aim and one with an explanatory aim, while Fricker's project has an evaluative aim. As such, they are engaged in different kinds of projects: Craig a descriptive project and an explanatory project and Fricker an evaluative project.

A philosopher can have multiple goals in a broad Ph-inquiry: she might want to accurately describe the content of her concept of interest *and* explain why its content is as it is *and* evaluate whether the concept serves some function well *and* ameliorate the concept to better serve this function. But a single project cannot have multiple aims. Rather, such a philosopher should be understood as being engaged in multiple projects at the same time, as using philosophical methods to undertake a number of different Q-inquiries, which are all part of the same broad Ph-inquiry. These projects might be undertaken using the same method, as in the case of Craig, who uses hypothetical conceptual genealogy as the method both for his Q-inquiry into 'What is *knowledge*?' and his Q-inquiry into 'Why does *knowledge* have the intension and extension it has?' In that case, the projects would be distinguished from each other by the relevant Q-inquiry.

Haslanger too writes of 'descriptive' and 'ameliorative projects' in philosophy, but she means something different by these terms than I do. For Haslanger, a question of the form 'What is X?', where X is a concept, can be answered in three different ways, and these ways of answering the question she calls *conceptual*, *descriptive* and *ameliorative projects*. A conceptual project answers the question 'What is *knowledge*?', for example, by appealing to *a priori* methods, such as introspection, to articulate the intension of "*our* concept of knowledge" (2006: 96): the concept of knowledge that is actually used by our epistemic community in various areas of life. A descriptive project, in contrast, is "concerned with what objective types (if any) our … vocabulary tracks", and will start by "identifying paradigm cases" in which the concept applies, before "draw[ing] on empirical (or quasi-empirical) research to explicate the relevant kind of type to which the paradigms belong" (95). An ameliorative project, on Haslanger's understanding, is one that asks questions like "What is the point of having the concept in question; for example, why do we have a concept of knowledge or a concept of

belief? What concept (if any) would do the work best?" Such a project might involve introducing a new concept to serve a particular functional role for the epistemic community, and stipulating its content in terms of its playing this role (95-6).

I do not mean by 'descriptive project' or 'ameliorative project' what Haslanger means by these terms. Haslanger's categories of conceptual projects, descriptive projects and ameliorative projects are more akin to my categories of intension-first, extension-first and function-first approaches (§2.2), respectively; though they do not align perfectly with these categories, as my three approaches are broad methodologies, while Haslanger's three projects seem more like individual methods. Haslanger's conceptual and descriptive projects would both count as descriptive projects on my categorisation, as they are projects with descriptive aims: roughly, conceptual projects aim to describe the intuitive intensions of our concepts, while descriptive projects aim to describe the extensions of our concepts iff those concepts track 'objective types', for example if they are natural kind concepts. Evaluative and ameliorative projects, on my categorisation, would be grouped together as ameliorative projects under Haslanger's categorisation. Haslanger's categories do not have an equivalent for my explanatory projects.

3. Objections and replies

In this section, I raise and respond to some objections to my account of the function-first approach. The first batch of objections (§3.1) concern whether my account of the function-first approach retains the benefits of Hannon's. I argue that it does. The second batch (§3.2) concern whether concepts can have the kinds of functions that the function-first approach assumes they do. I argue that they can, and in particular that they can have proper functions. I then raise a problem for this picture of concepts' functions – how are we to identify concepts' proper functions? – which I defer answering until the next chapter.

3.1. Does this methodology retain the benefits of Hannon's method?

Objection. My characterisation of the function-first approach as a methodology makes it less precisely demarcated than Hannon's account of the function-first approach as a method. Hannon characterises the function-first approach as a three-step method, and describes each step in some detail; while I say that the function-first approach is a

methodology, and includes any method for theorising about concepts that a function-first philosopher might use. Thus Hannon's characterisation of the function-first approach is preferable to mine, because it is more precise.

Response. The content of this objection is correct: my characterisation of the function-first approach does make it less precisely defined than does Hannon's. But I see this as a feature, not a bug, of my account. It is an advantage of my account that it both captures the central insight that motivates Hannon, namely the function-first insight, while being broad enough to unite other philosophers who are similarly motivated under the function-first banner. My characterisation of the function-first approach is sufficiently precise to distinguish it from intension- and extension-first approaches, without being so narrow as to exclude the work of Craig, Haslanger and, as I will argue in the next chapter, many other philosophers who are guided by the function-first insight from counting as function-first philosophers. For this reason, my account of the function-first approach is better situated than Hannon's in relevant literatures, making room for philosophers who have influenced Hannon, as well as his contemporaries who are motivated by similar concerns, who ask similar questions, and so on, in their theorising about concepts.

Objection. A crucial motivation for Hannon and Craig's methods is that they are, compared to conceptual analysis, relatively invulnerable to counterexample. If the function-first methodology is as broad as I say it is, do we have reason to think it will retain this advantage? If not, why should anyone take a function-first approach to a concept?

Response. Two methods that are particularly useful for function-first philosophy are conceptual reverse-engineering and conceptual engineering, and both of these have the advantage over conceptual analysis of being relatively invulnerable to counterexample. These two methods are the topic of the next chapter, but I will briefly outline the methods here to show why this is so.

Conceptual reverse-engineering is a method for finding the function of a concept. In conceptual reverse-engineering, a theorist looks to how a concept is used in a case that is typical of its use: to what needs the concept serves in this case, and what problems are solved for the relevant epistemic community by their use of the concept. The function of the concept in all cases is identified with its function in this case. (In the next chapter, I will explain how the conceptual reverse-engineer identifies or constructs this kind of typical case.) The conceptual reverse-engineer can then

develop an account of the concept's intension and extension that is based around its being able to serve this function in typical cases. An account of the concept's intension and extension that is derived from this method is relatively immune to counterexample, compared to an account derived from conceptual analysis. The kinds of counterexamples that threaten an account of a concept derived from conceptual reverse-engineering are examples in which a concept demarcated by the proposed intension and extension cannot serve the concept's function in a typical case. In cases that are sufficiently atypical, a concept demarcated by the proposed intension and extension being unable to serve the concept's function will not undermine that intension and extension's claim to being those of the concept of interest. These are not the kinds of cases that explain why the epistemic community has the concept in the first place, so the concept's non-functionality in these cases does not show that this function-first account of the concept is wrong: the concept may well have the intension and extension proposed, as this is what enables the concept to serve its function in typical cases.

Conceptual engineering is a method for improving our concepts. One might engineer a concept to better serve a function, or one might engineer it for other reasons, for example to eliminate some conceptual defect such as vagueness. In conceptual engineering, a theorist revises or replaces a concept with a view to improving the concept. Conceptual engineering is also relatively invulnerable to counterexample, compared to conceptual analysis. For if the new, or newly revised, concept excludes from its extension something which intuitively ought to be in it, or includes in its extension something that intuitively ought not to be in it, this will be acceptable to the conceptual engineer if the case is not one that is relevant to the reason for which the engineer revised or replaced the concept.

For example, Haslanger replaces the ordinary concept *woman* on which all and only objects that fall under this concept have a particular cluster of bodily characteristics, viz. vagina, uterus, ovaries, and so on, with a concept on which all and only objects that fall under the concept *woman* are subjugated in particular ways based on their being imagined or perceived to have this particular cluster of bodily characteristics. Haslanger's motivation for this replacement is that the new concept can better serve as an "effective [tool] in the fight against injustice"; more specifically, that it will allow us to more effectively:

1. Identify and explain persistent inequalities between males and females;

2. Be sensitive to similarities and differences between males and females, including identifying interlocking sources of oppression, for example the intersectionality of race, class and gender;
3. Track how gender is implicated in a broad range of social phenomena, for example explore whether art, science or law are 'gendered';
4. Allow us to develop an understanding of women's agency that will aid feminist efforts to empower social agents (2000: 36).

On Haslanger's new concept *woman*, some people who we ordinarily think of as women will not count as such. For example, the Queen of the United Kingdom might not count as a woman, on Haslanger's concept *woman*, because she is not subjugated in the relevant ways: she is protected from this treatment by her immensely privileged, and incredibly unusual, social situation. But this does not trouble Haslanger, as her aim in engineering the concept *woman* "is not to capture what we do mean" in our use of the word 'woman', but rather to produce a concept that can better serve these political purposes (34). For the purposes of Haslanger's "critical feminist inquiry, oppression is a significant fact around which we should organize our theoretical categories", with the consequence that non-oppressed females won't count as women; but this does not prevent Haslanger's concept *woman* from serving its intended purposes, as "relative to the feminist … values guiding our project – they are not the ones who matter" (46). What is important is not whether Haslanger's concept *woman* can capture in its extension all those people who we would pre-theoretically think of as women. What matters is, if one such person is excluded from the extension of *woman* (or alternatively, someone we wouldn't pre-theoretically think of as a woman is included in the extension of Haslanger's concept *woman*), whether their being excluded (or included) "is in conflict with the feminist values that motivate the inquiry" (46). Insofar as there is no conflict, the counterexample does not undermine Haslanger's project.[2]

I will discuss both conceptual reverse-engineering and conceptual engineering in much more detail in the next chapter. But for now, I wish to make the point that two methods that we will see are very useful for function-first philosophy are methods that

[2] As we will see in §3.4 of the next chapter, not all counterexamples to Haslanger's project are like this. In particular, she takes very seriously Jenkins's (2016) worry that her concept *woman* excludes some trans women from its extension.

share the central advantage of Hannon's method over conceptual analysis, namely, relative invulnerability to counterexample.

3.2. Can concepts have functions?

Objection. The function-first approach assumes that concepts can have functions. Further, it assumes that concepts can have functions beyond that of denoting, and therefore allowing us to think and talk about, their objects. The kinds of functions that concepts are assumed to have are supposed to be able to explain why the concept exists or is in use in a particular community. For example, Craig and Hannon argue that the reason we have the concept *knowledge* is because this concept functions to meet our need to flag good informants. On this picture, concepts must be able to have something like the functions that Larry Wright describes, where an item X has a function Z iff (a) X is there because it does Z, and (b) Z is a consequence or a result of X's being there (1973: 161). 'Is there' here can mean either 'exists', or 'is in use in this context'. Following Peter Graham (2014), we should understand functions as belonging to types of items, rather than tokens, so that a token X that is sufficiently malformed to be able to do Z can still have the function of doing Z; a function it is unable to achieve.[3] Herman Cappelen calls these kinds of functions "central functions" (2018: 181), and argues that it is not plausible that concepts have central functions beyond denoting their objects. If this is true, then the function-first approach is a non-starter: we cannot better understand, evaluate, and so on, our concepts with an eye to their function if they cannot have the relevant kinds of functions.

Here is Cappelen's argument. If concepts have central functions, these functions are either relative to context or they are not. We cannot make sense of the notion of context-relative central functions. Consider an example: what is the central function of the concept *woman* in each of these uses?

1. There were three women on the flight.
2. There are more women than men in my class.

[3] Graham's example: a heavily malformed heart cannot pump blood, and so does not exist because *it* can pump blood, thereby failing both of Wright's conditions. Yet the type heart of which this heart is a member has the function of pumping blood because there is a feedback mechanism which takes past token hearts as inputs and produces or maintains hearts because those past token hearts pumped blood. So the type heart exists because it (tokens of the type) pumps blood (2014: 19).

3. Women's shoes are more expensive than men's.
4. Women tend to do more housework than men.
5. Women get breast cancer more often than men.
6. Women tend to smoke less than men.

In each of these sentences, the concept *woman* has a "stable function": it is used to denote women (Cappelen 2018: 182). But beyond denoting women, the concept *woman* is not doing anything that explains why it is there in each use. So Cappelen concludes that if concepts have central functions, they cannot be context-relative; rather, those who hold that concepts have central functions should be invariantist about those functions. However, the only true version of invariantism about concepts' central functions is one on which the central function of a concept is simply the "trivial" function of denoting its object, as "the only *universal*, i.e., stable, function of a concept 'C' is to denote Cs" (183). The function of the concept *woman* is to denote women, the function of *knowledge* is to denote knowledge, the function of *fish* is to denote fish, and so on. Thus, Cappelen concludes that the only central functions concepts can have are denoting functions.

Response. Cappelen's argument is unpersuasive, for at least two reasons. First, it is not at all clear that all concepts have the denoting functions that Cappelen claims it is "non-controversial" to assign to concepts (183); such concepts can nevertheless have a function that explains why they are used by an epistemic community. For example, Richard Joyce discusses the traditional Polynesian concept *tapu*, where an object that falls under this concept is such that agents are forbidden to touch it because the object has "a kind of uncleanliness or pollution" which "may pass to humans through contact, may then be transmitted to others like a contagion, and which may be canceled through certain ritual activities, usually involving washing" (2001: 1). The concept *tapu* fails to denote anything, because there is nothing in the world that is tapu: something could be tapu only if imbued with certain "supernatural and magical forces", which simply do not exist (4). But noting that the concept is defective in this way – it is supposed to, but fails to, denote – is perfectly consistent with recognizing that it serves an important function, or indeed many functions, for those who use it: it guides the concept-users' behaviour in various ways which are beneficial for them; for example, it places restraints on their touching dead bodies without extensive washing afterwards (see Buck 1910), which helps to prevent the spread of disease.

It is an open question whether some of our most philosophically interesting concepts fail to denote in the same way as *tapu*, yet nevertheless serve a function for an epistemic community. Joyce argues that moral concepts are like this. The intension of the moral concept *right*, for example, is as the moral realist takes it to be, according to Joyce: an action φ's being (morally) right means that an agent has a reason to φ regardless of whether φ-ing would further her ends, where furthering her ends includes satisfying any desires and interests she has; that is, if it is morally right to φ, then one ought to φ, where this is a categorical, not a hypothetical, imperative (2001: 56). However, Joyce argues, there are no actions that fall under the extension of this concept. So *right* fails to denote. Nevertheless, moral concepts like *right* serve an important function: they motivate helping behaviours, which contribute to a person's reproductive fitness, and do this better than concepts on which the only normativity right action has is to do with desire satisfaction (135-40). Thus Cappelen's claim that "the only universal, i.e., stable function of a concept 'C' is to denote Cs" (2018: 183) is not obviously true for some concepts, including some philosophically interesting concepts such as *right*.

Second, Cappelen gives too short shrift to contextualism about central functions: the view that concepts could have different central functions in different contexts. He considers only an extreme contextualist position, whereby we expect a concept to have a different central function in each instance of its application. But one could endorse a more stable contextualism about central functions, according to which the same concept has different central functions in different domains of use (compare the versions of epistemic contextualism developed by Greco (2008) and Hannon (2015)). Consider again Cappelen's example of the concept *woman*. One could be a contextualist about the central function of *woman* because one thinks that this concept has a particular central function in the medical domain, where it is used (let's suppose) to denote adult humans with some or all of a particular cluster of bodily features, such as a vagina, uterus, ovaries and so on; another central function in the civil domain (for example, when used to design censuses, or plan the development of public spaces) where it's used (again, let's suppose) to denote any human who identifies as a woman. (Such a contextualism about the function of *woman* would sit well with Jennifer Saul's contextualism about 'women' and its cognates, according to which 'X is a woman' is true in a context C iff X is a human and relevantly similar (according to the standards

in C) to most of those possessing all the biological markers of the female sex (2012: 201). In some contexts, such as medical contexts, this might require having some, most or all of the biological markers of the female sex; in other contexts, "sincere self-identification as a woman will be sufficient" (203).) This view is not invariantism about central functions, but it nevertheless allows that the central function(s) of a concept are stable: within a domain of use, a concept's central function will be set, so the concept won't have a different central function on each instance of its application.

Still, the function-first philosopher ought to be able to identify these context-relative but stable functions in a convincing way. One way to do this is to appeal to Ruth Millikan's notion of proper function. On Millikan's view, an item has a proper function if it is a member of a "reproductively established family", and its proper function is whatever its ancestors did that contributed to the reproductive success of the family, which in turn explains the existence of this member of the family (1984: 28). Biological entities, like hearts, have proper functions: humans have hearts because hearts pumped blood in our ancestors, thereby enabling their survival; hearts that successfully pumped blood were passed down through generations via natural selection, as humans without hearts that successfully pump blood did not live long enough to pass on their genes; so our having hearts now is explained by hearts being what pumped blood in our ancestors; thus the proper function of the human heart is to pump blood. Millikan argues that concepts can have proper functions too, and concepts that have survived for generation after generation have done so "because they correlate with functions" (31).

Amie Thomasson argues that we can identify the proper functions of our concepts by asking what we are able to do with those concepts "that we couldn't do, or couldn't do as effectively, without them" (2020: 445). We can ask this in a domain-relative way. For example, we can ask what the concept *woman* in the medical domain allows us to do that we couldn't do, or couldn't do as easily, if we lacked it. Supposing that the concept *woman* in the medical domain is used to denote adult humans with a particular configuration of bodily features, it is plausible that this concept has as its proper function to group together some class of humans who are more likely than any other group to experience certain kinds of health problems, which in turn makes it easier and more efficient to study, diagnose and treat these problems. That the concept *woman* in the medical domain has this proper function explains its continued use in medical contexts, for example to name specific departments in hospitals, or to

disseminate medical information to the public (see for example National Health Service 2019).

Cappelen objects to Thomasson that the only kinds of functions that could be concepts' proper functions are the "trivial" denoting functions he has already countenanced:

> The reason 'salmon' is useful for us is that it can be used to talk about salmons (or denote salmons). The reason 'freedom' is useful is that it can be used to talk about freedom. We care about salmons and freedom and so we have words that enable us to talk about them. (2018: 187)

(Note that Cappelen here talks in terms of words, not concepts, but the same point can be made *mutatis mutandis* about concepts: the concept *salmon* is useful because it enables us to talk (and think) about salmon; similarly for *freedom*.) He argues there is too much variability in our use of concepts to make plausible that they have a proper function beyond denoting. We can use *freedom*, for example, "to undermine freedom or to promote it or to discuss it or to make fun of it" (187). There is "no limit" to what we can do with this concept, and further, what we do with it "will vary between contexts and over time" (187). Then we cannot point to any particular use of a concept to find its proper function, the function that contributed to its survival across generations and which explains its use today, beyond denoting. Therefore, proper functions are not the kind of functions that the function-firster needs concepts to have.

Thomasson responds to Cappelen that the kind of variability in our use of concepts that he cites does not undermine their claim to having proper functions (beyond denoting), for the analogous objection would never be accepted for biological entities like hearts. For example, that people all over the world use cow hearts as a foodstuff does not undermine that the proper function of a cow's heart is to pump blood in a cow's body. Cappelen's objection assumes, incorrectly, that we can identify the proper function of an item "just by looking to anything that can be or has been done with the item in question" (Thomasson 2020: 446). But this is not the case: to identify an item's proper function, we must look to what is done by the item in certain privileged cases. However, we now face the problem of specifying what these privileged cases are. I will put off answering this question until the next chapter, where I will argue that the same way of identifying or constructing typical cases in conceptual reverse-

engineering can be used to identify the privileged cases from which we can glean an item's proper function.

3.3. Spurious proper functions

In understanding concepts as having proper functions, I will run up against another problem, one that arises for etiological accounts of function in general. Mark Bedau (1991) noticed that Wright's account of functions allows that non-living, inorganic materials can have functions, when intuitively they cannot. Recall that, for Wright, an item X has a function Z iff:

(a) X is there because it does Z, and

(b) Z is a consequence or a result of X's being there.

Bedau discusses a case described by Richard Dawkins of clay crystals that build dams in streams. These clay crystals lay down patterns in streams, and layers of sediment stack up on top of each other according to those patterns. Consequently, the crystals replicate themselves, eventually building a dam in the stream. Once the dam is built, it withstands the flow of the stream, and cannot be washed away. So the clay crystals (X) build dams (Z), they are there because they build dams, and the dams' being built is a consequence of the crystals being there. As such, the clay crystals meet both of Wright's conditions for having a function. But it is not plausible that these crystals have the function of building dams. As Graham puts it, "building dams is just something these crystals *do*, not something they are *supposed* to do" (2014: 19). There's nothing the crystals are supposed to do. They have no function.

Millikan's conditions for an item's having a proper function similarly generate the result that the clay crystals have the function of building dams. Recall that, for Millikan, an item X has a proper function Z iff:

(a) X is a member of a reproductively established family, and

(b) Doing Z contributed to the reproductive success of the family.

The clay crystals replicate themselves, and replication is a form of reproduction; thus the clay crystals constitute a reproductively established family, meeting condition (a). The clay crystals that make up the dam, at least in its later stages, exist because the pattern laid down by the earlier clay crystals caused layers of sediment to stack on top of each other, thereby forming a dam; thus building a dam is what the ancestors of the later clay crystals did which contributed to the reproductive success of the family, in

turn explaining the existence of the later members. So the problem of the clay crystals is a problem for Millikan's, as well as Wright's, account of function.

Graham's solution is to follow Peter McLaughlin (2001) in modifying Wright's account of function to include a benefit or welfare condition. In order for Z to be a function of X, Z must be a means to some good or benefit for the system containing X, and this benefit must be relevant to the feedback mechanism that explains why X exists in the system. Functions aren't just explanatory effects, rather, they are "explanatorily *beneficial* effects" (Graham 2014: 20). The heart's pumping blood help's the heart's system, the human body, to survive, thereby doing that system a good. So pumping blood is the function of the heart. But the clay dam "doesn't have a good" because it has "no ends", and nothing can be good for a system with no ends (20). So building dams is not good for the system containing the crystals – the dam – and as such, building dams is not the function of the crystals. This leads Graham to the following account of functions. A function X in a system S is X iff:

(a) X does Z in S.

(b) X benefits S.

(c) X exists in S because X benefits S (X is the product of a feedback mechanism involving the beneficial character of Z to S). (2014: 20).

Adding a benefit condition to a Millikanian account of proper function would yield the following. A proper function of X in S is Z iff:

(a) X is a member of a reproductively established family.

(b) Doing Z in S contributed to the reproductive success of the family.

(c) Z benefits S.

However both of these accounts face a problem, in that they rule out the possibility of nefariously functional items: items which have a function, but a function that negatively impacts the system of which that item is part. In the next chapter, we will look at a philosophical method for uncovering the functions of concepts: conceptual reverse-engineering. It is an assumption of conceptual reverse-engineering that a concept may be revealed to have a function that we do not reflectively endorse, and indeed which we reject as bad, just as well as a concept may have a function that we do endorse. To use Matthieu Queloz's (2021) terminology, a conceptual reverse-engineering project can *debunk* as well as *vindicate* our practices with some concept. For example, our concept *virgin* might serve the function of sexually subjugating women; under patriarchy, an item's serving this function might well contribute to its

continued use – its reproductive success. But sexually subjugating women is not beneficial for humans. It doesn't do us any good. So this function doesn't benefit the system in which the concept *virgin* is part: the human communities which use this concept. One might argue that it does the system of patriarchy good. But just as clay dams don't have ends and so cannot be benefited, neither does patriarchy. Patriarchy is a social structure under which people live, but it is not itself constituted by people, nor by groups of people.

Here is another solution to the clay crystals problem, one which doesn't rule out the possibility of nefariously functional items. Instead of adding a benefit condition, we can add a condition that the system S of which the item X is part is something that can be benefited or harmed. This would yield the following account of proper function. A proper function of X in S is Z iff:

(a) X is a member of a reproductively established family,

(b) Doing Z in S contributed to the reproductive success of the family, and

(c) S can be benefited or harmed.

Note that my condition (c) is weaker than a benefit condition: in order for Z to benefit S, it must be the case that S can be benefited or harmed; but it is not the case that, in order for S to be capable of being benefited or harmed, Z benefits S. In this way, this account demands less of an item for it to be functional. But it still solves the clay crystals problem: although the clay crystals (X) are a member of a reproductively established family (thus meeting condition (a)), and although building dams (Z, and following Graham, we can understand the dams as the systems S in which this function is manifested) contributed to the reproductive success of the family (thus meeting condition (b)), dams are not capable of being benefited or harmed, so the clay crystals do not meet condition (c) for having a proper function.

4. Conclusion

In this chapter, I offered an account of the function-first approach to concepts as a methodology for conceptual philosophy: a collection of methods that a philosopher can make use of in theorising about concepts. I argued that, for a method to fall under the function-first banner, it must in some way give priority to a concept's function over its intension and extension. What this looks like will depend on the aim that the method is serving. Where the method serves a descriptive aim, the relevant kind of priority is observational priority; where it serves an explanatory aim, it is explanatory priority;

where it serves an evaluative aim, it is normative priority; and where it serves an ameliorative aim, it is teleological priority. I illuminated the methodology by contrasting it with two alternative methodologies, intension-first and extension-first approaches to concepts, which give priority (in the relevant way) to intension over extension and function, and extension over intension and function, respectively. On my picture, the function-first approach is a broad methodology, encompassing many methods, without any deep similarity between all of them.

I then responded to a number of potential objections to my account of the function-first approach to concepts. The first objection said that Hannon's characterisation of the function-first approach is preferable to mine, first because it is more precise, characterising the function-first approach as a three-step method; and second because it gives philosophers a reason to engage in the function-first approach: his method is relatively invulnerable to counterexample, compared to traditional conceptual analysis; but on my understanding, the function-first approach is a broad methodology, which might not share this advantage. I responded first that my characterisation of the function-first approach is exactly as precise as I want it to be: it is precise enough to distinguish the function-first approach from intension- and extension-first approaches, but broad enough to encompass a number of methods that are relevantly similar to Hannon's under the function-first banner. Second, I argued that two methods that are particularly useful for a function-first philosopher are conceptual engineering and conceptual reverse-engineering, and both of these methods have the advantage over conceptual analysis of relative invulnerability to counterexample. The second objection questions whether concepts can have functions beyond the trivial function of denoting their objects. I argued that, first, not all concepts have denoting functions; and second, that some concepts can have proper functions beyond denoting, but we won't be able to identify those functions just by looking to any use of the concept. Rather, we must look to a concept's use in certain privileged cases, to be explored in the next chapter.

Chapter 2. Conceptual reverse-engineering and conceptual engineering

1. Introduction

In this chapter, I describe two methods that are extremely useful for function-first philosophers, and which will underpin my projects in this thesis: conceptual reverse-engineering and conceptual engineering. I argue that conceptual reverse-engineering is an empirical method for confirming or disconfirming hypotheses about the functions of concepts; and that conceptual engineering is a method for improving our conceptual world by revising, replacing or abandoning defective concepts, improving non-defective concepts, or creating new concepts to serve functions currently going unmet by existing concepts. I offer taxonomies of both methods: various ways of partaking in these methods. I introduce the 'Strawsonian challenge' to conceptual engineering, and distinguish two versions of the challenge. On the first, the conceptual engineer must show how, in revising or replacing a concept, she has not changed the subject in such a way that she is no longer engaged in the same philosophical project. On the second, the conceptual engineer must show that it is possible to change a concept's intension or extension and still have the same concept with which one began. I consider several responses to both versions of the challenge, and argue that, for my purposes, the second challenge needn't be solved, and the first challenge can be successfully addressed using different arguments in different cases. Discussion of problems facing conceptual reverse-engineering are left until the next chapter, where I consider four conceptual reverse-engineering projects on *knowledge* in depth.

2. Conceptual reverse-engineering

In this section, I outline the method of conceptual reverse-engineering, whereby a theorist looks to a case in which a concept is used to glean the function of the concept. In §2.1, I argue that the case to which a conceptual reverse-engineer looks must be a *typical* case: a case representative of cases in which a given community uses a concept. In §2.2, I offer a taxonomy of sub-methods of conceptual reverse-engineering. In §2.3, I explain the value of conceptual reverse-engineering for function-first philosophers.

2.1. What is conceptual reverse-engineering?

Conceptual reverse-engineering is a method for finding the function of a concept. In conceptual reverse-engineering, a theorist looks to what a concept does for some group of agents in a particular case: what needs it serves, what problems are solved for the group by their use of the concept. The function of the concept for the community that uses it is then identified with what it does for this group of agents in this case. That the concept has this function can then shed light on the 'shape' of the concept – its intension and extension – because the concept must have an intension and extension that enables it to serve its function.

To understand conceptual reverse-engineering, it is helpful to consider reverse-engineering of a more familiar kind: that of artefacts. Suppose an alien encounters a hammer for the first time, and has no idea what it is for: what purpose(s) hammers serve for humans that explains why humans invented them. To learn about hammers and their purpose(s), the alien looks to a situation in which a hammer is being used; say, a situation in which a human is using a hammer to drive nails into a piece of wood. The alien generalises from this case that the function of hammers is to drive nails into things: this is what hammers do for humans that explain why they were invented, thus why they exist (why hammers 'are there'; cf. §3.2 of the last chapter). Having identified the function of a hammer, the alien can make sense of why hammers have the shape they do: why they have heavy heads (so they can hit nails with enough force to drive them into other objects); flat faces (a larger surface area means a better chance of hitting the nail); and ergonomic handles (to make them easier to hold). Thus the alien learns about the hammer itself by discovering its function, which she did by looking to how the hammer was used in a particular case.

But here is a problem. Now that humans have hammers, we can use them for all kinds of purposes. For example, a group of people might use a hammer as a bat in a game of rounders. If the alien had come across this situation, she might well have inferred that being used to hit balls in games is *the* function of hammers: the purpose hammers serve for humans that explains why we invented them, thus why they exist. This generalisation could also explain why hammers have the shape they do, though not as well as its actual function: the handle must be ergonomic so that the player can hold it easily; the head must be heavy so it can repel the ball; and so on. But this isn't the function of hammers. By looking to how a hammer is used in a particular case and making a generalisation from that case, the alien fails to discover the function of

hammers. Rather, she has seen a hammer being used in an atypical way, and made a mistaken inference that being used in this way is the function of hammers. This would, in turn, lead her to have strange beliefs about hammers themselves. For example, how would she understand the hammer's claw?

This suggests a general problem for conceptual reverse-engineering. If we look to some situation in which a concept is used, it may be being used in an atypical way, just as the hammer in the game of rounders is being used in an atypical way. Then we can't reverse-engineer the function of the concept from that situation of its being used. But if we don't know the function of the concept (which we don't: this is precisely why we want to reverse-engineer it), then how can we tell whether the way the concept is used in a given situation is typical, and thus whether we can reverse-engineer the function of the concept from this instance of its use?

Hannon makes this objection to conceptual reverse-engineering. He argues that it is "doubtful [that] we can glean the point of our epistemic concepts from our linguistic practices featuring the words that express those concepts" (2019: 23). For just as we can use hammers, once we have them, for various purposes other than the one for which they were designed, once we have a concept, we can use it for various purposes beyond the one that explains why the concept is there. For example, Hannon holds that the function of the concept *knowledge* is to flag good informants; we have the concept *knowledge* because it serves this purpose for us. But now that we have the concept, we can use it for other purposes: "I might try to comfort a friend who is experiencing hardship by saying 'I know that you'll get through this.' … My use of 'knows' in this case is intended to provide reassurance, not to identify a reliable informant" (23). If our alien tried to glean the function of *knowledge* from this interaction, she would end up with a false belief about its function. This, in turn, would lead her to have false beliefs about *knowledge* itself: for example, that one can have knowledge without good evidence, just because one has a sincerely held belief.

But this is not a knock-down argument against conceptual reverse-engineering. Rather, it is an argument for being discerning about the cases to which we look when trying to reverse-engineer a concept's function. Instead of looking to just any case in which a concept is used, the theorist needs a *typical* case. Typicality here is not a matter of frequency: we might frequently use hammers to break ice, say; we might frequently use screwdrivers to open paint cans, stethoscopes as reflex hammers, pipe cleaners for arts and crafts; and we might frequently use the concept *knowledge* to

comfort friends, but this doesn't make any of these uses typical in the relevant sense. Rather, the relevant sense of 'typical' is that of meaning 'representative of the relevant type'. A case in which a hammer is used is typical if it is representative of cases in which hammers are used. That is, it is typical if it is the kind of case that explains why we have hammers. Similarly, a case in which a concept is used is typical if it is representative of cases in which the concept is used. That is, it is typical if the concept is being used in a way that explains why it is there.

But wait. Isn't this just to say that the kinds of cases from which we can reverse-engineer a concept (or any other item) are just those cases in which the concept (artefact, etc.) is serving its function? Conceptual reverse-engineering was supposed to be a method for *finding* the function of a concept. Yet it seems that conceptual reverse-engineering is not possible unless we already know the concept's function, as otherwise we cannot know whether the instance of use from which we reverse-engineer the concept's function is typical or not. The method looks hopelessly circular.

Here is a way of understanding conceptual reverse-engineering that avoids vicious circularity. We can think of conceptual reverse-engineering as a method for confirming or disconfirming hypotheses about the function of a concept. The first step of the method will be to offer a plausible hypothesis about the function of the concept (cf. Craig 1990: 2; Hannon 2019: 24). One can then identify or construct a case in which the concept is used to serve this function. The next step of the method will be to theorise about what the concept must be like in order to serve this function in this case: what intension and extension it must have. If the concept demarcated by this intension and extension is relevantly similar to the intuitive concept of interest – the concept with the intension and extension that, pre-theoretically, we take it to have – then one has a good claim to having identified the function of the concept. But if the concept one ends up with is too dissimilar to the intuitive concept, this is reason to think one's original hypothesis was false. In this way, conceptual reverse-engineering either confirms or disconfirms one's original hypothesis about the function of the concept. There is no vicious circularity here: rather than beginning by assuming that the concept has a particular function, one simply tries out the hypothesis on which it has that function, and sees what follows.

In order to formulate a plausible hypothesis about the function of a concept, the theorist must already have a grasp on the various practices that the relevant epistemic community has with the concept: the various ways that the community uses the

concept to structure their thought, talk and action. Queloz calls this a "conceptual practice" (2021: 3). In particular, she must be able to make evaluative judgements about these conceptual practices: to be able to judge roughly which of these are typical and which are atypical, though she needn't think that these judgements won't later be overturned. One thing she will appeal to in making these judgements will be frequency, although, as noted, frequency isn't the same thing as typicality, in the relevant sense. If the theorist very frequently comes across people using a concept in a particular way, this gives her reason to try out the hypothesis that the concept functions to serve whatever need, or resolve whatever problem, it does in that conceptual practice. Compare: if nine times out of ten, when you come across someone with a hammer, they are using it to drive nails into things, this gives you reason to try out the hypothesis that the function of hammers is to drive nails into things. As J. L. Austin tells us, although "ordinary language is not the last word … remember, it is the *first*" (1956: 11).

Another resource to which the theorist can appeal in making her judgement about (a)typicality is the following counterfactual conditional: if the community lacked this concept, would they still be able to meet the need/solve the problem that the concept is meeting/solving in this case, and do so as easily and as efficiently? (Compare Cappelen (2018: 187) and Thomasson (2020: 448), who both suggest that we can, at least sometimes, identify the function of an item by asking what that item "enables us to do that we couldn't do as effectively or as efficiently without it".) If the answer is "Yes", this is good reason to think that this is not a typical case: the concept is not being used in a way that explains why it is there. Compare: the group of rounders-players who are using the hammer would still be able to meet their need to hit the ball if they lacked the hammer; but it would be much harder to drive a nail into an object with any other item in a toolbox.

One final test for typicality appeals to explanatory priority. An item's having a particular primary function should explain how it can be used to serve other purposes. A typical use of a screwdriver is to screw in a screw; this explains why it has the shape that it has: in particular, the flat edge that slots into the screw; and its having this shape explains why it is used to open paint cans. A typical use of a hammer is to drive nails into things; this explains why it has the shape that it has: in particular, why it has a heavy head and an ergonomic handle; and its having this shape explains why it is used to break ice. Similarly, a test for whether a use of a concept is typical is whether the concept's being used in this way can explain other ways that it is used. (Compare

Fricker's test for a 'paradigm case' of a conceptual practice that it can explain "the nature of the practice in all its diversity" (2016: 166): that other forms of the conceptual practice can be explained in terms of this "explanatorily basic" form (165).) For example, the use of the concept *knowledge* to comfort friends cannot explain why it is also used to flag good informants: people who will tell you that P because it will comfort you are not good informants regarding P, in the sense that matters to Craig and Hannon, and which I discuss in the next chapter. But the use of *knowledge* to flag good informants can explain why we can also use it to comfort friends: if *knowledge* functions to flag good informants, then it must have a particular shape, for example, someone who falls under the concept *knower* must have good evidence for what they believe; if I have good evidence that you will get through this, you should feel comforted that you will get through this.

We are now in a position to resolve the problem considered at the end of the last chapter: that of identifying the 'privileged' cases from which an item's proper function can be gleaned. Privileged cases are just typical cases. A case in which a token cow's heart is demonstrating the proper function is a typical case: a case that is representative of the type of case that explains why this token heart is there. The case of a token cow's heart pumping blood through a cow is typical: if this cow's heart were not there, the need to pump blood in a cow would not here be met. The case in which a token cow's heart is being used as a foodstuff is not typical: if people did not have cow's hearts to eat, they would eat something else. Further, that the type cow's heart functions to pump blood explains why the type is as it is: for example, why it is a muscle; its being a muscle explains why it is also used as a foodstuff: muscles are tasty.

Craig's approach to *knowledge* can be understood as conceptual reverse-engineering, as I have characterised the method. Craig begins by offering a "prima facie plausible hypothesis" about the function of *knowledge*: "what the concept of knowledge does for us, what its role in our life might be"; then develops a picture of what the concept must be like if it is to serve this function: "what a concept having that role must be like, what conditions would govern its application" (1990: 2). He then compares the concept that his method generates to the "intuitive" concept *knowledge*: the intuitive intension and extension of the concept. If there is too significant a mismatch between his concept and the intuitive concept, he takes this to falsify his original hypothesis: "should this method reach a result quite different from the intuitive intension, or one that yielded an extension quite different from the intuitive extension,

then, barring some special and especially plausible explanation of the mismatch, the original hypothesis about the role that the concept plays in our life would of course be the first casualty" (2). Otherwise, he takes the original hypothesis about the function of *knowledge* to be confirmed. Craig's hypothesis is that *knowledge* functions to flag good informants: we have a need to share information, thus a need to flag other subjects as reliable informants, and meeting this need is the function of *knowledge*. I will evaluate Craig's hypothesis, and his overall project, in the next chapter.

Fricker's "paradigm-based explanation" of blame can also be understood as conceptual reverse-engineering.[4] Fricker aims to illuminate our concept *blame* by "making explicit its most basic roles in our life" (2016: 165). She begins her project by imagining "a realistic conception of the most simple and basic form of the extant practice – a paradigm of the phenomenon". This paradigm will constitute an "explanatorily basic form" of the practice (165). But how does she identify the paradigm of our practices with *blame*? She notes that these practices demonstrate a huge amount of diversity. We sometimes apply the concept to ourselves, as in constructions like 'I blame myself for my marriage failing.' We sometimes apply *blame* in second-personal interactions, as in constructions like 'I blame you for your embarrassing behaviour at the conference.' And we sometimes apply *blame* in third-personal interactions, as in 'The government is to blame for the latest economic crisis.' Further, our practices with *blame* demonstrate diversity in their affective colour: sometimes an attribution of the concept is "little more than a dispassionate judgement that someone is blameworthy", while on other occasions "it is a judgement invested through-and-through with the deepest moral emotion" (167).

Given this diversity, Fricker doesn't point to any actual case in which we use *blame* as a paradigm. Rather, she "constructs" her paradigm case (166), by isolating features of some of our actual practices with *blame*. She hypothesises that amongst our actual practices with *blame*, "there is a basic second-personal interaction of X blaming Y for an action, motive or attitude (or lack thereof) from which other variant practices can be seen as derivative". Fricker calls this interaction "Communicative Blame" (167). In Communicative Blame, the point of X's blaming Y is to "inspire remorse in the wrongdoer, where remorse is understood as a pained moral perception of the wrong one has done" (167). This is Fricker's paradigm case. Her hypothesis will

[4] Queloz (2021) likewise understands Craig and Fricker as engaged in conceptual reverse-engineering.

be confirmed to the extent that she is able to account for or other practices with *blame* as derivative from this case. I am not going to evaluate Fricker's project of showing how other *blame* practices relate to the paradigm case. I wish to note simply that Fricker is well-understood as a theorist engaged in conceptual reverse-engineering, where conceptual reverse-engineering is, in turn, understood as an empirical method for confirming or disconfirming hypotheses about the function of a concept.

2.2. A taxonomy of conceptual reverse-engineering

In a co-authored paper (ms), Angela O'Sullivan and I offer a taxonomy of sub-methods of conceptual reverse-engineering on which methods are organised along two axes: first, whether the method gets its typical case through direct or indirect abstract representation; second, whether this representation has a time-axis. I will explain each of these axes in turn.

The aim of conceptual reverse-engineering is to confirm or disconfirm a hypothesis about the function of a concept. This is done by looking to what the concept does in a typical case. There are two ways to isolate a typical case of the conceptual practice. The first is to identify such a case from among actual instances of the conceptual practice; the second is to construct a hypothetical case. But for both approaches, the theorist must represent the case abstractly, including in her representation only features that are relevant, and omitting any merely incidental features. For example, it might be that whenever a given community makes use of the concept *justice*, all members of the community are wearing underwear. However, this is incidental to the conceptual practice: whatever the community is doing with the concept *justice*, they could do just as well without wearing underwear. Then it will not help the conceptual reverse-engineer to theorise about *justice* that she includes this feature in her representation of the typical case.

This distinction between two ways of isolating a typical case corresponds to a distinction found in the philosophy of science between *abstract direct representation* (ADR) and *modelling*. On the dominant picture of models, a model is an abstract indirect representation of a phenomenon (Weisberg 2007; Godfrey-Smith 2009). It is an *abstract* representation because it does not represent all features of the phenomenon. Rather, the modeller abstracts away all features of the phenomenon that are not relevant for her purposes; she "omits" irrelevant features from her representation of the phenomenon (Elliot-Graves 2020). The abstractly represented

real-world phenomenon is called the "target system" of the model (Weisberg 2007, 2013; Peschard 2010; Elliot-Graves 2020). A model is an *indirect* representation because its starting point is a description of the features of the target system with which the modeller is interested. She constructs the model out of this description, and then theorises about the model, only later (if at all) applying her findings to the target system itself. In contrast, in ADR, the theorist studies the target system directly. That is, she isolates some features of a real-world phenomenon, and theorises directly about this abstracted version of the phenomenon.

To see the difference, consider Michael Weisberg's examples of modelling vs. ADR. Weisberg gives Vito Volterra as an example of a modeller. Volterra (1926) wanted to understand why the population of sharks, rays and other predators in the Adriatic Sea increased, while the population of squid, cod and lobster decreased, during the First World War, despite the latter being fished much less than usual during this period (Weisberg 2007: 208). To do this, Volterra imagined a biological system composed of one population of predators and one population of prey, and attributed to each of these populations just a few properties, which he described using mathematical expressions. He then studied the dynamics of these two populations, and applied his findings to the real-world phenomenon: the populations of sharks, rays and other predators, and of squid, cod and lobster, in the Adriatic Sea during WWI. Volterra is a modeller because he constructs a representation of the real-world phenomenon of interest by stipulating properties of the phenomenon, representing these properties and then studying the representation, only later applying his findings to the target system.

In contrast, Dimitri Mendeleev's construction of the Periodic Table of Elements is an instance of ADR. Mendeleev wanted to develop a system for classifying the known elements, but there were many ways this could be done: in terms of their density, conductivity, ductility, melting point, or other chemically important properties (Weisberg 2007: 212). He decided to focus on atomic weight, valency (the combining ratio of an element; for example, carbon is tetravalent: it can combine with four equivalents of hydrogen), and isomorphism, where elements are isomorphic if families of salts containing chemically similar but distinct metals form similar crystal shapes (Brock 1992: 158). The ordering of the elements in terms of their weight and properties with which he ended up is what we now call the Periodic Table of Elements. The Periodic Table is thus an *abstract* representation of the real-world phenomenon, the known elements, because it includes only some features of the phenomenon (atomic weight,

valency, isomorphism), and abstracts away others (density, conductivity, and so on). It is nevertheless a *direct* representation of the real-world phenomenon: Mendeleev represented the elements directly, without the mediation of a model. The Periodic Table was not a mere organisational device for Mendeleev: it allowed him to make correct predictions that elements would be discovered in the future that would correspond to gaps in his table (Scerri 2001, 2006). Weisberg argues that this is a significant theoretical achievement, which Mendeleev was able to arrive at because of his ADR. Mendeleev had "no empirical knowledge that there were any empty slots to be filled", rather, he needed first "to hypothesize the existence of the missing elements by analyzing the theoretical structure he had created" (2007: 214). From studying his ADR, Mendeleev "was able to use the trends posited by the Periodic Table to make predictions about the properties of the 'missing' elements" (214).

O'Sullivan and I thus give the first choice-point for a method of conceptual reverse-engineering as whether that method represents its typical case using modelling or ADR. Hannon's (2019) method (to be discussed in much more detail in §2.4 of the next chapter) is an example of conceptual reverse-engineering that involves modelling. He begins his inquiry by describing the basic needs and abilities of a community of creatures like us, but who lack the concept *knowledge*. This is not a description of an actual case, but a hypothetical one. From this description, Hannon imagines a situation in which these creatures have a need to share information, and introduces the concept *knowledge* to meet this need; he then theorises about what this concept must look like in order to meet these needs. This leads him to posit a picture of the concept, which he then compares to the concept *knowledge* that we actually have. Thus Hannon constructs a model of our actual conceptual practices with *knowledge*: he describes features of creatures relevantly similar to us, but omits some features of our actual situation (most importantly, that we have the concept *knowledge*); he imagines a case in which these creatures have a particular need, and introduces a concept whose intension and extension enables it to serve this need; he later applies his findings about *knowledge* in this imagined case to our actual conceptual practices with *knowledge*.

In contrast, Fricker's conceptual reverse-engineering of *blame* involves ADR. Fricker begins her investigation of blame by isolating "the most simple and basic form of the extant practice" (2016: 165): that of communicative blame, whereby one person wrongs another, and in response the other person "lets [her] know with feeling that

[she is] at fault for it" (171). Fricker abstracts away all non-essential features that are present when one person blames another person for some bad action, but her central case is not hypothetical. Rather, it is an abstracted version of our actual dealings with the concept of blame. There are actual cases in which A wrongs B, and B lets A know, with feeling, that A wronged her. There is, of course, more to each actual case of blaming than this minimal description has it. But this is a description of actual practices we have involving the concept *blame*.

The second choice-point for a method of conceptual reverse-engineering is whether or not the model or ADR used includes a time-axis. Both ADR and modelling can be synchronic or diachronic: both can, but need not, involve a time-axis. Hannon's model and Fricker's ADR are synchronic, as they do not involve a time-axis. O'Sullivan and I interpret Craig's genealogy of *knowledge* as representing his central case using a diachronic model. Craig attempts to elucidate our actual concept *knowledge* by constructing a state of nature, consisting of creatures much like us, but who lack a concept of knowledge. Again, as for Hannon, the creatures that Craig describes are not supposed to be (descriptions of) real people. Rather, they are purely hypothetical people who resemble us in relevant ways (they are social, use language, need food and shelter, and so on) but are dissimilar to us in other ways (namely, lacking the concept *knowledge*). Craig theorises about why an ancestor of our concept *knowledge* would emerge in this state of nature, and considers what this ancestor concept *protoknowledge* (so-called by Kusch 2009) would look like. He then elaborates on this initial model by introducing further needs we should predict creatures in the state of nature to have, and imagines how the concept would evolve over time in response to these needs. Insofar as the end concept resembles our intuitive concept *knowledge* in relevant respects, Craig's original hypothesis about the point of the proto-practice is confirmed; the more plausible his genealogy – for example, the more basic and generic are the needs he introduces – the better his claim to having identified the function of our concept *knowledge*.

Finally, we offer Friedrich Nietzsche's genealogy of *punishment* as an example of a conceptual reverse-engineering project that involves diachronic ADR. Nietzsche is concerned with identifying the function(s) of our concept *punishment.* To do this, he tells a story about the origins of the concept and its evolution over time, which abstracts away from particular events to capture the essential features of historical changes. For this reason, he makes no reference to actual historical events in his genealogy of

punishment, but still conceives of himself as engaging with real history (see for example *GM* II §13-14). Unlike diachronic modellers, who begin their projects by constructing a hypothetical typical case and de-idealise by introducing real history only later in the model, if at all, on our reading Nietzsche begins by attempting to describe the real origins of our conceptual practice(s) with *punishment*. Just as synchronic ADR aims to capture the relevant features of a real-world conceptual practice by abstracting away its incidental features, so diachronic ADR aims to capture the real evolution of a concept by abstracting away from incidental features of its historical development. In the end, Nietzsche's diachronic ADR of *punishment* reveals it to have no single function, but to be "overladen with functions of all kinds" (*GM*, II, §14): there is no one function that explains why the concept *punishment* is there. As such, the genealogy makes it "impossible to say for sure *why* we actually punish" (§13): the complexity of the history of *punishment* makes us unable to identify a typical case from which we can glean the concept's function, and thereby come to better understand the concept.

To summarise, the taxonomy of sub-methods of conceptual reverse-engineering that O'Sullivan and I offer turns on two choice-points for representing the typical case: using ADR or modelling, and omitting or including a time-axis. Thus, on our taxonomy, all conceptual reverse-engineering is the reverse-engineering of a function from a typical case of the conceptual practice, and a conceptual reverse-engineer can choose to represent her typical case as:

1) a synchronic model, as in Hannon's approach to *knowledge*;
2) a diachronic model, as in Craig's genealogy of *knowledge*;
3) a synchronic ADR, as in Fricker's paradigm-based explanation of *blame*; or
4) a diachronic ADR, as in Nietzsche's genealogy of *punishment*.

2.3. The value of conceptual reverse-engineering for function-first philosophy

Conceptual reverse-engineering is a crucially important method for function-first philosophers. To be able to theorise about a concept via an examination of its function, the theorist must have a good idea of what its function is. Conceptual reverse-engineering, I have argued, should be thought of as a method for confirming or disconfirming hypotheses about the function of a concept. Thus the method of conceptual reverse-engineering is useful for the descriptive project of answering the question "What is the point of the concept?" that is central to the function-first approach (see Haslanger 1999: 95; Gardiner 2015: 31; Hannon 2019: 12). Once this

function is identified, the theorist can engage in the further descriptive project of "illuminat[ing] the contours of the concept itself" in terms of its ability to serve this function (Gardiner 2015: 31; see also Greco and Henderson 2015: 1). Conceptual reverse-engineering is also useful for explanatory and evaluative projects. By reconstructing the needs which the concept serves, and showing how the concept serves those needs, conceptual reverse-engineering can help explain why creatures like us have the concept in question. This is a crucial part of Craig (1990) and Hannon's (2019) approaches to *knowledge*, which I discuss in more detail in the next chapter; as well as Williams's (2002) approach to *truthfulness*. By making explicit the needs which a concept serves, conceptual reverse-engineering is also useful for the evaluative project of assessing whether those needs are ones which we reflectively endorse. This is a central part of Haslanger's ameliorative approach to race and gender concepts (1999, 2012a), as well as Fricker's (1998: 170) account of the secondary functions of *knowledge* (discussed in §2.4 of the last chapter). Conceptual reverse-engineering is thus a useful method for conducting many kinds of Q-inquiries that function-first philosophers are interested in.

Conceptual reverse-engineering shares a key advantage that Hannon claims for his method over conceptual analysis: relative invulnerability to counterexample. Conceptual reverse-engineering aims to find the function of a concept by examining the need(s) that the concept serves in a typical case. The theorist then uses this function to shed light on the shape of the concept: its intension and extension. Our pre-theoretical intuitions about the intension and extension of the concept play a role in conceptual reverse-engineering: the concept we end up with should resemble the intuitive concept to a sufficient degree. But some deviation from the intuitive concept is permitted, so long as the conceptual reverse-engineer can explain why our intuitions about the concept's intension or extension do not line up exactly with the intension and extension demarcated by the reverse-engineering project. Counterexamples are possible: there may be cases that fall under the intuitive extension of the concept, but not the extension we end up with at the end of our project; or cases that do not fall under the intuitive extension of the concept, but do fall under the extension of the concept at the end of the project. But these don't undermine the account of the concept generated so much as provide an explanatory challenge: the reverse-engineer must be able to plausibly explain the mismatch.

Such counterexamples can come to undermine a conceptual reverse-engineering project, however, if there is some alternative hypothesis about the function of a concept that, when plugged into a conceptual reverse-engineering project, generates a concept that is not subject to these counterexamples, yet does everything else at least as well: for example, which just as well or better explains a wide variety of our conceptual practices; which just as well or better explains other purposes to which we put the concept as derivative of its primary function; and so on. But to the extent that a conceptual reverse-engineer can explain away counterexamples, in the absence of alternative reverse-engineering projects that do not face these counterexamples but are otherwise equally or more plausible, these counterexamples will not undermine the conceptual reverse-engineering project. In contrast, any counterexample to a project of conceptual analysis undermines that project to some degree.

Conceptual reverse-engineering projects face a different kind of counterexample to conceptual analysis projects, however: cases where a concept fails to serve its posited function, and cases where the concept's posited function is served by some other concept. Whether these counterexamples undermine the project is again a matter of, first, whether they can be explained away, and second, whether alternative hypotheses about the function of the concept are available that do not face these counterexamples while doing everything else just as well.

One way of explaining away these kinds of counterexamples is to appeal to Millikanian proper function. If a concept's function is a proper function, then its function is whatever its ancestors did that contributed to the reproductive success of this family of concepts (the concept of interest plus its ancestors), which in turn explains the existence of this member of the family. An item can have a proper function without manifesting this function in all its uses, and even if, in some cases, some other item serves this function better. The cow's heart functions to pump blood even when it is outside the cow's body, on a human's plate; and a cow's heart functions to pump blood even if Daisy the cow has a mechanical heart, not a cow's heart. Similarly, a concept like *knowledge* could function to flag good informants even if some knowers are not good informants, and sometimes good informants are not knowers. So long as the cases in which the concept doesn't serve its hypothesised function, or in which some other concept serves this function, can be explained away, and there is no alternative reverse-engineering project of the concept in the offing which doesn't face

these counterexamples, these kinds of counterexamples don't undermine a conceptual reverse-engineering project.

3. Conceptual engineering

In this section, I outline the method of conceptual engineering, another method that is useful for answering the kinds of questions in which function-first philosophers are interested. In §3.1, I argue for a broader characterisation of the method than that which is developed by Herman Cappelen: where Cappelen holds that conceptual engineering must begin by identifying some defect in an existing concept, I argue that this isn't so. Rather, conceptual engineering is an ameliorative method, where the amelioration in question can either be fixing some defective concept, improving an existing concept, or constructing a new concept that better serves a function than any existing concept. These latter two forms of conceptual engineering don't begin from any conceptual defect. In §3.2, I discuss Rudolf Carnap's method of explication, and argue that this is a sub-method of conceptual engineering. In §3.3, I offer a taxonomy of different kinds of conceptual engineering. In §3.4, I explain the value of conceptual engineering for function-first philosophers.

3.1. What is conceptual engineering?

Cappelen is responsible for popularising the term 'conceptual engineering', though he did not introduce it. Cappelen (2018: 4) credits Simon Blackburn (1999) with introducing the term, however David Chalmers argues that credit ought to go to Richard Creath (1990), who "made a big deal of Carnap as a conceptual engineer" (2020: 6). But just as 'function-first philosophy' is a new name for something philosophers have been doing for some time, the name 'conceptual engineering' is (relatively) new, while the method is not. Rudolf Carnap's method of explication is taken by many to be a kind of conceptual engineering (Brun 2016; Cappelen 2018: 11; Cappelen and Plunkett 2020; Eder 2021: 4979), as is Haslanger's ameliorative method (Cappelen and Plunkett 2020; Brigandt and Rosario 2020), to name some influential examples.

Cappelen defines conceptual engineering as "the process of assessing and then ameliorating our concepts" (2020: 132), as well as, more broadly, "the process of assessing and improving our representational devices" (2018: 3), which may or may not be concepts. Conceptual engineering, on Cappelen's picture, begins with

identifying some defect in a concept or other representational device. Concepts (or other representational devices, like words) can be defective in many ways. They can be semantically defective, for example if they are incoherent or vague. This is what Kevin Scharp (2013) argues regarding the concept *truth* – specifically, that *truth* is inconsistent – and this defect motivates his revisionary project of replacing *truth* with two distinct concepts, *ascending truth* and *descending truth*. Concepts can be morally, socially or politically defective. For example, if the concept *marriage* excludes same-sex couples, this might have the morally, socially and/or politically bad effects of excluding these couples from moral, social and/or political goods or rights to which opposite-sex couples have access (Cappelen 2018: 34). And concepts can have cognitively objectional effects, for example if they license the use of generics that interfere with the cognitive processes of those who endorse them, leading them to endorse mistaken generalisations about social kinds (Leslie 2017). Once a defect in a concept has been identified, the conceptual engineer begins ameliorating the concept, either by revising or replacing the concept, or abandoning it altogether. This final strategy counts as ameliorative as it eliminates a defective concept from our conceptual repertoire, thereby improving the repertoire overall (Cappelen 2018: 35).

However, Cappelen's characterisation of conceptual engineering is too narrow. For Cappelen, conceptual engineering always begins with the identification of a conceptual defect. But as Mona Simion and Christoph Kelp note, eliminating a defect is not the only way of improving something, and "it is widely agreed in the theory of normativity that in order to justifiably embark on a certain project, such as a conceptual engineering project, all that's needed is improvement, not fixing a defect" (2020: 987). For this reason, Simion and Kelp argue that conceptual engineering needn't consist only in fixing defective concepts or eliminating such concepts from our conceptual repertoire; rather, we should broaden our understanding of the method so that it includes any way of "improving the world of concepts" (988).

Chalmers offers a broader characterisation of conceptual engineering than Cappelen. Chalmers develops his characterisation by looking to dictionary definitions of 'engineering', identifying a common core among these definitions of 'engineering' as "the process of utilizing knowledge and principles to design, build and analyze objects" (2020: 2). Invoking compositionality, and applying this to concepts, Chalmers attempts a first pass at a definition of 'conceptual engineering' as "the process of designing, building and analyzing concepts" (2). But this definition is potentially

confusing, as the notion of 'analysing concepts' is, for contemporary Anglo-American philosophers, closely tied to the method of conceptual analysis. So Chalmers suggests we replace talk of 'analysing concepts' with that of 'evaluating concepts'. Further, he suggests that "maybe 'implementing' is better than 'building' where concepts are concerned" (2). He doesn't give an argument for this, but the idea is likely something like the following: if by 'designing a concept' we mean specifying its intension, then this is really all it takes to 'build' the concept, too. The relevant work for bringing the concept into the world (as, for example, building a bridge brings its design into the world) is implementing the concept: using it, and encouraging others to use it, too. (Cappelen would call this work "conceptual activism", rather than conceptual engineering (2018: 60).) Making these tweaks, Chalmers gets the following definition of 'conceptual engineering' as "the process of designing, implementing and evaluating concepts" (2).

Chalmers's characterisation of conceptual engineering is broader than Cappelen's, as it includes both improving existing concepts and creating new ones as projects within conceptual engineering. Chalmers writes that, regarding more familiar forms of engineering, it would be bizarre to rule out the creation of new objects as part of engineering: though "fixing and improving our deficient bridges" is certainly part of bridge engineering, "it's not the only part, and it's probably not the most important or exciting part." Rather, the project of building new bridges is, for Chalmers, the "paradigm case of bridge engineering" (2020: 6). Applying this insight to conceptual engineering: though improving existing concepts is a kind of conceptual engineering, so too should be the creation of new concepts.[5] Thus Chalmers distinguishes between two kinds of engineering projects: *de novo* engineering, which is "building a new bridge, program, concept, or whatever", and re-engineering, which is "fixing or replacing an old bridge, program, concept, or whatever" (6). *De novo* conceptual engineering is the creation of a new concept; conceptual re-engineering is the revision or replacement of

[5] An important difference between bridge engineering and conceptual engineering is that we already had many concepts prior to engaging in (at least deliberate) conceptual engineering, but we didn't have lots of bridges around before engaging in bridge engineering. This might make us sceptical whether Chalmers's claim that building a new bridge is the 'paradigm case' of bridge engineering has any bearing on what would be the 'paradigm case' of conceptual engineering: improving existing concepts or creating new ones. But this doesn't matter for our purposes: neither project needs to be more paradigmatic than the other for these distinctions to be meaningful and useful.

an old one. For Chalmers, it isn't necessary for either project to begin by identifying a defect in an existing concept.

3.2. Explication

As noted in the previous section, Carnap's method of explication is taken by many to be an early example of conceptual engineering. Explication, for Carnap, is "[t]he task of making more exact a vague or not quite exact concept used in everyday life or in an earlier stage of scientific or logical development, or rather of replacing it by a newly constructed, more exact concept" (1947: 7-8). Explication starts by identifying a conceptual defect: the concept the theorist begins with, which Carnap calls the *explicandum*, is "vague or not quite exact". The theorist fixes this defect by replacing the *explicandum* with another, more exact, concept, the *explicatum*, which can be used in place of the *explicandum* in relevant contexts. The *explicandum* and the *explicatum* are different concepts, with different intensions and perhaps different extensions. These concepts needn't closely resemble each other in either intension or extension, as long as the *explicatum* can play the same role(s) in relevant theoretical contexts as the *explicandum*. Carnap writes, "it is not required that an explicatum have, as nearly as possible, the same meaning as an explicandum; it should, however, correspond to the explicandum in such a way that it can be used instead of the latter" (1947: 8).

Georg Brun understands Carnapian explication as involving two phases. First, the *explicandum* must be identified as clearly as possible. This might involve disambiguating the word or term used for the concept, if it is ambiguous. Brun calls this expression the "explicandum-term", noting that Carnap does not distinguish the *explicandum* and *explicatum* concepts from the terms we use for them (2016: 1215; see also Eder 2021: 4979). However, the theorist won't be able to exactly identify the *explicandum*, as it is inexact (Carnap 1950: 4); indeed, inexactness in the *explicandum* is the conceptual defect that one is trying to fix in giving the explication. As such, the *explicandum* must be characterised more informally, for example by pointing to cases in which the *explicandum* either clearly does or clearly does not apply. In the second phase of explication, an *explicatum* is introduced. This phase involves specifying rules for the use of the *explicatum* in the target system of concepts (3). For example, the theorist might give a definition of the *explicatum*, although less strict methods of concept introduction are also acceptable.

Notice that I said 'an *explicatum*', not 'the *explicatum*'. Since the *explicandum* is not exact, there is not one single *explicatum* that could be used in its place. Rather, there will be a number of candidate *explicata*, which will be more or less adequate, from which the theorist can choose. Adequacy of an *explicatum* is, for Carnap, a matter of satisfying the following four criteria to a sufficient degree. The first is similarity to the *explicandum*, in the sense that the theorist must be able to use the *explicatum* instead of the *explicandum* in relevant contexts (1950: 5). The second is exactness: rules for the use of the *explicatum* within the theory must be formulated as exactly as possible, "for instance, in the form of a definition … so as to introduce the explicatum into a well-connected system of scientific concepts" (7); these rules should eliminate ambiguity (4) and must not lead to paradox or contradiction (1963: 935); and the *explicatum* must be less vague than the *explicandum*, in the sense of there being fewer cases where it is unclear whether the *explicatum* applies than there are regarding the *explicandum* (1950: 5). Third, the *explicatum* must be as fruitful as possible: using the *explicatum*, the theorist must be able to formulate many laws and generalisations (6). If two candidate *explicata* meet similarity, exactness and fruitfulness to a similar degree, then the final criterion of simplicity can be used to choose between them. This includes both how simple are the rules for using the *explicatum*, for example its definition, as well as how simple are the forms of the laws which can be formulated using the *explicatum* (7; see also Brun 2016: 1215).

As an example of explication in action, Carnap offers biologists' replacement of the everyday concept *fish* with the concept *piscis*. The concept *fish* that is used by, for example, fisherpeople and cooks, has an intension whose content is something like 'animal living in water', and so includes in its extension whales and seals (1950: 6). This concept is not sufficiently exact for scientific purposes. As such, biologists replaced the concept *fish* with the more exact concept *piscis*, which means "aquatic [vertebrate that has] gills through life and limbs, if any, in the shape of fins" (Nelson 2006: 2). (Note that the biological concept *piscis* is still referred to by biologists using the *explicatum*-term 'fish'; Carnap uses different terms for the *explicandum* and the *explicatum* simply to "avoid confusion" (1950: 6).) *Piscis* is more exact than *fish*, in that, first, it is given an explicit definition, and second, it will have fewer borderline cases (consider: how much of an animal's life must be spent in water for it to count as an animal that lives in water? A seal doesn't spend all its life in water, but nevertheless counts as a fish on the *explicandum fish*. Does a duck count as a fish on this inexact

concept?) *Piscis* cannot be used in place of *fish* in all contexts. For example, if a fisherperson wanted to talk about all the animals living in the sea, *piscis* could not serve this purpose as well as *fish*. But *piscis* can be used in place of *fish* in contexts that are relevant to biologists. The concept *fish* has thus been "succeeded by [*piscis*] in this sense: the former is no longer necessary in scientific talk; most of what previously was said with the former can now be said with the help of the latter (though often in a different form, not by simple replacement)" (6). The concept *piscis* is more fruitful than *fish*: it can be better "brought into connection with other concepts on the basis of observed facts" than can *fish*, and more laws can be formulated using *piscis* than using *fish* (6).

Carnap calls *explicanda* "pre-scientific" and *explicata* "scientific" (for example 1945: 513; 1950: 3, 5, 6). His terminology shouldn't lead us to think that, for Carnap, explication is a method for science alone. Carnap thinks of explication as available to, and already used by, theorists of different kinds: scientists, yes, but mathematicians and philosophers, too. Some of his examples of explication come from philosophy, for example he discusses Alfred Tarski (1933) as explicating the concept *true* (1950: 5), and Gottlob Frege and Bertrand Russell as explicating natural numbers (17) and definite descriptions (1947: 8). For this reason, Brun suggests reading Carnap's use of 'pre-scientific' and 'scientific' as meaning 'pre-theoretical' – or better, 'less theoretical', as one can explicate a concept that is already part of a theoretical system if it is inexact (Carnap 1950: 7) – and 'theoretical', respectively (Brun 2016: 1217).

Haslanger's ameliorative method can be thought of as explication. Explication involves concept replacement, rather than concept revision: the *explicatum* is a distinct concept to the *explicandum*. Haslanger seems to think of her method as involving concept replacement, rather than revision, too. Haslanger writes:

> Some analytical6 projects are oriented towards theoretical concepts: the concept X is explicitly introduced or adopted as a theoretical tool within a larger inquiry, where the emphasis in determining the content of the concept is placed on the theoretical role it is being asked to play. But an analytical approach is also possible in exploring non-(or less-)theoretical concepts if we are willing to accept an answer to the question "What is X?" that does not exactly capture our

[6] Recall from §2.1 of the previous chapter that Haslanger later renames her method the 'ameliorative approach', and projects undertaken using this method 'ameliorative projects'.

> intuitive concept of X, but instead offers a neighboring concept that serves our legitimate and well-conceived purposes better than the ordinary one. (1999: 468)

When applying Haslanger's ameliorative method to the question 'What is X?' for some theoretical concept X (where a 'theoretical concept' is a concept that is introduced into a theoretical context to play some specified role; compare Carnap use of 'scientific' to describe *explicata*), a concept is 'explicitly introduced or adopted'; when taking this approach to a less theoretical concept (compare Carnap's use of 'pre-scientific'), one 'offers a neighbo[u]ring concept' to the everyday one to be used for (some of) the same purposes. In either case, the output of an ameliorative project is a different concept to that which constituted its input: the original concept is replaced by another which better serves some particular purpose(s). But it should be noted that Haslanger's ameliorative method is only a kind of explication if the output concept is introduced, even in the 'less theoretical' case, into a system of concepts, and rules for its use within this concept are specified.

Chalmers classes Carnapian explication as a kind of conceptual re-engineering. Recall his distinction between *de novo* engineering and re-engineering: "*De novo* engineering is building a new bridge, program, concept, or whatever. Re-engineering is fixing or replacing an old bridge, program, concept, or whatever" (2020: 7). On this way of making the distinction, Chalmers is correct to hold that "[c]ertainly the Carnapian explication literature is very much a literature on re-engineering" (7): explication is conceptual re-engineering, as it is a method for replacing existing ('old') concepts.

However, I think Chalmers's distinction between *de novo* (conceptual) engineering and (conceptual) re-engineering is inadequate, for two reasons. Both turn on his including *replacing* an old concept, bridge, program, or whatever with a new one as a kind of re-engineering. The first reason is that the distinction between *de novo* engineering and re-engineering will be blurry. Replacing an old concept, bridge, whatever with a new one will often (in the case of bridges, pretty much always) involve constructing a new concept, bridge, whatever. Then re-engineering will often involve *de novo* engineering, thus the distinction is not robust. The second, and in my opinion the more substantial, reason is that Chalmers's way of drawing the distinction has odd consequences for the ontologies of concepts, bridges, whatever. If replacing an old concept, bridge, whatever with a new one counts as a re-engineering of that concept, bridge, or whatever, then we will have to say that the new concept, bridge, whatever is

simply the old concept, bridge, whatever, re-engineered. Applied to non-conceptual objects, that sounds very odd, especially if the original object continues to exist.

Chalmers himself offers an example that demonstrates this oddness:

> Take the Tappan Zee Bridge, just up the Hudson River from New York City. The old Tappan Zee bridge is still there for now. They're building a new bridge in the same location as the old bridge. Is that *de novo* engineering because it's a new bridge? Or is it re-engineering because it's a replacement? For my purposes I'm going to count that kind of thing as re-engineering, because the new bridge is being used to fix an old bridge (2020: 7).

Chalmers holds that this is a "hard case" that shows that it is "not totally straightforward" to draw and maintain the boundary between *de novo* engineering and re-engineering (6-7). But this doesn't seem to me to be a 'hard case' whatsoever. A new bridge without the old bridge's problems has been built, but the old bridge, with all its problems, still exists (or at least it did at the time of Chalmers's writing; it has since been demolished). These are two different objects, on pretty much any way that we might draw up individuation conditions for objects: they are constituted by entirely different material at all times at which they exist; they came into and will go out of existence at different times; to demolish one would not be to demolish the other, so they have different modal properties. But claiming that the new bridge is a 're-engineering' of the old bridge seems to imply that the new bridge is identical to the old bridge: it is literally that bridge, just 're-engineered' – just *fixed*. Thus Chalmers seems committed to a very odd claim about the ontology of bridges: there aren't two different bridges here, despite what our best ways of individuating objects suggest.

Even worse for Chalmers: the new Tappan Zee Bridge doesn't 'fix' the old bridge at all. The only way to make sense of Chalmers's claim that the new bridge is 'being used to fix an old bridge' is that the new bridge offers a solution to the same problem to which the old bridge once, but no longer adequately, provided the solution: crossing the Hudson River at roughly this place. But this is to fix the problem, not the bridge itself. Then Chalmers's justification for calling this 're-engineering' falls flat: the new bridge isn't being used to fix the old bridge. What we should say about this case, I think, is that there is no re-engineering going on, because replacement is not re-engineering.

Chalmers could respond that the sense in which *de novo* engineering is *de novo* is not that its object is new, and the sense in which re-engineering is re-engineering is

not that its object is not new. Rather, what is new in *de novo* engineering is the overall project. Recall Chalmers's definition of engineering as "the process of utilizing knowledge and principles to design, build and analyze objects" (2). He could hold that *de novo* engineering occurs when this process begins from scratch to solve a problem, improve a situation, and so on, for the first time. Re-engineering will be returning to an engineering project that has already begun. This might involve constructing new objects, but this will nevertheless be part of an old project: a process of utilising knowledge and principles to design, build and analyse objects to serve some purpose that is already ongoing.

This line of argument is not compatible with the view that Chalmers articulates in his paper. He writes that "[*d*]*e novo* engineering is building a new bridge, program, concept, or whatever. Re-engineering is fixing or replacing an old bridge, program, concept, or whatever" (7). What is new in *de novo* engineering *is* the bridge, the program, the concept; what is not new in re-engineering is the bridge, the program, the concept. So making this kind of argument would be to revise his original conception of (conceptual) engineering. But that's okay: perhaps this better captures the underlying distinction between different kinds of engineering that Chalmers is interested in; then he would have reason to revise his position in this way.

However, I think there is reason to doubt that this better captures the distinction that Chalmers is interested in. For on this way of making the distinction between *de novo* engineering and re-engineering, genuine *de novo* engineering would be a relatively rare occurrence, both in more familiar kinds of engineering and in conceptual engineering. In order for an instance of engineering to count as *de novo* engineering, not only must there not already be an existing engineered object, for example a bridge, program or concept, there must have been no attempt to serve the particular purpose(s) for which such an object has been developed. No one may have tried to cross the Hudson River at that point before; no one may have tried to do with a concept the kinds of things which the newly constructed concept is intended to do.

This isn't true for some of Chalmers's central examples of *de novo* engineering. For example, he takes the creation of the concept *supervenience* to be an example of *de novo* conceptual engineering. In creating this concept, philosophers "weren't particularly trying to fix or replace other concepts" (7). Nevertheless, the concept was created as a way of solving problems which philosophers had tried to solve with other concepts. *Supervenience* was supposed to be able to do "a lot of philosophical work

that previous concepts like identity might have been hoped to do" (4). But that *supervenience* was created within the context of already-existing problems with already-existing (but inadequate) solutions does not undermine its claim to being *de novo* engineering, for Chalmers:

> If you squint really hard, you might say that supervenience is intended as a replacement for identity. But that's not quite right. The concept of identity is doing fine. It's just that there's a job people were using identity for, in some reductive projects, that people then tried to use supervenience to do. (7)

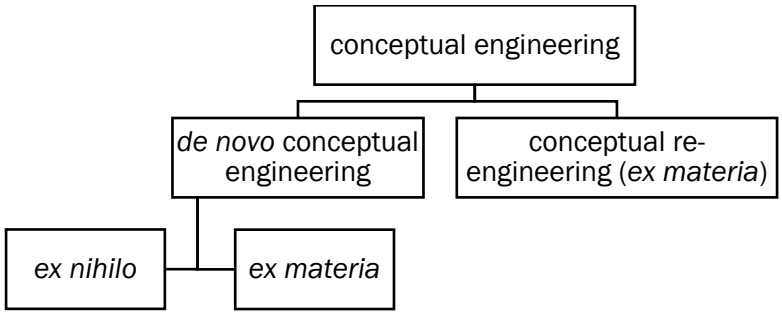
However, on the characterisation of *de novo* engineering that we are considering, this is exactly what would prevent this from being *de novo* conceptual engineering, rather than conceptual re-engineering. In philosophy, it will be hard to find cases where no one has ever tried to solve the kinds of problems which a new concept is introduced to fix. Hence it will be hard to find cases of *de novo* conceptual engineering. But for Chalmers, "*de novo* conceptual engineering is often the most fruitful kind" of conceptual engineering (16). This suggests that this way of making the distinction between *de novo* (conceptual) engineering and (conceptual) re-engineering does not better capture the underlying distinction in which Chalmers is interested.

3.3. A better taxonomy of conceptual engineering

I suggest that we use '*de novo* (conceptual) engineering' and '(conceptual) re-engineering' in the way that Chalmers explicitly suggests, except without replacement counting as re-engineering. Then *de novo* (conceptual) engineering will be building a new bridge, program, concept, and so on, while (conceptual) re-engineering will be fixing or otherwise improving an existing bridge, program, concept, and so on. The construction of the new Tappan Zee Bridge will not count as bridge re-engineering, but as *de novo* bridge engineering; and Carnapian explication will not count as conceptual re-engineering, but as *de novo* conceptual engineering. Although the new Tappan Zee Bridge serves the same purpose as the old bridge – to allow people to cross the Hudson River at roughly this point – it isn't a re-engineering of the old bridge, as it doesn't fix the old bridge, but replace it. Similarly, the *explicatum* in explication serves the same purpose, or one of the same purposes, as the *explicandum*, but it isn't a re-engineering of the old concept: it doesn't fix, but replace, the old concept. In contrast, the ongoing repairs to Edinburgh's North Bridge are an instance of bridge re-engineering; while the

expansion of the concept *marriage* to include same-sex couples is an instance of conceptual re-engineering (at least this is the most natural way of understanding this case; however in §4.1 I consider whether it is possible to change a concept's intension and extension in this way and still have the concept with which one began).

Further, I suggest we introduce two terms to capture another distinction that Chalmers is interested in. Chalmers wants to capture in his *de novo* engineering/re-engineering distinction the following idea: sometimes, there is some item (such as a concept) already in use that we want to improve or replace, but other times we have needs which are not currently served by any existing item. This can be captured by introducing a distinction between *ex nihilo* and *ex materia* (conceptual) engineering. When we are creating a new item (for example, a new concept) where previously there was nothing, we are doing *ex nihilo* (conceptual) engineering; when we are either fixing or improving an existing item (such as a concept) or creating a new item to replace an old one, we are doing *ex materia* (conceptual) engineering. All (conceptual) re-engineering will be *ex materia* (conceptual) engineering: one cannot re-engineer what isn't already there. *De novo* (conceptual) engineering can be either *ex materia* or *ex nihilo*. A diagram may be helpful:



Carnapian explication will count as *de novo* conceptual engineering *ex materia*: the theorist begins with a defective concept, the *explicandum*, and replaces this with a concept that doesn't have the defect, the *explicatum*, which can serve (some of) the same purpose(s). Likewise, Haslanger's ameliorative approach to the concept *woman* is *de novo* conceptual engineering *ex materia*: a concept *woman* already exists, but doesn't serve Haslanger's feminist goals; she introduces a new concept *woman* to replace the existing concept which can serve those goals.

I suggest that an example of *de novo* conceptual engineering *ex nihilo* is the development of concepts for non-binary gender identities: gender identities that go beyond the discrete categories *man* and *woman*. For example, Robin Dembroff provides an account of the "gender kind" genderqueer (2020: 2). Though gender kinds, like other social kinds, are metaphysical rather than conceptual entities, Dembroff articulates a concept of that gender kind and thus can be understood as doing conceptual engineering. Dembroff is not the first person to use the term 'genderqueer', but this language is still relatively new: Dembroff cites trans activist Riki Wilchins as coining the term in the 1990s. We can understand the project of establishing a new concept *genderqueer*, which Dembroff has taken up but which has been in the works for some decades, as an instance of *de novo* conceptual engineering *ex nihilo*: before trans activists and theorists began thinking in these terms, there were not concepts serving the needs in the relevant epistemic communities that led to the creation of this gender concept; specifically, the need for "tools for understanding … new and quickly evolving gender identifications" (Dembroff 2020: 3), as well as the political goal of destabilising the dominant conception in Western cultures of gender as an exhaustive binary consisting of the discrete categories *man* and *woman* (16).

3.4. The value of conceptual engineering for function-first philosophy

A function-first philosopher has clear use for conceptual engineering. If considerations about the function(s) of a concept are at the forefront of a theorist's approach to a concept, then it is natural for her to ask whether the concept in which she is interested is serving its legitimate function(s) as well as it could. If she determines that it is not, she ought to try to find out whether and how the concept could be revised to better serve its function(s), or whether it could be replaced by a different concept that better serves the same function(s). If it can be so revised or replaced, a function-firster might try to do just that. Then she is engaged in conceptual engineering, on the characterisation of the method outlined here.

Cappelen and David Plunkett hold that even if a philosopher doesn't engage in any revisionary work herself – that is, even if she doesn't attempt to revise or replace the concept in order for the concept's legitimate function(s) to be better served – simply evaluating a concept "in an ameliorative spirit" is enough to be engaged in conceptual engineering (Cappelen 2020: 132). Cappelen and David Plunkett characterise conceptual engineering as having three steps:

1. The assessment of concepts,
2. Reflections on and proposals for how to improve concepts,
3. Efforts to implement the proposed improvements (2020: 3).

A conceptual engineer needn't engage in all three steps. Instead, some conceptual engineers "focus on discovering defects, some on ameliorative strategies, others on conceptual activism, and yet others want to do the whole shebang" (3-4). (Recall that on my understanding of conceptual engineering, there needn't be any conceptual defects for conceptual engineering to be possible and appropriate.) If we endorse this three-step picture of conceptual engineering, then it is even clearer that function-first philosophers have use for conceptual engineering: simply asking the kinds of questions that motivate taking the function-first approach to a concept is enough to make a theorist a conceptual engineer.

Conceptual engineering shares the advantage that conceptual reverse-engineering has over conceptual analysis of being relatively invulnerable to counterexample. Though our pre-theoretical intuitions about a concept's intension and extension play some role in conceptual engineering, this role is much less significant than in conceptual analysis. Further, the role of intuitions is less significant here than in conceptual reverse-engineering, too. For it is only in conceptual re-engineering that our intuitions about which objects fall under the extension of the concept should be taken into consideration: the post-revision concept at the end of a process of conceptual re-engineering should be recognisable as the same concept with which the theorist began. But even this doesn't mean that we cannot tolerate any differences in the intuitive extension of the concept pre-revision and its explicit extension post-revision. Indeed, difference in these two extensions is sometimes exactly what we want: in revising the concept *marriage* so that it now applies to same-sex couples where previously it only applied to opposite-sex couples, we have changed the extension of the concept; but the concept is still recognisable as the same concept.

In *de novo* conceptual engineering, the theorist introduces a new concept. Where this is a replacement of an already existing concept – that is, in *ex materia de novo* conceptual engineering – the concept needn't be recognisable as the same concept with which the theorist began, because it isn't the same concept. Counterexamples to *de novo* conceptual engineering projects are still possible, but these counterexamples cannot appeal to our pre-theoretical intuitions about what should and should not fall under the extension of the ordinary concept, as this concept

is no longer at work. There must be some other reason for thinking that an object excluded from the explicit extension ought to be included, or an object included in the explicit extension ought to be excluded; for example, that the concept could better serve its function if this object was included/excluded.

For example, recall from §3.1 of the previous chapter that it is a consequence of Haslanger's conceptual engineering of the concept *woman* that the Queen of England may fail to count as a woman, as she does not seem to be subordinated in the relevant way. This does not undermine Haslanger's conceptual engineering project, because her new concept *woman* better serves the purpose she wants it to: helping us to identify and overcome gender-based oppression in society. What matters for Haslanger is not whether her concept *woman* can capture in its extension all and only those people who we pre-theoretically think of as women. What matters is whether someone's being excluded (or included) from the concept "is in conflict with the feminist values that motivate the inquiry" (46). In contrast, Haslanger takes very seriously the criticism of her project put forward by Katharine Jenkins (2016: 398-400), that her concept *woman* excludes some trans women from its extension. This is because she does see this exclusion, unlike that of the Queen of England's, as in conflict with the feminist values motivating her project; thus Haslanger takes this counterexample to represent a genuine problem for her account of gender (Haslanger 2020: 236-7). So conceptual engineering is not totally invulnerable to counterexample, but it is not sufficient for a case to constitute a counterexample to a conceptual engineering project that our pre-theoretical intuitions about the extension of the ordinary concept would include/exclude some object that the end-product concept excludes/includes.

4. Objections and replies

In this section, I consider two objections to conceptual engineering that have both been articulated under the name of the "Strawsonian challenge" or "Strawson's challenge" to conceptual engineering (see Cappelen 2018: 104; Thomasson 2020: 442). The first objection says that to change a concept in order to solve some philosophical problem, or advance some philosophical debate, is to change the topic in such a way that one is no longer talking about the same problem or engaging in the same debate. As such, the objection goes, we cannot solve philosophical problems or advance philosophical debates using the method of conceptual engineering. The second objection says that

concepts are individuated by their intensions and extensions, so the idea that we can change a concept's intension or extension and still have numerically the same concept is coherent. Then conceptual engineering, or at least conceptual re-engineering, is impossible: any revision of a concept is, in fact, abandonment of the concept in favour of a new one. These objections are related, so I consider them together (§4.1). I do not raise any specific objections to conceptual reverse-engineering in this chapter. I will consider and respond to some objections to conceptual reverse-engineering after discussing in depth some examples of the method in the next chapter (in §3).

4.1. The Strawsonian challenge to conceptual engineering

Two different challenges to conceptual engineering have been articulated under the name of the 'Strawsonian challenge' to conceptual engineering. The first version says: in changing a concept to solve a philosophical problem, advance a philosophical debate, and so on, we change the topic; thus we cannot solve the same problem, engage in the same debate, and so on, that was originally under consideration; then conceptual engineering cannot solve philosophical problems, advance philosophical debates, and so on. The second version says: in changing a concept's intension or extension, we destroy that concept and create a new one, so conceptual engineering, or at least conceptual re-engineering, is incoherent. Call these the 'topic-preservation' challenge and the 'incoherence' challenge, respectively. The topic-preservation challenge is a generalised form of Peter Strawson's challenge to Carnap's method of explication. The incoherence challenge has been articulated by Cappelen and Thomasson. I survey and evaluate some replies to both versions of the challenge.

Strawson objects to Carnap that to explicate a concept in order to solve a philosophical problem is simply to change the subject, and as such, cannot solve the original problem. This objection has two aspects. The first specifically concerns using explication to solve *philosophical* problems: Strawson holds that explication involves replacing philosophical concepts with scientific concepts, which don't and cannot solve the problems that philosophers are interested in. This part of the objection is easily resisted. The second aspect of the objection is that changing a concept means changing the subject in such a way that one is no longer talking about the same things as those which featured in the original formulation of the problem, and thus cannot constitute a solution to the very same problem. This part of the objection is harder to resist, but it can be done.

Here is Strawson's articulation of the objection. He writes that "it seems prima facie evident that to offer formal explanations of key terms of scientific theories to one who seeks philosophical illumination of essential concepts of non-scientific discourse is to do something utterly irrelevant – it is a sheer misunderstanding, like offering a text-book on physiology to someone who says (with a sigh) that he wished he understood the workings of the human heart" (1963: 505). On Strawson's reading of Carnap, to explicate a concept is to take a philosophically interesting concept and replace it with a non-philosophical, thus non-philosophically interesting, concept: a scientific concept. Strawson holds that scientific concepts cannot "serv[e] just the same purposes" as philosophical concepts, so any attempt to solve philosophical problems by explicating concepts would result in "something so radically different from the original [concept] that it could no longer be said to be fulfilling the same purpose, doing the same thing" (1963: 505). Therefore, explication is not a useful method for solving philosophical problems: to explicate "is not to solve the typical problem, but to change the subject" (506).

Here we see both aspects of Strawson's objection: that scientific concepts do not serve the same purposes as philosophical ones, so cannot be used in their place; and that to change a concept is to change the topic in such a way that one can no longer be engaged in the same project (for example, attempting to solve the same problem) as one was previously. The first aspect is easily resisted when we recall from §3.2 that Carnap's use of 'scientific' is best understood as meaning 'theoretical': part of a theory, but not necessarily a scientific theory. Then to explicate a philosophically interesting concept does not mean replacing the concept with a non-philosophical concept: philosophical concepts are theoretical concepts. However, the second aspect cannot so easily be done away with. I will consider some responses that have been made to this aspect of Strawson's challenge, but a generalised version of the challenge that is meant to apply to conceptual engineering more broadly, not just to explication.

Haslanger is concerned with Strawson's challenge insofar as it applies to her own ameliorative method. She articulates a version of the objection that does not refer to '(pre-)scientific concepts', and so more obviously applies to her method: "In asking what race is, or what gender is, our initial questions are expressed in everyday vocabularies of race and gender, so how can we meaningfully answer these questions without owing obedience to the everyday concepts?" (2000: 34). Generalising the problem to apply to all revisionary approaches to concepts, she writes that "it isn't

entirely clear when a project … is no longer even revisionary but simply changes the subject" (34). Similarly, Cappelen articulates a version of Strawson's challenge that applies to conceptual engineering in general: "Change of extension and intension … is a change of topic, so revisionary projects are bound to fail. Even if the revisions succeed, they do not provide us with a better way to talk about what we were talking about; they simply change the topic" (2018: 100). Though Cappelen here talks in terms of 'revision', the objection applies to both conceptual re-engineering and *de novo* conceptual engineering: both revising a concept and replacing it with a different concept involve a change of extension and intension, and so (the objection goes) a change in topic; thus neither kind of conceptual engineering project can "provide us with a better way to talk about what we were talking about", both "simply change the topic". Further, this means that there cannot be substantive disagreement (as opposed to merely verbal disagreement, see Chalmers 2011) about whether the product of a conceptual engineering project is an improvement of the concept of interest: if this objection is correct, then those who make use of the pre-engineered concept and those who make use of the post-engineered concept are simply talking past each other, rather than engaging in debate over a single subject matter (see Sawyer 2020: 384).

Another way of understanding the Strawsonian challenge to conceptual engineering is as undermining the very possibility of revising concepts. If the identity of a concept is determined by its intension and extension, this challenge says, then it is not possible to revise concepts: any change in intension or extension means abandoning the old concept and creating a new one. In other words, any change to a concept's intension or extension means numerical change of concept. This challenge is articulated by Cappelen:

> *You can't improve on a concept by changing its intension and extension because that very idea is incoherent*. Concepts have their intensions and extensions *essentially*. So a change in intension or extension always involves abandoning a concept, and can never be an improvement of the old concept. (2018: 104)

As well as by Thomasson:

> … one cannot improve on a concept by changing its intension and extension, since concepts possess these essentially, ensuring that any such changes leave

> us with new concepts rather than improvements of the old concepts. (2020: 442)

Stephen Koch calls this the "metaphysical" interpretation of the Strawsonian challenge (2021: 5).

The first way of understanding the Strawsonian challenge to conceptual engineering, the 'topic-preservation' challenge, applies to both kinds of conceptual engineering projects: it says that we cannot use conceptual engineering to solve philosophical problems, nor indeed to participate in any philosophical projects, as this would mean changing the topic so that the original problem or project is no longer under consideration. The second way of understanding the Strawsonian challenge, the 'incoherence' challenge, applies only to conceptual re-engineering: it says that the very notion of conceptual re-engineering is incoherent; any change to a concept's intension or extension means replacement of the original concept with a new one.

Note that Strawson does not make the incoherence challenge to Carnap's explication. This makes sense: Carnap doesn't claim that explicating a concept leaves one with the same concept with which one began. Indeed, he explicitly denies this: in explication, one replaces the pre- or less-theoretical concept, the *explicandum*, with a more exact theoretical concept, the *explicatum*, which can be used in place of the *explicandum* in relevant contexts. So the incoherence challenge is not applicable to explication. Nor is it applicable to *de novo* conceptual engineering in general. It is an objection to the possibility of conceptual re-engineering: changing a concept's intension or extension and still having (an improved version of) the same concept.

4.2. Cappelen's samesaying response

Cappelen responds to the topic-preservation challenge to conceptual engineering that conceptual engineering can be topic-preserving: conceptual engineers can use the end-product of their conceptual engineering project to talk and think about the same topic that motivated the project. His argument appeals to "samesaying": the phenomenon of different speakers having said the same thing. Two speakers A and B are "samesayers" for Cappelen if there's some proposition P such that A and B have both said that P (2018: 107).

Cappelen's response from samesaying has two steps. First, he argues that two speakers A and B can samesay using a sentence 'Fa', even if the extension of 'F' differs

in A's speech to B's. He gives the example of sentences involving gradable adjectives: adjectives that can apply in greater or lesser degrees, such as 'cold', 'tall', or 'interesting'. Gradable adjectives are context-sensitive: whether a gradable adjective (say, 'tall') is true of some object (say, a 6'0" man) depends on the context of utterance (say, in the pub vs. at basketball try-outs). Cappelen holds that, whatever are the mechanisms that determine the truth-conditional contribution that gradable adjectives make to the sentences in which they feature, these mechanisms are "very fine-grained", such that "two utterances of a sentence like 'S is interesting' in two contexts will almost always vary at least a little bit in their extensions", and since intensions are functionally determined by extensions, in their intensions too (109). Nevertheless, "we can often describe two people who utter the sentence 'A is an interesting theory' in different contexts as *having said the same thing*. They both said that A is an interesting theory" (109). Generalising, Cappelen holds that, although much of our language is context-sensitive, and the mechanisms of context-sensitivity are largely opaque to us, "we can still use disquotational reports with confidence" (110). He concludes that samesaying is possible even between speakers for whom a term differs in extension and intension. Second, Cappelen argues that since samesaying is possible where extensions and intensions differ, so is talking about the same topic, as "[s]ameness of topic goes hand in hand with samesaying" (108). Putting together these two steps, we get Cappelen's response to the topic-preservation challenge to conceptual engineering: it isn't the case that "conceptual engineering involves a change of topic because it involves a change of intension and extension", because change of intension and extension does not mean change of topic (113).

Cappelen's response is not, and is not intended as, a response to the incoherence challenge. He argues that conceptual engineering can preserve topic, but not that conceptual re-engineering is possible: that we can change the intension or extension of a concept and still have the same concept at the end of the process. His response to the topic-preservation challenge is to say that conceptual engineering doesn't entail changing the topic, so long as those who use the word 'women', for example, at the end of a conceptual engineering project on the term can samesay when they use 'women' with speakers at the beginning of the project. Cappelen says nothing about the identity conditions for concepts, such that changing a concept's extension or intension needn't mean abandoning the concept.

In fact, Cappelen puts forward a theory of conceptual engineering on which this method does not involve concepts at all. On Cappelen's theory of conceptual engineering, what is engineered is not a concept, but the non-conceptual world. Conceptual engineering is not about, for example, the concept *marriage*, the concept *person*, the concept *torture*. Instead, "conceptual engineering is about the world. It is about … marriage, persons, torture" (137). Conceptual engineers, on Cappelen's view, are engaged in "object-level change: we're changing what gender, freedom, salad, marriage, etc. *are*" (137). He calls this the "worldly description" of conceptual engineering (138). Nevertheless, a conceptual engineer who endorses a picture of conceptual engineering on which it is concepts that are engineered can still make use of Cappelen's samesaying response to the topic-preservation challenge: she can hold that, regardless of whether the concept is preserved in conceptual re-engineering, speakers who make use of the pre-engineered concept may be able to samesay with the relevant terms as speakers who use the engineered concept. For example, speakers who make use of Haslanger's concept *woman* may be able to samesay with the term 'woman' as those who make use of the pre-engineering concept, and to the extent that these different speakers can samesay with 'woman', Haslanger's conceptual reverse-engineering project preserves topic. I will return to Cappelen's proposal in §4.4.

4.3. Appeal to function

An alternative response to the topic-preservation challenge to conceptual engineering is suggested in Strawson's original formulation of the objection. He writes that "the result of attempting [explication] would be something so radically different from the original that it could no longer be said to be fulfilling the same purpose, doing the same thing" (1963: 505). This suggests that if the conceptual engineer could show that the end-product of her project *is* fulfilling the same purpose, doing the same thing as the concept with which she began, then she could show that she has not changed the topic. That is, the conceptual engineer could appeal to the preserved function of her concept of interest to argue that topic has similarly been preserved.

This is the kind of response to the challenge that Thomasson makes. She appeals to function to identify "a sense in which we remain on topic across changes in intension and extension", which is that in making these changes to a concept, we "aim to solve the same problem or pursue the same goals" as were originally under

consideration (2020: 442). As example, she discusses how the concept *marriage* has changed over the last 50 or so years. In 2022, our concept *marriage* includes same-sex couples. But in 1972, not only did *marriage* not include same-sex couples in its extension, but arguably its intension was such that this would be impossible: *marriage*, as a matter of conceptual necessity, was a union between a man and a woman. *Marriage* now has an intension and extension different to that which it had in 1972. But Thomasson argues that this change can be considered an improvement of the old concept *marriage* rather than a change of topic, because the new concept serves the same "legitimate and desired function" as the old one: "to mark a range of close relationships that we would help protect by offering a special legal and social status" (443). This function is still served, and indeed served even better, by the expanded concept *marriage*, whose extension includes same-sex couples "that are otherwise similar in character to those previously included in the extension". Because this function is served by both the earlier and the later concept *marriage*, there is an important sense in which topic has been preserved: we can think and talk about the same subject matter using the pre-amelioration concept *marriage* and the post-amelioration concept *marriage*. Thus "we can see the change as a conceptual improvement, rather than a mere changing of topic" (443).

It is not clear whether Thomasson intends this as a solution to both versions of the Strawsonian challenge, or just the topic-preservation challenge. She sometimes writes as if she intends to appeal to the function of concepts as an alternative way of individuating concepts, for example she writes that Strawson's challenge demands "a way of understanding *concept* that can preserve the sense in which people are 'talking about the same subject' over time – not just 'changing the subject'. To do this, it seems we might do better to look to function and historical continuity in individuating concepts than to rely on precise intensions and extensions" (442-3). This suggests that she intends her response to be to both versions of the Strawsonian challenge: concepts are to be individuated by their functions, so where there is preservation of function there is not a numerical change of concept; and where there is no change of concept, there is no change of topic, either – we are talking about the same thing (the same concept) before and after amelioration. Yet she never goes so far as to explicitly say that, on her picture, the concept itself has been preserved over (function-preserving) change to its intension and extension, just that topic is preserved.

Michael Prinzing explicitly argues that we should individuate concepts in terms of their functions, and for the reason that this enables us to address the topic-preservation challenge to conceptual engineering, which he calls the "Discontinuity Objection" and articulates thus:

> A common charge faced by conceptual engineers is that they do not advance their respective debates because they merely change the subject. If you change a concept, this objection claims, then you equivocate and fail to address the pertinent issue. [Conceptual engineering] is therefore philosophically uninteresting or irrelevant. (2018: 855)

Prinzing argues that what is essential to a concept is the function that it serves within some conceptual repertoire (867), thus revisions to a concept that preserve its function do not change that concept essentially. On Prinzing's view, concepts are "functional kinds, like can-openers, software programs, or bodily organs", and as such are "individuated by the functions that they serve" (867). If concepts are individuated in terms of their functions, a concept can persist through a change in its intension or extension, so long as that change preserves its function. If the concept that is the product of a conceptual engineering project is numerically the same concept with which that project began, then we are still talking about the same thing at the end of the project as we were at the beginning. As such, there has not been a change in subject. Then the Discontinuity Objection (what I have been calling the 'topic-preservation challenge') "cannot be a sweeping dismissal of [conceptual engineering], as it does not properly apply to *all* conceptual changes" (858). Changes to a concept that preserve its function "don't lead to discontinuity in subject, inquiry, or communication" (858).

Though Prinzing intends this as a response to the topic-preservation challenge, it clearly is a response to both challenges we have been considering: the topic-preservation challenge and the incoherence challenge. The incoherence challenge says that concepts have their intensions and extensions essentially, and as such any change to a concept's intension or extension destroys that concept and produces a new one in its place; thus conceptual re-engineering, whereby a concept is revised without being replaced, is incoherent: there can be no qualitative change to a concept without numerical change. Prizing's "function theory" of concepts (866), on which concepts are individuated in terms of their functions, denies that a concept's intension

or extension is essential to that concept. What is essential to a concept is its function. Change to a concept's intension or extension doesn't necessarily involve numerical change of concept, provided that the change doesn't mean a change in function.

However, the view that Prinzing endorses (and to which Thomasson seems sympathetic), that we should individuate concepts in terms of the function that they serve, faces a problem: different concepts can serve the same function, even within one conceptual repertoire. Recall Thomasson's example of *marriage*, which she holds serves the function of "mark[ing] a range of close relationships that we would help protect by offering a special legal and social status" (2020: 443). The function theory of concepts says that we should then individuate this concept in terms of this function. But in England and Wales, another concept serves the exact same function as does *marriage*: the concept of civil partnership. That is, both *marriage* and *civil partnership* serve the function of affording a special legal and social status to couples. (They don't have all the same legal consequences, for example property is not automatically transferred from one partner to the other in case of the first partner's death in a civil partnership, unlike in a marriage. But this doesn't undermine the claim that both concepts function to afford a special legal and social status to couples.) But we presumably don't want to say that *marriage* and *civil partnership* are the same concept. For one thing, their extensions are radically different. In fact, they are disjoint sets: a couple can be married in England or Wales only if they are not in a civil partnership, and *vice versa* (Civil Partnership Act 2004 and the Civil Partnership, Marriages and Deaths (Registration etc) Act 2019. Same-sex civil partnerships can be converted into marriages, but this immediately dissolves the civil partnership; see the Marriage Same Sex Couples Act 2013). This is presumably not the minor difference in extension and intension that Thomasson has in mind when talking about not relying on "precise intensions and extensions" in individuating concepts. Such a radical difference in extension and, because the extensions are necessarily disjoint, in intension too, cannot plausibly be found in one and the same concept. As such, the function theory of concepts, on which concepts are individuated in terms of their functions, is untenable.

4.4. Richard on concept individuation

Mark Richard (2019) develops an alternative picture of concept individuation to the function theory, on which concept identity can persist through (some) changes in extension or intension, thus on which change in concept does not necessitate change

of concept. On this picture, concepts are closely related to word meanings – in fact, Richard often uses 'concept' and 'meaning' interchangeably, though he does eventually distinguish concepts from meanings. The meaning of a word, for Richard, is what someone must know in order to be a competent speaker with respect to that word (cf. Higginbotham 1992). For Richard, this is a matter of meeting the expectations of other speakers. For example, to know the meaning of the word 'red', it is not sufficient that one knows that the predicate 'is red' is true of all and only red things (so meaning does not reduce to reference). Rather, one must know what are the expectations that other competent language-users will have of those who use the word (including themselves): they will expect them to know that red is a colour, that colours are perceptible, that other colours include blue and yellow, and many other facts (Richard 2019: 59).

Richard unpacks these expectations in the language of presupposition and common ground, two notions he borrows from Robert Stalnaker. To presuppose a proposition P is not to believe P, but rather "to be disposed for certain purposes to act as if one believed it" (Richard 2019: 65). The common ground in a group of language-users G is the set of all the propositions P such that each member of G presupposes P, each presupposes that each presupposes P, each presupposes that each presupposes that each presupposes P, and so on (55; cf. Stalnaker 2002, 2014). The meaning of a word in a community of language-users is, on Richard's account, a set of all the propositions P such that:

1. Users of the word presuppose P,
2. Those users expect their audiences to recognise (1),
3. (1) and (2) are common ground for users of the word,

plus any generic claims to the effect that:

4. Users of the word use it to φ,
5. Those users expect their audiences to recognise (4) (2019: 72).

Richard calls this the word's "interpretive common ground", or ICG for short (65).

The presuppositions that make up a word's ICG will change over time as, for example, new facts about the world are discovered. Thus word meaning changes over time. Returning to the example of 'red', it is plausibly part of the ICG for 'red', at least for adult users of the word, that objects appear to be certain colours when they reflect light at particular wavelengths. For this information to become part of the ICG of 'red', it had to be discovered by physicists, then disseminated throughout society through

education. This process took time, and during this time, the meaning of 'red' changed: this process led to new propositions being in the ICG of 'red', thus to a change in the meaning of 'red'.

Richard argues that, even though word-meanings are ICGs, there can be gradual change *in* an ICG without there being change *of* meaning, just as we can gradually change the planks that make up a ship without thereby changing one ship for another. There are two ways that a word's meaning can change in a population between two times t_1 and t_2. First, the way a word is used in a population at t_2, and in particular the presuppositions associated with that word, might be so different from the way the word was used at t_1 that "it is inappropriate to say that [the word's] meaning in [that population] is the same" (2019: 106). Richard calls this *change of meaning*. Second, the claims that make up the word's ICG may shift between t_1 and t_2. Richard calls this *change in meaning*. Meanings can persist through change in meaning, but not change of meaning. Change of meaning is numerical change. Change in meaning is merely qualitative change.

To test whether a given change to an ICG was change of meaning or only change in meaning, Richard appeals to the notion of *fluid conversation*:

> Meaning as we are conceiving it is what grounds linguistic competence. The natural measure of such competence is the ability to engage in fluid conversation. If this measures competence, then insofar as a change in a property or relation of a word does not by itself impede fluid communication between the word's users and their audience, that change shouldn't count as change of meaning. (122)

That is, change in an ICG that does not prevent a language-user who uses the word with its new ICG from fluidly communicating with one who uses it with the old ICG is only change in meaning, not change of meaning. These speakers are using a word with the same meaning, even though the propositions making up the word's ICG are different for each of them.

A word's ICG can change because its intension or extension changes. Richard gives the example of 'pasta', whose meaning, he argues, at one point necessarily included being made from wheat flour and water (124-5). Now, people make pasta from all kinds of flours, including chickpea flour and lentil flour. Suppose that Peter knows that pasta is now made from different kinds of flours, and so does his father.

However, Peter also knows that some people make pasta out of agar, while his father doesn't know this. It is not part of the ICG of 'pasta' for either Peter or his father that pasta is necessarily made from wheat flour and water. But it may well be part of the ICG of 'pasta' for Peter's father that pasta is necessarily made from *some kind of flour* plus water, while this isn't part of its ICG for Peter. The extension and intension of 'pasta' is thus different for Peter than for his father. But supposing that Peter and his father only ever eat bog-standard wheat flour pasta, it's unlikely that this difference in ICG, which is due to a difference in intension and extension, will make them unable to fluidly communicate using the word 'pasta'. So 'pasta' can have the same meaning for Peter and his father, on Richard's view, even though its extension and intension (and ICG) is different for each of them.

I have been talking in terms of meaning individuation, but in order to resolve the incoherence challenge, I need to talk in terms of concept individuation. As noted, Richard tends to use the words 'meaning' and 'concept' interchangeably. He himself acknowledges that he has "been free and easy, going back and forth between talk of a word's meaning and of the concept it expresses" (128). But he does distinguish the two. He introduces the notion of *uses* of lexical items, such as words:

> Assuming that there is something in the way vocabulary and constructions are realized in the mind of the individual speaker that corresponds to the way lexical entries in the dictionary are divided: the entry for 'hit', for example, is divided into two subcategories – nouns and verbs – with each subcategory constituted by sub-entries (hit, v. 1. to strike with (something held in) the hand; 2. to impact; 3. …). Call these divisions in the lexical knowledge uses. (133)

For a community able to communicate with each other, the individual language-users of words will be *coordinated* with each other in that they associate roughly the same ICGs with these uses (133, see also 63). Idealising, Richard assumes that this coordination "more or less groups the community's uses into equivalence classes", and these equivalence classes, for Richard, are concepts (134). The uses that make up a concept for a linguistic community might vary, but not by much: "there will be a rough uniformity in them – too much divergence will, at least over time, result in coordination disappearing" (134). Thus, for Richard, a concept is an equivalence class of uses of words by members of a linguistic community; the equivalence relation that binds this class is between the ICGs that users each associate with a use of a word.

Richard individuates concepts, similarly to meanings, in terms of intercommunication. There may be a difference in the extensions that two language-users associate with the concept *dog*; for example, one of them thinks that Chihuahuas are dogs and so includes them in *dog*'s extension, while the other thinks that Chihuahuas are rats and so does not include them in *dog*'s extension. But so long as the ICGs which these language-users associate with this use of 'dog'[7] are the same, then both their uses can feature in the equivalence class that makes up the concept *dog*, and so both language-users can partake in the same concept. Difference in extension is thus insufficient for difference of concept, on Richard's view.

This picture of concept individuation suggests what we might think of as an ontological version of Cappelen's samesaying response to the Strawsonian challenge. Cappelen's response is to the topic-preservation challenge, the challenge to account for topic-preservation, but not concept-preservation. As Richard individuates concepts in terms of language-users' abilities to communicate fluidly using that concept, that two speakers are samesayers in Cappelen's sense regarding (for example) 'pasta' suggests that they partake in the same concept *pasta*. This can be so even if the extension or intension of the concept is different for each of them. Then adopting this picture of concepts makes possible the following response to the incoherence challenge to conceptual engineering: it isn't the case that any change to the extension or intension of a concept is change of concept (change in concept does not necessitate change of concept), so conceptual re-engineering is not incoherent.

Such a response to the incoherence challenge is, of course, heavily theoretically laden. One must commit oneself to a substantial picture of what a concept is. On this picture, concepts are closely tied to word-meanings, thus to languages. To possess the concept *dog*, one must understand the word 'dog' in the sense of grasping the word's ICG. But this means that someone who doesn't speak English doesn't have the concept *dog*, as they don't understand 'dog'. They might have a very similar concept, in terms of intension and extension; for example, a concept *perro* or *chien* that is an equivalence class of uses of 'perro' and 'chien', respectively. But they don't have the concept *dog*. Thus Richard's account of concepts is not compatible with any picture of concepts on which an English and a Spanish speaker can have the same concept *dog* even though

[7] 'Dog', like 'hit', has different uses, akin to the different lexical entities in the dictionary: 'dog' as a noun means the animal, 'dog' as a verb means to follow someone closely. I am interested in the noun use.

only the former understands the word ‘dog’ (for example, the picture endorsed by Simion and Kelp 2020: 986-7).

Richard’s picture of concepts faces a further, and more substantial, problem. It is necessary and sufficient for partaking in the same concept, for Richard, that two language-users can fluidly communicate using that concept. Crucially, this condition can be met even if the extension or intension of the concept is slightly different for each of them: change in concept is not change of concept. But, surely, enough changes in concept would build up to a change of concept. Suppose that two language-users A and B can fluidly communicate using a concept X, even though the intension and extension of X is slightly different for each of them. For A, something is X iff it meets some conditions (1), (2) and (3), while for B, something is X iff it meets conditions (2), (3) and (4). Suppose that B can also fluidly communicate about X with a language-user C, for whom something is X iff it meets conditions (3), (4) and (5). C, in turn, can fluidly communicate with D, for whom something is X iff it meets conditions (4), (5) and (6). But A would not be able to fluidly communicate with D: there is nothing in common in A and D’s understandings of concept X. Then ‘partaking in the same concept’ is not a transitive relation. As such, it is not an equivalence relation. But this is very odd: sameness relations are equivalence relations.

A similar problem arises for Cappelen’s notion of samesaying: the relation of *saying the same thing*. Suppose that A and B can samesay using the term ‘x’, because for A, ‘x’ is true of an object iff that object meets conditions (1), (2) and (3); while for B, ‘x’ is true of an object iff it meets conditions (2), (3) and (4). B and C can samesay with ‘x’, because for C, ‘x’ is true of an object iff it meets conditions (3), (4) and (5). And C and D can samesay with ‘x’, because for D ‘x’ is true of an object iff it meets conditions (4), (5) and (6). But A and D will not be able to samesay with ‘x’: there is nothing in common in their understandings of the meaning of ‘x’, no condition that they both think is necessary for ‘x’ to be true of some object. Thus samesaying is not transitive, and so not an equivalence relation. Again, this is a problem: sameness relations are equivalence relations. We see the problem more vividly when we reframe the problem in non-technical terms: on Cappelen’s view, A and B say the same thing using ‘x’; B and C say the same thing using ‘x’; C and D say the same thing using ‘x’; but A and D do not say the same thing using ‘x’. Therefore, ‘having said the same thing’ is not transitive, so not an equivalence relation.

This problem seems to me sufficient reason to think that both Cappelen's response to the topic-preservation objection, and the response from Richard's picture of concepts to the incoherence objection, fail. Sameness relations are equivalence relations, so accounts of *saying the same thing* and *partaking in the same concept* on which these are not equivalence relations are inadequate. Then Cappelen hasn't shown that two speakers can samesay using 'x' even if 'x' differs in intension and extension for these speakers, and the response from Richard's theory of concepts hasn't shown that the same concept can persist through changes in intension and extension. Thus these responses don't solve the topic-preservation objection and the incoherence objection, respectively.

However, something can be salvaged from these responses. To solve the topic-preservation objection, we don't need it to be the case that two speakers can say exactly the same thing when one partakes in the pre-engineered concept and the other in the post-engineered concept. It is sufficient that these speakers say something *similar enough* to be able to fluidly communicate. Call two speakers A and B *similarsayers* regarding a term 'x' iff they can fluidly communicate using 'x'. For A and B to be similarsayers with 'x' doesn't require that 'x' has the same intension and extension for A and B, as demonstrated by the example of 'pasta'. So long as users of a pre-engineered concept X can similarsay with the term 'x' as users of the post-engineered concept X', the conceptual engineering project has not changed the topic in such a way that speakers cannot partake in the same debates, attempt to solve the same problems, and so on. Then conceptual engineering can preserve topic.

Similarsaying won't be a transitive relation, as A's use of 'x' can be similar enough to B's use of 'x' for A and B to similarsay with 'x', and B's use of 'x' can be similar enough to C's use of 'x' for B and C to similarsay with 'x', without A's use of 'x' being similar enough to C's use for them to similarsay with 'x'. As such, it is not an equivalence relation. But this is fine: similarsaying is not a sameness relation, so we shouldn't expect it to be an equivalence relation. So this response to the topic-preservation challenge is not subject to the objection just raised against Cappelen's samesaying response, that it articulates a sameness relation that fails to be an equivalence relation.

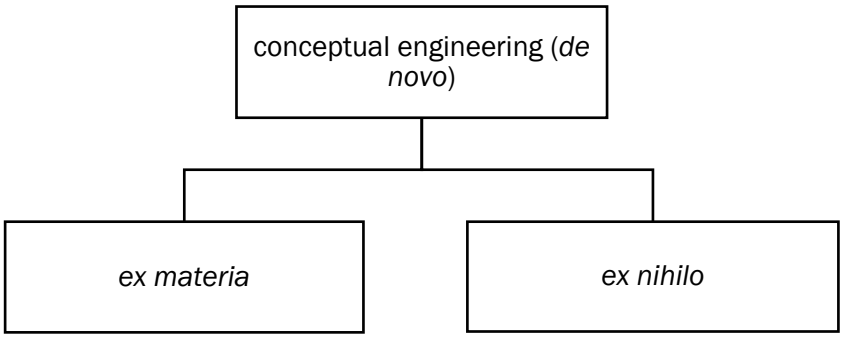
4.5. Simion and Kelp's bullet-biting response

Simion and Kelp offer a response to the topic-preservation challenge, which in turn suggests a response to the incoherence challenge. They argue that one way of responding to the objection that ameliorating a concept "seems tantamount to simply turning one's back on the old philosophical problem" is to bite the bullet and accept that this is exactly what one is doing. But this bullet-biting is entirely legitimate, as "walking away from a problem isn't always a bad thing. On the contrary, sometimes it is exactly what one needs to do in order to make progress" (2020: 991). For example, they suggest that there is a sense in which Copernicus "turn[ed] his back" on the problem that vexed Ptolemy, "to wit, explaining planetary motions within a geocentric window." But we should be thankful that he did turn his back on the problem: this was "exactly what needed to be done in order to make progress in astronomy" (991). The conceptual engineer, then, can simply accept that she is changing the subject when she revises her concept. But this is exactly what she ought to do in order to make progress in philosophy (or whatever discipline her conceptual engineering project lies within). This response does not solve the problem so much as dissolve it.

Extending this thought, the conceptual engineer could make a similarly bullet-biting response to the incoherence challenge. In changing the extension and intension of a concept, she has not improved that concept but abandoned it. But perhaps this is just what needs to be done in order to make the kind of progress she cares about. For example, perhaps in revising the concept *marriage* to include same-sex couples, we didn't improve the original concept *marriage*, but simply replaced it. But this new concept *marriage* is better suited for the legitimate purpose of marking a range of close relationships that we help protect by affording a special legal and social status. So replacing the original concept *marriage* with this new concept is exactly what we should have done. Then it may be that conceptual re-engineering is not possible. But that's just fine: we can understand all conceptual engineers as engaged in *de novo* engineering, and legitimately so.

If we accept this response, then the taxonomy from §3.3 should be revised, as all conceptual re-engineering is really *ex materia de novo* conceptual engineering. Our revised taxonomy would thus be as follows: all conceptual engineering is *de novo* conceptual engineering; *de novo* conceptual engineering can either be *ex nihilo*, if a new concept is created where previously there was no concept, or *ex materia*, where a

new concept is created where there previously was another concept. Another diagram may be helpful:



Examples of (*de novo*) conceptual engineering *ex materia* are Carnap's explication of *fish*, Haslanger's amelioration of gender concepts, and the replacement of the concept *marriage* on which same-sex couples cannot fall under its extension with the concept *marriage* that includes same-sex couples. An example of conceptual engineering *ex nihilo* is the creation of the new gender concept *genderqueer*. Just as for Simion and Kelp's response to the topic-preservation objection, this response to the incoherence objection dissolves, rather than solves, the problem. This response grants that conceptual re-engineering is incoherent, but those projects that we previously thought of as conceptual re-engineering projects can be redescribed as instances of *de novo* conceptual engineering.

The bullet-biting response to the topic-preservation challenge is plausible for some conceptual engineering projects, and less plausible for others. Simion and Kelp are right that, regarding some problems, turning our back on them is the best thing we can do to make progress in our broader projects; the case of Copernicus is one such case. But there are some problems that we don't want to turn our backs on, but nevertheless want to use conceptual engineering to solve. Revisionist theories have been offered for a huge number of concepts about which we care a great deal: *moral responsibility* (Brandt 1979, Railton 1986), *truth* (Scharp and Shapiro 2017), *personal identity* (Parfit 1984), and *free will* (Vargas 2013), to name a few. It is not at all obvious that we would be right to turn our backs on prior philosophical problems that involved *moral responsibility*, *truth*, and so on. Rather, these philosophers surely hope that, in revising these concepts, they can make progress on these very same problems. So the bullet-biting response to the topic-preservation challenge cannot serve as a blanket response to the topic-preservation challenge.

To take the bullet-biting response to the incoherence challenge, however, doesn't require holding that we have, for every conceptual engineering project, turned our back on the original problem (or inquiry, or conversation, or whatever). The bullet-biting response to the incoherence challenge simply accepts that any change to a concept's intension or extension is a numerical change of concept. This view can be combined with a view on which numerical change of concept doesn't necessarily mean changing the topic. That is, we can take the bullet-biting response to the incoherence challenge without taking the bullet-biting response to the topic-preservation challenge, at least as a maximally general response that is meant to cover all conceptual engineering projects.

4.6. The importance of (dis)solving the Strawsonian challenges for my project

For my purposes, it does not matter whether the incoherence objection to conceptual re-engineering is soluble, as the conceptual engineering project I undertake (in Chapter 5) is best thought of as an instance of explication: I argue that we ought to replace the less theoretical concept *doubt* with the more exact concept *epistemic anxiety*. Then it does not matter to me whether conceptual re-engineering is possible; I am not trying to do conceptual re-engineering. Regarding the topic-preservation objection, it might be that different responses are appropriate for different cases. Perhaps sometimes the right thing to say is that there has not been a change in topic over the course of a conceptual engineering project, because speakers who have the pre-engineering concept X1 can similarsay using 'x' with speakers who have the post-engineering concept X2. But perhaps on other occasions, the right thing to say is that the conceptual engineer has changed the topic, and she was right to do so: that was exactly what was required to advance the debate. I will return to this issue as it affects my own conceptual engineering project in Chapter 5.

5. Conclusion

In this chapter, I described two methods that are valuable for function-first philosophers, and which underpin my projects in the rest of this thesis: conceptual reverse-engineering and conceptual engineering. Conceptual reverse-engineering is a method for finding the function of a concept, and proceeds as follows: the theorist offers a plausible hypothesis about the function of the concept; she abstractly

represents a typical case in which the concept is used to serve this function; she theorises about what the concept must be like in order to serve this function in this case: what intension and extension it must have; and she then compares the concept demarcated by this intension and extension to the intuitive concept of interest. If the concepts are relevantly similar, she has a good claim to having identified the function of the concept of interest, but if they are sufficiently dissimilar, this is reason to think her original hypothesis was false. I offered a taxonomy of sub-methods of conceptual reverse-engineering in terms of whether the typical case is represented using a model or ADR, and whether this representation includes a time-axis. Conceptual engineering is a method for improving our conceptual world, by either revising, replacing or abandoning defective concepts; improving non-defective concepts; or constructing new concepts to serve particular purposes. I offered a taxonomy of sub-methods of conceptual engineering in terms of whether a method begins with a concept or not (*ex materia* vs. *ex nihilo* conceptual engineering), and whether the concept with which one begins is revised or replaced (conceptual re-engineering vs. *de novo* conceptual engineering). I introduced the Strawsonian challenge to conceptual engineering, and distinguished two versions of the challenge: the topic-preservation challenge and the incoherence challenge. I evaluated some responses to both. I concluded that the topic-preservation challenge can be successfully addressed in different ways in different cases, and that for my purposes, the incoherence challenge needn't be solved.

The discussions in this chapter have been largely schematic, and examples of different conceptual reverse-engineering and conceptual engineering projects have not been fleshed out in great detail. In the next chapter, I consider in more depth four conceptual reverse-engineering projects: Craig (1990), Hannon (2019), Kelp (2011) and Klemens Kappel's (2010) approaches to the concept *knowledge*. Having considered these four projects in depth, I will raise and respond to some problems that arise for conceptual reverse-engineers.

Chapter 3. The point of *knowledge*

1. Introduction

In this chapter, I evaluate four conceptual reverse-engineering projects on the concept *knowledge.* These projects are undertaken by Edward Craig (1990), Michael Hannon (2019), Christoph Kelp (2011) and Klemens Kappel (2010). The projects are distinguished from each other by the hypotheses about the function of *knowledge* that they test, and the types of model they construct. Craig and Hannon both test the hypothesis that *knowledge* functions to flag good informants, while Kelp and Kappel test the hypothesis that *knowledge* functions to signal when inquiry should come to an end. Craig and Kelp construct diachronic models, while Hannon and Kappel construct synchronic models. In §3, I argue that each type of model-based conceptual reverse-engineering, diachronic model-based conceptual reverse-engineering and synchronic model-based conceptual reverse-engineering, faces a problem that the other type does not. Diachronic model-based conceptual reverse-engineering can be seen as committing something like the genetic fallacy (§3.1), while synchronic model-based conceptual reverse-engineering cannot demonstrate whether our concepts are "practically necessary" for us (see Queloz 2021: 58): that is, whether our concepts are such that, given the kinds of creatures we are in the kinds of environments in which we live, we must have those concepts (§3.2). I argue that theorists interested in conceptual reverse-engineering therefore have reason to engage in both kinds of conceptual reverse-engineering project, as they do not face each other's limitations. As such, a hypothesis about a concept's function that is confirmed by both types of conceptual reverse-engineering project is more robustly confirmed than one confirmed by only one type (§3.3). In §4, I present three advantages of the inquiry-stopper picture of the function of *knowledge* over the informant flagging picture: that the inquiry-stopper picture is more robustly supported than the informant flagging picture (§4.1); that the inquiry-stopper picture explains why there should be some modal condition on knowledge (§4.2); and that the inquiry-stopper function is explanatorily prior to the informant-flagging function, in that *knowledge* having the former function would explain how it could also serve the latter, but not *vice versa* (§4.3). I end (§5) by considering some recalcitrant data for the inquiry-stopper picture: cases in which a subject who intuitively knows that P seems to be epistemically permitted to inquire into

whether P. These cases are Jessica Brown's (2008) Surgeon (§5.1) and Elise Woodard's (2021) Locked Door (§5.2). I argue that both can be explained on the inquiry-stopper picture. Brown's surgeon, I argue, is plausibly *not* permitted to inquire into whether P, epistemically speaking, but required to inquire into whether P by professional norms that bind her in her role as a surgeon. Thus this is a case in which two domains of normativity clash. The subject in Woodard's Locked Door can be understood as dropping her belief, thus her knowledge, that the door is locked when she goes back to check whether it is locked. Then she doesn't permissibly inquire into whether P while knowing that P, because she doesn't inquire into whether P while knowing that P.

2. Reverse-engineering *knowledge*

In this section, I consider Craig, Hannon, Kelp and Kappel's conceptual reverse-engineering projects concerning *knowledge*. Each of these theorists utilise models, rather than ADRs. Instead of directly but abstractly representing some actual practice that we have with *knowledge*, each theorist lists some facts about creatures like us, then from these facts constructs an imaginary situation featuring creatures about whom these facts are true, but who do not have our concept *knowledge*. This imaginary situation is the target system about which each of them theorises. The final stage of each theorist's method is to apply their findings about the target system to our actual conceptual practice with *knowledge*. These theorists' conceptual reverse-engineering projects differ along two axes: whether they use diachronic or synchronic models, and what hypothesis about the function of *knowledge* they test. Craig and Kelp construct diachronic models, while Hannon and Kappel construct synchronic models; and Craig and Hannon test the hypothesis that *knowledge* functions to flag good informants, while Kelp and Kappel test the hypothesis that *knowledge* functions to flag when inquiry should come to an end.

2.1. Craig's project

Craig's project involves constructing a diachronic model. He lists some facts about the kinds of creatures we are – that we live socially, have certain physical needs, and so on – and constructs a 'state of nature' case out of this description. This case isn't a direct representation of our conceptual practice with *knowledge*, as in the state of

nature, the concept *knowledge* does not exist. Craig hypothesises that these creatures would have to develop a concept like *knowledge* in order to meet a need they would have to flag good informants. The model is diachronic because it represents change over time: as time goes on, the creatures in the state of nature face different practical pressures, and this causes their concept to change to look more like the intuitive concept *knowledge*. Craig calls his method that of "practical explication" (1990: 8), intentionally invoking Carnap's method of explication. But as I argued in the previous chapter, Craig's method is a kind of conceptual reverse-engineering, while Carnapian explication is a kind of conceptual engineering. To avoid confusion, I will follow Georgi Gardiner (2015: 39) and Matthieu Queloz (2021: 52) in calling Craig's method 'hypothetical genealogy'.

Craig's hypothetical genealogy of *knowledge* proceeds as follows. He begins by offering a "plausible hypothesis" about the function of *knowledge*: what the concept "does for us, what its role in our life might be" (1990: 2). His hypothesis is that *knowledge* functions to flag good (in the attributive sense of 'good'; see Geach 1956) informants: people who are such that, if they tell us that P, we should take it from them that P. That *knowledge* serves this purpose for us, Craig hypothesises, explains why we have this concept. To test this hypothesis, Craig constructs "an ordinary situation" (2) from a description of the kinds of creatures we are: that we have certain basic needs for food, water and shelter; that we need true beliefs in order to meet these needs; that we can get true beliefs through our "on-board" sources, such as perception and reasoning, or by asking other people (11). His ordinary situation consists of creatures who are like us in these respects, but unlike us in that they do not have the concept *knowledge*. Following Kelp (2011: 62), let's call them our 'ancestors'. He then asks what needs our ancestors would have that would be going unmet, such that they would be led to develop this concept, or one very close to it.

As noted, our ancestors, like us, need true beliefs in order to satisfy their basic needs. True belief thus has survival value for them. They can build up a "primary stock" of true beliefs using their "on-board" sources of perception and reasoning (11). But often, some ancestor won't easily be able to get a true belief using her on-board sources: she won't be in the right place to see whether P, for example, or she won't have reasoning skills as sharp as other ancestors. It would therefore be hugely advantageous for our ancestors if they could "tap the primary stocks of their fellows" (11); that is, if they could make use of each other as informants. On any given issue,

some informants will be better placed than others to supply their fellows with true beliefs: “Fred, who is up a tree, is more likely to tell me the truth as to the whereabouts of the tiger than Mabel, who is in the cave” (11). As such, our ancestors have a need to evaluate informants, and in particular, to flag that some informants are good, and should hence be used. They would thus develop a concept that functions to “flag good informants” (11). What would this concept look like?

Craig answers this by considering what properties an ancestor who wants to find out whether P would want a potential informant to have. First, she would seek someone who is available to her, here and now. Second, she will want him to be willing to tell her whether P – as Craig puts it, the “channels of communication” between inquirer and informant “should be open” (85). Third, she will want what he tells her to be true, and an informant will not “in general tell [her] the truth unless he (the informant) holds a true belief”. Thus she wants an informant who satisfies the following condition: “Either *p* and he believes that *p*, or not-*p* and he believes that not-*p*” (12). That is, she wants an informant who has a true belief whether P. Further, she will want someone who is detectable as having, or at least as being very likely to have, a true belief whether P; for otherwise she wouldn’t think to ask this person whether P. For example, Fred up the tree is detectable as very likely to have a true belief on the whereabouts of the tiger: our ancestor can see that he is in the right place to see where the tiger is. Craig doesn’t think that there is a property that is both detectable and which strongly correlates with having true belief, thus he leaves this condition schematic: the informant must satisfy the condition of having “any detectable property which has been found to correlate closely with holding a true belief as to whether *p*” (25). He adds that this correlation must be lawlike, as an accidental correlation between having the property and holding a true belief won’t support inference to new cases, and “inference to the new, as yet untested case is precisely what the inquirer needs the correlation for” (25). A natural way of putting this is that “the correlation between possessing [the property] and being right about *p* must be reliable” (25). Just how reliable the correlation must be depends on the inquirer’s concerns. For matters of grave importance, the inquirer might want possession of the property to guarantee, or near enough guarantee, holding a true belief whether P; where the inquirer’s concerns are less pressing, possession of the property’s making true belief more likely than not will be enough (86). Thus we get the following picture of what our ancestors would seek in an informant:

1. He must be accessible to the inquirer here and now,
2. Channels of communication are open between inquirer and informant,
3. He must have a true belief about whether P,
4. He must be as likely to be right about whether P as the inquirer's concerns require,
5. He must be detectable to the inquirer as satisfying (4).

Note that condition (4) is, in an important sense, derivative of condition (5): it is because the informant should be detectable to the inquirer as sufficiently likely to be right given her purposes that he should also *be* this likely to be right. The verb 'detect' is factive: one can only detect that P if P. Thus, an informant is detectable as being as likely to be right about whether P as the inquirer's concerns require only if he is as likely to be right about whether P as the inquirer's concerns require.

Our ancestors' need for a concept to flag good informants is thus a need for a concept that applies to all and only those informants who satisfy (1)-(5). Following Martin Kusch (2009: 65), call this concept '*protoknowledge*'. *Protoknowledge* resembles our intuitive concept *knowledge* in some respects. In particular, it is part of our ordinary thinking that *knowledge* requires true belief, so condition (3) of *protoknowledge* will be present in *knowledge*, too. However, it is not part of our ordinary thinking that *knowledge* requires being accessible to a given inquirer here and now, nor being willing to tell that inquirer whether P. We think that someone in Melbourne, Australia may very well know what the weather is like in Melbourne today, even though she is not accessible to me, now, in Edinburgh; and that a student can know the answer to an exam question without being willing to tell her classmate. Thus (1) and (2) are not part of our intuitive concept *knowledge*. Neither are (4) and (5). We don't think that whether S knows whether P is as closely tied to a given inquirer's concerns as (4) has it. Nor do we think that knowing requires being detectable to others as being as likely to be right as their concerns require, or indeed being detectable to others as meeting any other condition. Craig gives an example of "the secretly studious milkman, who actually knows the answer to the abstruse question that is bothering you", but "nothing about him gives the slightest hint that it would be anything but stupid to ask him" (82): the milkman fails to satisfy (5), yet our intuitive concept *knowledge* applies to him. As such, a gap opens between our intuitive concept *knowledge* and the concept *protoknowledge* that arises in the state of nature.

To bridge this gap, Craig anticipates further practical pressures that would arise for our ancestors, and models how *protoknowledge* would change in response to these pressures. He calls this process "objectivisation" (82). One such pressure comes from the need to recommend informants to each other. Often, an ancestor will be less able than others to recognise that some informant would be a good informant for her. As such, our ancestors have a need to recommend informants to each other. For example, if our ancestor wants to find out the result of a football match, "there may be nothing about Fred to suggest to [her] that he would be a good person to ask. But Fred's friends are aware that he was at the ground when play finished; so their advice [to ask Fred] will help [her]" (88). In seeking a recommendation for an informant, an inquirer isn't asking to be recommended an informant who, without this recommendation, would have struck her as a good informant. Quite the opposite: the inquirer wants to be recommended someone who, without this recommendation, she would not have recognised as a good informant. This motivates a weakening of the detectability requirement (5): though an inquirer may still "hope for … something that he could recognise" in a potential informant that correlates well with true belief, satisfaction of this condition won't be "embodied in the public concept that now develops" (90).

Neither will condition (1) be embodied in this public concept. For whether someone is a good informant, objectively speaking, won't be tied to whether they are available to some particular inquirer here and now. Craig holds that objectivisation will significantly weaken condition (2): the objectively good informant needn't be willing to tell the inquirer whether P, but she should be willing to tell some person or other whether P (92-3). Finally, objectivisation will change condition (4) to eliminate reference to the inquirer's particular concerns. For there will be pressure on users of *protoknowledge* to collect information while it is available to store for future use, without being aware of when, why, by whom the information will be used in the future. This will turn (4) into a strong reliability condition, the satisfaction of which requires "a very high degree of reliability", for the informant must be "acceptable even to a very demanding inquirer" (91). We thus end up with the following concept at the end of the process of objectivisation, which we can call '*objectivised protoknowledge*'. *Objectivised protoknowledge* applies to some informant if:

1. He has a true belief whether P,
2. He will tell someone or other whether P,
3. He is very likely to be right about whether P,

4. He is detectable to someone or other as satisfying (3).

Again, as for *protoknowledge* pre-objectivisation, the 'likely to be right' condition is derivative of the detectability condition. It is because the objectively good informant whether P must be detectable to someone or other as very likely to be right that he also must be very likely to be right about whether P – 'detect' is factive.

The final stage of Craig's method is to compare this concept to our intuitive concept *knowledge*. As noted, it is part of our ordinary thinking about *knowledge* that knowledge requires true belief, thus (1) is shared by both *objectivised protoknowledge* and *knowledge*. Craig holds that condition (3), or something very like it, is also part of our intuitive concept: this explains the popularity of reliabilism in epistemology (1990: 31). However, our intuitive concept *knowledge* does not seem to include conditions (2) or (4): we don't think that it is required for S to know that P that S will tell anyone whether P, nor that she is detectable as being very likely to be right whether P. Thus there is not a perfect match between *objectivised protoknowledge* and *knowledge*. But recall that Craig is not attempting to articulate necessary and sufficient conditions for satisfying *knowledge*. Rather, he is attempting to make plausible that our concept *knowledge* is as it is because it serves our need to flag good informants. That there are cases where a knower would not be an objectively good informant, or *vice versa*, doesn't undermine Craig's hypothesis, so long as these are sufficiently "exotic cases" (16). In the next section, I will consider some cases that make trouble for Craig's thesis, and which don't seem to be particularly exotic.

2.2. Recalcitrant data for Craig's picture

Christoph Kelp (2011) offers two cases in which our intuitions about whether a subject knows come apart from whether they satisfy the concept that is the output of Craig's project. Kelp does not put forward these cases as straightforward counterexamples to Craig's thesis. Craig's hypothetical genealogy, and indeed the broad method of conceptual reverse-engineering of which Craig's hypothetical genealogy is a sub-method, does not aim to output necessary and sufficient conditions for a concept, thus is not vulnerable to counterexample in the same way that conceptual analysis is (see §2.1 of Chapter 1). Rather, Kelp takes these cases to constitute "recalcitrant data" for Craig (2011: 58): they are cases that are "not easily explicable by Craig's account" (61). These cases undermine Craig's hypothesis insofar as an alternative hypothesis about the function of *knowledge*, when plugged into a conceptual reverse-engineering

project, would generate a concept that resembles our intuitive concept *knowledge* at least as well as Craig's, and would more easily be able to account for these cases. Kelp takes himself to offer such a hypothesis; I will discuss this in the next section.

Before introducing the cases, it is important to flag that what makes these cases recalcitrant is *not* that they are cases in which *knowledge* does not serve the function of flagging good informants. An item X can have a function F even if we sometimes use X for other purposes, and sometimes have other items that serve F. For example, the function of hammers is to drive nails into things. That this is the function of hammers is not undermined by the fact that we can use hammers for other purposes, such as breaking ice; nor that, in a pinch, one might have to use some other item to drive nails into things, such as a rolling pin. Rather, these cases constitute recalcitrant data for Craig because they are each a case in which only one of the two concepts *objectivised protoknowledge* and *knowledge* apply. This undermines Craig's original hypothesis in showing that the concept that is the output of his conceptual reverse-engineering project does not closely resemble our intuitive concept *knowledge* in certain ways.

Kelp outlines the concept that emerges in Craig's state of nature slightly differently to me. On Kelp's characterisation, Craig's (pre-objectivisation) *protoknowledge* applies when an informant meets the following conditions:

PK-1 The informant tells one the truth on the question.

PK-2 The informant is as likely to be right about P as one's concerns require.

PK-3 The informant is detectable by one as likely (enough) to be right about *P*.

PK-4 The channels of communication between oneself and the informant are open.

PK-5 The informant is accessible to one here and now. (Kelp 2011: 55)

Kelp's PK-5 is condition (1) in my outline of Craig's *protoknowledge*, PK-4 is my (2), PK-2 is my (4) and PK-3 is my (5). Kelp outlines the objectivised version of the concept, *objectivised protoknowledge*, as applying to an informant when:

OPK-1 The informant tells one the truth on the question.

OPK-2 The informant is highly likely to be right about P. (2011: 58)

Kelp doesn't add objectivised forms of PK-3 to PK-5 to his account of Craig's concept *protoknowledge*, but he takes it that weaker versions of these conditions survive objectivisation (57-8). Note that belief does not feature in Kelp's reconstruction of Craig's *objectivised protoknowledge*. Kelp holds that OPK-1 "approximates the true belief condition on the familiar concept of knowledge", as it "remains the case that,

typically, an informant won't tell one the truth unless he also has a corresponding true belief" (58). But he doesn't take it to be a "conceptual truth about good informants" that they have true beliefs (60), thus OPK-1 doesn't make reference to belief and OPK-2 should be interpreted as saying that the informant is highly likely "to give the right answer", rather than "to have a true belief on the issue" (60).

Kelp presents two cases that constitute recalcitrant data for Craig. The first is one in which a subject intuitively counts as a knower, but is not such that Craig's *objectivised protoknowledge* applies to him; the second is one in which a subject intuitively doesn't count as a knower, but Craig's *objectivised protoknowledge* applies to him. Here is the first:

> *Seal of Confession*. Don Camillo is the priest at the local parish. The members of the parish, who are all devout believers, regularly come to Don Camillo to confess their sins. As an ordained priest, Don Camillo is bound by the seal of confession. That is to say, he must not divulge information about his confessors' sins in any way or for any reason and cannot be forced to break this obligation even by the authorities. (2011: 59)

Intuitively, Don Camillo comes to know all kinds of things about what his parishioners have been up to through their confessions. But he isn't an objectivised protoknower, because he doesn't satisfy even a weak descendent of the 'channels of communication' condition PK-4: because Don Camillo is under the obligation to keep his knowledge to himself, and he takes that obligation very seriously, "[h]is channels of communication simply would not be open to anyone at any time" (59).

As discussed, Craig's method generates accounts of concepts that are relatively invulnerable to counterexample, when compared to traditional (reductive) conceptual analysis. That there is some case in which our intuitions about whether the concept applies diverge from the verdict issued by the "explicit" concept (Craig 1990: 1) that is the output of the method needn't worry Craig if the case is sufficiently "freakish" (14). But Kelp argues that Seal of Confession is not freakish; rather, the case illustrates a very general phenomenon, "that of knowledge under the seal of confession" (2011: 59). Once we consider "how common cases of knowledge under the seal of confession are, not to mention cases of professional secrecy in general, it becomes clear that … Seal of Confession [is] not at all freakish" (59-60).

But what makes a case 'freakish' or otherwise is not how 'common' such cases are, in the sense of there being numerically many of them. What matters is how typical the case is, where the relevant sense of 'typicality' (recall from Chapter 2, §2.1), is that of meaning 'representative of the relevant type'. Is Seal of Confession typical of how *knowledge* is used, in the relevant sense? One reason to think that this case is not typical is that Don Camillo is bound by professional, religious and/or moral norms that affect what he is able to do with his knowledge. In particular, he is bound by a professional, religious, and perhaps moral norm not to share what he knows. Given that, on Craig's account, the function of *knowledge* is to flag good informants, and Don Camillo is prevented from being a good informant because of professional, religious and/or moral norms, then Craig would think that this case cannot be representative of our broader practices with *knowledge*.

The general phenomenon of secret-keeping might constitute more worrying data for Craig's thesis. Jesús Navarro (forthcoming) offers an account of secrets as interruptions in the social flow of knowledge: some subject knows that P, but deliberately withholds this information from another. Navarro has it that S can only keep P a secret if S knows that P. That is, secret-keepers are knowers. Secret-keepers who are unwilling to share their information with anyone make trouble for Craig's account. Consider:

> *Secret Crush*. Since she first saw him on television, Elsie has had a huge crush on Gordon Ramsey. She is very embarrassed about this and consequently keeps it secret from all her friends. In fact, she would keep it secret from anyone: there is no one she would be willing to tell about her crush on Gordon Ramsey.

Intuitively, Elsie knows that she has a crush on Gordon Ramsey. But as she is unwilling to share this information with anyone, she doesn't satisfy even a weak descendent of PK-4. Unlike in Kelp's Seal of Confession, there are no professional, religious or moral norms in play that undermine this case's claim to typicality. Thus we have a case of a knower who isn't an objectivised protoknower that does not seem 'freakish' in the way that Seal of Confession can be seen to be.

Kelp's second case is one in which a subject to whom the concept *objectivised protoknowledge* applies intuitively doesn't count as a knower:

> *Secret Sect*. Dick is a member of a secret sect and for that reason shares the sect's belief that our planet is gradually warming. However, this belief is not held on the basis of scientific findings but is instead grounded in the sect's belief (also shared by Dick) that global warming is the result of God's decision to punish humanity for the fornicatory practices that in recent times have become so outrageously widespread among his once beloved sheep. Since the sect is secret, Dick is not allowed to assert its beliefs. For that reason he has adopted a policy of asserting on the relevant issues in accordance with what the experts in the field have to say. Fortunately, Dick is a government spokesman on environmental issues and thus particularly well acquainted with expert views on global warming. (60)

Dick counts as an objectivised protoknower, Kelp holds, because he will tell one the truth whether the planet is slowly warming, satisfying OPK-1, and given his office and policy, he is highly likely to tell one the truth on this question, satisfying OPK-2. Further, given his office, he is detectable as satisfying OPK-2, his channels of communication are open to many, and he is accessible by a wide range of people. Thus he satisfies objectivised versions of PK-3 to PK-5. But intuitively, Dick doesn't know that the planet is slowly warming, as his belief is "highly irrational," and "held for reasons that only the raving mad would conceivably take to support it" (62). So Dick is an objectivised protoknower without being a knower.

Note that, on my characterisation of Craig's *objectivised protoknowledge*, Secret Sect does not obviously constitute recalcitrant data for Craig. My characterisation of Craig's *objectivised protoknowledge* includes a true belief condition (1) in place of Kelp's OPK-1; as such, the condition that an objectively good informant must be 'very likely to be right' is naturally understood as meaning 'very likely to have a true belief'. Given the way that Dick formed his belief whether the planet is slowly warming – namely, believing on the basis of the teachings of a cult – it's not the case that he is very likely to hold a true belief on this question. Forming beliefs about global warming on the basis of the teachings of a cult is not a reliable method of belief formation. So though he satisfies OPK-2, on Kelp's reconstruction of Craig's *objectivised protoknowledge*, he doesn't satisfy condition (3) on mine.

Kelp notes that Craig might be able to explain away Secret Sect as a freakish case. For Dick is an unreliable believer, given the way he formed his belief – namely, on the basis of the teachings of a cult – and "cases in which unreliable believers are

reliable asserters are bound to be atypical" (61). As such, he accepts that Secret Sect doesn't "seal the case" against Craig (61). In fact, Kelp holds that even Seal of Confession, "which cannot obviously be dealt with in the same way" – namely, by insisting on reconstructing Craig's *objectivised protoknowledge* with appeal to belief – "does not refute [Craig's hypothesis], either by itself or in conjunction with Secret Sect" (61). Craig's method of hypothetical genealogy permits some mismatch between the intuitive concept and the concept that is the output of an application of the method. This mismatch constitutes recalcitrant data for the original hypothesis about the concept's function, but doesn't disconfirm it, unless an alternative hypothesis generates a concept that resembles the intuitive concept at least as closely, while easily accommodating the recalcitrant data. Unfortunately for Craig, Kelp advances an alternative hypothesis about the function of *knowledge* which he takes to do just that.

2.3. Kelp's project

Kelp makes use of Craig's method of hypothetical genealogy to test an alternative hypothesis about the function of *knowledge*. Kelp acknowledges that our ancestors will need concepts to evaluate informants, but points out that this is not their only conceptual need: "they also need concepts to evaluate various inquiries agents undertake" (62). The alternative hypothesis that he tests is that *knowledge* meets this need: the function of *knowledge* is "to flag when agents may appropriately terminate inquiry into a given question" (62).

Starting in the same state of nature situation as Craig, Kelp imagines that one of our ancestors is interested in the question whether P and sets out to inquire. What properties would this ancestor want himself to have upon terminating inquiry? Kelp suggests the following:

> **PK-A** He has formed a belief on whether P,
>
> **PK-B** His belief whether P is true,
>
> **PK-C** His belief whether P stems from a source that is as trustworthy on the question whether P as his concerns require. (62)

As such, our ancestors would develop a concept that applies when these conditions are met; again, following Kusch (2009), call this concept '*protoknowledge*'.

This concept *protoknowledge* approximates our intuitive concept *knowledge* at least as closely as does Craig's concept *protoknowledge*. PK-A corresponds to the belief condition on *knowledge*, PK-B to the truth condition, and PK-C is "recognisable

as an ancestor of the familiar reliability requirement" on *knowledge* (62). But however the reliability condition on *knowledge* is substantiated, it won't be relative to a particular inquirer's concerns in the same way that PK-C is. As Kelp notes, "given a suitable cost-benefit balance of being right, a true belief acquired from a barely trustworthy source can qualify as protoknowledge even though it could not qualify as knowledge" (62). Thus there is a gap between Kelp's *protoknowledge* and our intuitive concept *knowledge*. But even at this stage of Kelp's model, the gap is not as great as that between Craig's *protoknowledge* and the intuitive concept *knowledge*, as Kelp's *protoknowledge* doesn't have analogues of Craig's detectability, accessibility, or 'channels of communication' requirements. After all, "why should it matter to whether one has adequately terminated inquiry that one is detectable to be right on the issue, that one is willing to share one's results, and that one is accessible to others" (63)? As such, there is less work for Kelp to do in explaining how *protoknowledge* would change to look more like our intuitive concept *knowledge* than there was for Craig.

Kelp too appeals to objectivisation to explain why *protoknowledge* would morph into *knowledge* in his diachronic model. Kelp notes that it would be advantageous for our ancestors to store the results of their inquiries for their own future use, but at the point of inquiry they won't always be able to predict to what use they will put this information in the future. As such, "pressure towards an objective conception of adequately terminated inquiry arises, one that abstracts away from the agents' concerns at the time of inquiry", as their inquiries will be adequately terminated only when they would be satisfactory relative to "the concerns of their future selves, which may be very different and can also be opaque to them at the time of inquiry". (63). It would also be advantageous for our ancestors to inquire on each other's behalves, and to store the results of their inquiries in shared databases, to be drawn on subsequently by various members of the community. This puts further pressure towards objectivisation on *protoknowledge*, as inquirers must now terminate inquiry only when their results would be acceptable given the concerns of "other agents and groups of agents, present and future, which may be very different than the ones of the inquiring agent at the time of inquiry" (63). Objectivisation will therefore strengthen PK-3, such that only sources trustworthy enough to meet anyone's concerns will satisfy the condition.

Kelp's objectivisation story thus generates the following concept *objectivised protoknowledge*. An objectivised protoknower whether P will be such that:

OPK-A He has formed a belief on whether P,

OPK-B His belief is true, and

OPK-C His belief stems from a highly reliable source. (64)

This concept resembles our intuitive concept *knowledge* at least as well as Craig's *objectivised protoknowledge*. Further, as Kelp's concept doesn't include analogues of the objectivised versions of PK-3 to PK-5, neither Seal of Confession nor Secret Crush constitute recalcitrant data for Kelp: an objectivised protoknower whether P needn't be willing to share her information with anyone. And where Kelp argues Craig's *objectivised protoknowledge* concept will include "a strong reliability condition on agents as truth-tellers", Kelp's concept includes "a strong reliability requirement on agents' belief sources" (64). As such, Kelp's concept doesn't count Dick in Secret Sect as an objectivised protoknower, thus this case is also not a recalcitrant datum for Kelp.

2.4. Hannon's project

Hannon's sub-method of conceptual reverse-engineering involves three steps. It begins, like Craig's method, with a plausible hypothesis about the function of one's concept of interest. In order to be plausible, this hypothesis "must be compatible with certain facts about human life, such as facts about our physical environment, our social organization, our cognitive capacities, and the basic aims and interests humans typically have" (2019: 13). These facts will give rise to a "certain conceptual need", to be met by a concept that serves the hypothesised function (13-4). The second step is to "determine what a concept having this role would be like (i.e., what conditions would govern its application" (14). The third step is to "examine the extent to which the concept we have constructed matches our everyday notion" (14). That is, we compare the concept described in the second stage to the intuitive concept of interest. Hannon applies his method to *knowledge*, and his hypothesis about the function of *knowledge* is the same as Craig's: Hannon hypothesises that "the point of the concept of knowledge is to identify reliable informants" (35).

This looks very much like Craig's project. But Hannon's method differs to Craig's in that his representation of the facts about human life that generate the conceptual need to be met by the concept of interest does not include a time-axis. That is, it involves a synchronic, rather than a diachronic, model. Hannon writes that in order to "explain our contemporary conceptual equipment … we only need to make claims about actual facts as they are now" (53). It is true about us, actually and presently, that

"we need true beliefs about our environment to successfully guide our actions; reliable sources of information will lead to true beliefs; asking a good informant will often be the easiest way to acquire a true belief; on any issue some people will be more likely than others to provide a true belief" (53). We can model these facts by imagining a community of agents about whom these are also facts, but who do not have the concept *knowledge*. In order for these agents to get enough true beliefs, and not too many false beliefs, by using others as informants, they "must mark out those people on whom [they] should rely on from those [they] should not"; that is, they have a need to "distinguish reliable informants from unreliable informants" (36). But they recognise that they "don't always need an informant *here and now*, although [they] might at some point in the future. It is therefore in [their] shared interest to store useful information when it is available, since [they] do not always know when, why or under what circumstances it might be needed" (40). Thus they have a need to "assess the adequacy of informants for people and purposes beyond [their] own immediate concerns" – that is, they have a "need to identify good informants *in general*" (40). (Note that this doesn't mean that they have a need to identify people who are good informants about general matters – i.e., good informants about a wide variety of topics. Rather, it is a need to identify people who are good informant *for general purposes* – i.e., good informants relative to a wide variety of purposes to which the epistemic community might put one's information. One can be a good informant in this sense on one specific topic, or on a variety of topics.) The concept *knowledge*, Hannon holds, "derives from [this] need" (40).

What will *knowledge* look like, on Hannon's picture? Hannon doesn't offer his own characterisation of a 'good informant in general', but he does endorse that which arises from Craig's objectivisation story. That is, for Hannon, a 'good informant in general' with respect to P, is typically someone who satisfies the following two conditions:

1. She will tell the inquirer the truth whether P. (2019: 38)
2. She is highly likely to be right about P. (44)

   As well as "relax[ed]" versions of the following three conditions:

3. She is detectable by the inquirer (via some property X) as likely to be right about P.
4. She is accessible to the inquirer here and now.
5. Channels of communication between her and the inquirer are open. (38)

Though Hannon does not specify what the relaxed version of these conditions will be, we can assume that they will be something like:

3*. She is detectable to someone or other as likely to be right about P.

4*. She is accessible to someone or other.

5*. She is willing to share her information with someone or other.

The concept that is the output of Hannon's method thus resembles the intuitive concept *knowledge* as closely as does Craig's *objectivised protoknowledge*, because it is, for all intents and purposes, the same concept.

Despite insisting that his method is "not a genealogy and makes no reference to a fictional state of nature" (2019: 2), Hannon outlines Craig's hypothetical genealogy of *knowledge* in great detail (37-42). Indeed, Hannon's synchronic model is, essentially, the post-objectivisation stage of Craig's diachronic model (40-41). But without the pre-objectivisation stage of Craig's model doing the "loadbearing" work (53, cf. Craig 2007: 193) of showing how a concept like *knowledge* would inevitably arise among creatures like us in even a very primitive situation, that *knowledge* can meet the need to flag reliable informants for the agents in Hannon's model supports only the contingent claim that *knowledge* could function to meet this need for us.

This is because Hannon must build into his model much more detail than Craig must build into his model's original stage. In particular, Craig only need build into his state of nature that the inhabiting creatures need true beliefs about their environment to satisfy their basic needs for food, water and shelter. We cannot imagine being the kinds of creatures we are – human beings in our actual environments – without having this need. Hence Craig tells a story that makes plausible that any creatures sufficiently like us would have a need to flag reliable informants, which can be met by a concept that, given other practical pressures anticipatable from within his model, would change to look very much like the concept *knowledge* that we recognise as ours. But Hannon's model consists of creatures more complex than this. These creatures have already developed practices of sharing and pooling information. We can imagine creatures like us who do not partake in such practices. Indeed, some such creatures exist: hermits, for example. Then Hannon has not shown that the need for an informant-flagging concept is a very basic need for creatures like us. As such, he does not show that *knowledge*, or a concept very like it, would inevitably arise in connection with this need: he has not shown that the need itself is inevitable. He has failed to demonstrate the "practical necessity" (cf. Fricker 2019: 245; Queloz 2021: 33) that *knowledge* has for

us.[8] Inability to demonstrate practical necessity is a general problem for conceptual reverse-engineering projects that utilise synchronic models. I will return to this point in §3.2.

2.5. Kappel's project

Klemens Kappel also reverse-engineers the concept *knowledge* using a synchronic model. His method aims to answer "What is the point of X?" questions (2010: 71): it is a method for explaining why we have a given item, for example a concept, X in terms of what X does for us. Kappel, like Craig, calls his method "practical explication", though he doesn't think of himself as utilising the same method as Craig (72). For the same reason that I don't want to use this name for Craig's method, I don't want to use it for Kappel's: Kappel's method, when applied to concepts, is a kind of conceptual reverse-engineering, and explication is a kind of conceptual engineering. As it is a method for explanatory projects, I will call it 'practical explanation'.

Kappel's practical explanation is explicitly contrastive. It is not intended to explain why we have X rather than some other item Y that could fulfil the same function. For example, we can practically explain cars: we can explain why we have cars in terms of what they do for us. We need to move efficiently between different geographical locations. Our "natural ability to walk and run" is insufficient for meeting this need (72). Thus we have a need for fairly quick and efficient transport. Cars are our means of meeting that need. Thus, "we can say that fulfilling a certain need for transport is the point of cars" (71). That cars meet our need for transport explains why we have them. But it only explains why we have cars rather than "no means of transport apart from our natural ability to walk and run. It is not meant to explain why we have cars rather than some other equally efficient system of transport", such as an elaborate public transport system (72).

Kappel thus explicitly understands his method as generating only contingent explanations. The schematic form of practical explanation is as follows, for some explanandum E:

---

[8] 'Practical necessity', as used here, does not mean the same thing as it does for Bernard Williams. Williams uses the term to pick out the 'must' at work in contexts of everyday deliberation, whereby "[s]omeone deliberating in an everyday situation may conclude that a certain action is one that he must, or has to, do" (1982: 145).

1. Given a set of facts F, and a set of aims or interests I, we have a certain need N.
2. E is what actually fulfils N. (72)

The kinds of facts in F are typically facts about our environments, our biological or psychological constitution, and our social organisation. The aims and interests in I can be general or specific. General aims and interests are those that any creatures sufficiently like us would have; for example, aims for food, water, shelter. Specific aims and interests are those that we know, empirically, that we have; for example, aims to get married, own property, become celebrities. Thus both F and I can contain propositions that are both "empirical" and "contingent" (73). In that case, the need N that arises from the combination of F and I would likewise be contingent: that we have N is contingent on the facts F obtaining and our having aims and interests I, which we have only contingently. Finally, even given that we have N, E is posited as that which actually meets N, i.e. which meets N in the actual world. It could be the case that in other worlds in which our counterparts have N due to the combination of F and I, some other item, E*, is what meets N for them. Then practical explanation does not generate explanations for why we have some item X that hold in all worlds in which some set of basic practical facts about us – that our environments are as they are; that we have needs for food, water and shelter; and so on – are true. The method does not aim to show the 'practical necessity' of our having some item, for example some concept.

Kappel applies his method to the question 'What is the point of *knowledge*?' Relative to a set of facts F about our cognitive capacities and physical environments, and a set of practical aims and interests I, Kappel argues that we have a need for an "enquiry-stopper": a concept that flags when inquiry should come to an end (74). This need, Kappel holds, arises from three "trivial observations" about us, relative to F and I. The first is that "truth matters. For a range of important propositions we generally benefit from treating them as true just in case they are true" (74). Second, "without some form of enquiry, we cannot have truth. But enquiry is always costly" (75). Kappel is here talking about Q-inquiry (see Chapter 1, §2.4): inquiry into questions. He understands Q-inquiry as a process of ruling out various possible answers to Q. For example, inquiry into whether P will be a process of ruling out various possibilities in which P and various possibilities in which not-P; if the inquiry delivers the result that P, this will be because it has ruled out all but one possibility, and this is a P-possibility; if it delivers the result that not-P, this will be because it has ruled out all but one

possibility, and this is a not-P-possibility. Q-inquiry, so understood, is costly: it "generally requires resources in the form of time, energy and attention". Inquiry can thus, from the individual's point of view, "be considered a kind of risky investment" (75): the benefits of a given inquiry may, but may not, be worth the costs of embarking on it. The third observation is that inquiry "has no natural stopping point. Imagination may always reveal further conceivable but as yet uneliminated error-possibilities" (75).

These observations, taken together, demonstrate a need for "a way of expressing that enquiry has now taken one far enough, and that one shouldn't worry about remaining as yet uneliminated error-possibilities" (76). We need some way to "command a switch of our attention away from further uneliminated non-*p* possibilities, a way to urge that we simply take the truth of *p* for granted in our practical deliberation, as well as in our enquiry into other questions" (76). In talking about 'our' practical deliberation and 'our' inquiry into other questions, Kappel intends that some subject S having done enough in her inquiry to deliver the result that P (say) means that not only S but other subjects may take P for granted. As such, what we have a need for is a predicate that permits us to express the following kind of judgement:

> P, and S1 is in a sufficiently good epistemic position such that S1-S_n, given right circumstances of transmission, ought to take the truth of P for granted in their practical and theoretical deliberation. (79)

Kappel calls the predicate that will allow us to express this judgement the "K-predicate" (79), the state of a subject to which the K-predicate applies a "K-state", and the concept expressed by the K-predicate the "concept of a K-state" (80).

When will a subject be in the K-state? Kappel suggests that K-states will be factive, i.e., one can only be in the K-state with respect to P if it is true that P; that the K-state will involve belief; and that K-states "require some degree of justification or warrant, however understood" (80). Then the concept of a K-state will be similar to our concept *knowledge*, as it is part of our ordinary thinking about *knowledge* that S knows that P only if P is true, S believes that P, and S is justified or warranted in believing P (however this final condition is fleshed out). Kappel thinks that *knowledge* is what actually meets our need for a K-predicate, though he notes that he "ha[sn't] argued this further point explicitly" (80) – and indeed, he does not go on to do so. But at this point we are able to make the argument explicit. We have a need for a concept of a K-state. Our intuitive concept *knowledge* is relevantly similar to the concept of a K-state.

Thus our concept *knowledge* could meet our need for a concept of a K-state. Therefore: our concept *knowledge* does meet our need for a concept of a K-state. This is not a deductive argument. It is best thought of as an abductive argument: that our concept *knowledge* meets our need for a concept of a K-state is the best explanation for why we have it.

3. Synchronic vs. diachronic modelling

In this section, I argue that each type of model-based conceptual reverse-engineering – diachronic model-based and synchronic model-based conceptual reverse-engineering – faces a disadvantage that the other does not. Diachronic model-based conceptual reverse-engineering commits something like the genetic fallacy. Synchronic model-based conceptual reverse-engineering cannot reveal practical necessity in our conceptual practices. As such, I argue that it is worthwhile to engage in both types of model-based conceptual reverse-engineering: doing so can more robustly confirm one's hypothesis than engaging in each on its own.

3.1. Diachronic modelling and the genetic fallacy

The primary aim of an hypothetical genealogy is explanatory: one explains why we have some item in terms of its meeting a need of ours. However, many theorists who give genealogical explanations take these explanations to have normative import: for example, to demonstrate that some concept or practice is valuable because it serves some purpose that we reflectively endorse as worthwhile; or that it is not valuable, because it doesn't serve any purpose that we reflectively endorse. Bernard Williams calls genealogies of the former kind "vindicatory" (2002: 38) and those of the latter "subversive" (283, n.20). For example, Hobbes's (2008) state of nature story is intended to show that instilling a government with absolute power would serve the purpose of maintaining the social contract that allows for the peaceful coexistence of citizens; this function is one that we reflectively endorse, thus Hobbes's hypothetical genealogy of absolute authority is vindicatory. However, one might worry that insofar as hypothetical genealogists take their genealogical explanations to have normative import, they commit the genetic fallacy: from purely descriptive facts about how a concept or practice developed, or might have developed, they derive normative claims about the value of that concept or practice. Colin Koopman makes this objection to

Williams's genealogical explanation of *truth*, writing that Williams "commits the genetic fallacy in conflating genesis and justification" (2013: 20) when he "attempt[s] to deploy historical inquiry into the (actual or hypothetical) emergence of present practices in order to establish a normative evaluation of those practices" (87).

But what exactly is the genetic fallacy, and does it apply to hypothetical genealogy? The term 'genetic fallacy' was introduced by Morris Cohen and Ernest Nagel, who identify two distinct fallacies under this name. The first form of the genetic fallacy "takes a logical order for a temporal order" (1934: 388): it fallaciously infers from the fact that some feature of an item seems to us essential to that item that the feature must have arisen early in the item's history. For example, Cohen and Nagel lambast:

> … all attempts current in the eighteenth century, and still widely popular, to reconstruct the history of mankind … on the basis of nothing but speculations as to what must have been. The theories as to the origin of language or religion, or the original social contract by which government was instituted, which were based on empirical unsupported assumptions as to what 'the first' or 'primitive' man *must* have done are all historically untenable. (388-9)

These reconstructions of the history of some practice invariably assume that "the earlier stages [must have] been simpler, and the later stages more complex" (389). This way of thinking is appealing to us "because we can understand the present complex institutions better if we see them built up out of simpler elements". But that some way of thinking is appealing to us does not mean that it is accurate. As such, Cohen and Nagel call it an "inexcusable error" to conflate the temporal order in which some practice actually developed with the logical order by which we can best understand that practice (389).

Amia Srinivasan notes that it is this form of the genetic fallacy that poses a *prima facie* problem for vindicatory genealogies (2019: 129, n3). However, as Queloz points out, this is only for genealogies that are intended as "conjectural histories", rather than diachronic models (2021: 23). The hypothetical genealogies of Craig and Kelp are not speculations about the actual history of our practice with *knowledge*. Far from this: the method of hypothetical genealogy, understood as involving the construction of a diachronic model, invokes a state that is explicitly alleged not to have obtained, because it involves "unrealistic or unstable idealizations" (Queloz 2021:

214). Then this form of the genetic fallacy does not apply to Craig or Kelp's hypothetical genealogies of *knowledge*, which use models precisely to show that we can understand our "present complex institutions if we see them built out of simpler elements" (Cohen and Nagel 1934: 389). More generally, hypothetical genealogy does not commit the first fallacy that Cohen and Nagel identify.

The second fallacy identified by Cohen and Nagel involves "the supposition that an actual history of any science, art, or social institution can take the place of a logical analysis of its structure" (1934: 389-90). The error here is to identify the temporal order in which we came to understand some domain as an order that is inherent in that domain. For example, many theorems of geometry were discovered before there was any suspicion that these theorems were systematically connected. As such, "[t]he logical priority of the axioms [of geometry] to the theorems is … not identical with temporal priority in our apprehension or knowledge" (390). More generally, "[t]he temporal order in which we learn or acquire our knowledge is not, in general, the same as the logical order of the propositions which are constituents of that knowledge" (390).

A genealogical explanation of some concept or practice may seem to fall foul of this second genetic fallacy. Consider Miranda Fricker's interpretation of hypothetical genealogy. According to Fricker, in a hypothetical genealogy, "what is claimed about the State of Nature – for instance, that it contains a concept or practice with such and such features – is really a claim about what is *basic* (or 'core') in our *actual* concept or practice" (2019: 244). (This interpretation is well-supported by Craig's own writings; for example, he states that it is his "thesis that the method reveals the core of the concept as it is to be found now" (2007: 191).) As the concept or practice develops in the genealogical story, the form it takes is "increasingly contingent" (Fricker 2019: 244). Then some feature of the concept or practice arising temporally later in the genealogical story indicates that it is a "contingent cultural iteration of the basic paradigm", best explained as "derivative" of the core case (245), and thus not itself capable of explaining why we partake in the concept or practice. But is this really an instance of the genetic fallacy? Again, crucially, we must note that hypothetical genealogists are definitively not trying to reconstruct the actual history of a concept or practice. Indeed, on Fricker's understanding of hypothetical genealogy, temporal priority within the genealogy is "largely a metaphor for *explanatory priority*, regardless of what actually (in historical time) came first" (2019: 245). That some feature of the

concept comes earlier in the genealogical story does not mean it is posited to come earlier in actual history. Then hypothetical genealogy is not subject to the second of Cohen and Nagel's fallacies, either.

However, neither of Cohen and Nagel's fallacies are quite what contemporary theorists generally intend when they write of the 'genetic fallacy'. Rather, what is typically intended in levying an accusation of committing the genetic fallacy is that a theorist conflates the origins of some theory, claim, argument, and so on, with its normative standing. Following Hans Reichenbach, this is often put in terms of conflating the "context of discovery" of that theory, claim, and so on with its "context of justification" (1938: 6-7): the way it was actually discovered ("the thinker's way of finding the theorem" (6)) with whether it is in good epistemic standing. This is what Koopman is getting at when he says that Williams's genealogy "conflat[es] genesis with justification" (2013: 20). Queloz addresses whether this accusation is properly levelled at hypothetical genealogy. As conceptual practices are not discovered so much as developed or formed over time, Queloz talks about the "context of formation", rather than the context of discovery, of a conceptual practice (2021: 214). The charge against hypothetical genealogy would then be that it conflates the context of formation of a conceptual practice with its context of justification.

For a final time, we must remember that hypothetical genealogies do not attempt to reconstruct an item's actual history. As such, the hypothetical genealogy of some concept or practice won't conflate the actual context of formation of that concept or practice with its context of justification: the actual context of formation does not play any role in the hypothetical genealogy. But hypothetical genealogy, in giving a how-possibly story of a conceptual practice's genesis, generates a merely possible context of formation. A theorist could conflate this merely possible context of formation with the actual conceptual practice's (or the concept's) context of justification. This seems at least as problematic as conflating the conceptual practice's actual context of formation with its context of justification (or that of the concept it generates). For if we cannot infer anything normative from the actual history of some concept or practice, then surely we cannot infer anything normative from its merely possible history. So a hypothetical genealogist must justify any normative conclusions she makes just as much as a conjectural historian must do. In what follows, I will continue to talk in terms of the 'origins' of concepts and the 'context of formation' of conceptual practices, but

this should from now on be understood such that these origins and contexts can be merely possible.

Queloz argues that, once we get clear on what exactly is fallacious in this way of understanding the genetic fallacy, we will see that hypothetical genealogies do not necessarily commit the fallacy – though they can do so. He insists that the genetic fallacy is not simply treating facts about the formation of a conceptual practice as relevant to the normative standing of that practice. Rather, the fallacy is treating these facts as relevant to the context of justification *when they are not* (2021: 214). Facts about a conceptual practice's formation can be relevant to its normative standing, but "only if there is a *connection* between some aspect of the context of formation and the justification of the item in question" – that is, of the conceptual practice or concept (214). Whether a theorist commits the genetic fallacy in taking her hypothetical genealogy of a concept to have normative import thus depends on whether there is the right kind of connection between the conceptual practice's context of formation and its context of justification.

According to Queloz, genealogical explanations, even hypothetical ones, can reveal the right kind of connection between context of formation and context of justification. Queloz distinguishes three ways that hypothetical genealogies can reveal facts in a conceptual practice's context of formation to bear on its context of justification. Following Williams (2002), he calls these three kinds of genealogies "vindicatory", "non-vindicatory" and "subversive" (2021: 217). Schematically, in each kind of genealogy, a group G engages in a conceptual practice P, and the genealogy reveals that the best explanation for why G engages in P is that P is the result of some formation process F(P). In a vindicatory genealogy, F(P) offers reasons to engage in P over merely possible rivals to P, where "possible rivals to P are unrealized alternatives to P" which "notably include the abandonment of P" (216). The genealogy thereby justifies G in continuing to participate in P. For example, Craig's hypothetical genealogy of *knowledge* reveals that we have a practical need to flag good informants, and shows how a concept that meets this need would plausibly change to look like our concept *knowledge*. He then makes the abductive inference that our concept *knowledge* is what meets this need for us. If this is so, then we have reason to continue engaging with our conceptual practice *knowledge* over at least one possible rival to this practice, namely its abandonment: then nothing would meet our need to flag good informants. In a non-vindicatory genealogy, in contrast, F(P) fails to yield reasons to prefer P over

merely possible rivals to P, and thereby fails to justify the continuation of P. In a subversive genealogy, F(P) is "incriminating", in that it offers reasons against the continuation of P (217). Thus Queloz concludes that hypothetical genealogies can reveal the right kind of connection between context of formation and context of justification to avoid committing the genetic fallacy.

But a worry remains. A genealogy can make plausible that a conceptual practice arose in order to meet some need, without showing that the practice still meets that need. Consider: one could tell a plausible genealogical story according to which the practice of monogamous marriage arose to ensure the paternity of children and thereby secure the transference of familial wealth and property (see Engels 1884, Chapter 2, §4 for such a story). But even if it is true that marriage as we know it originated in order to serve this need, it doesn't follow that our practice of marriage now functions to ensure the paternity of children. Indeed, this is likely not the case. There are cheaper and more accurate ways of determining disputed paternity, most obviously the use of DNA kits. This suggests that the need to ensure paternity is now better served by other practices. Similarly, a hypothetical genealogist might make plausible that a concept could have arisen in order to serve a particular need, but this wouldn't show that the concept still serves this need. Consider again Craig's hypothetical genealogy of *knowledge*. Even if Craig makes plausible that *knowledge* might have arisen in order to meet our need to flag good informants (this best explains why it has the shape that it does), this doesn't show that *knowledge* still meets this need for us. We now have a number of epistemic concepts that could serve this function. Most strikingly, we have concepts of trustworthiness and reliability. Why not think that these concepts, now we have them, are what actually and presently meet our need to flag good informants?

Patrick Rysiew raises an objection in this vein to Craig:

> … notice that we have at our disposal other epistemic terms we can and do use in picking out good informants – 'trustworthy', 'reliable', 'always right', etc. But given that there are plenty of terms available for picking out informants of one or another degree of reliability, what's special about '*know(s)*'? (2012: 278)

However, Rysiew doesn't draw into his objection the diachronic aspect of Craig's picture, and as such his objection has limited force. For Craig could respond to Rysiew that his genealogy is intended to be non-vindicatory. Then it doesn't undermine his

genealogy of *knowledge* whether we use other concepts to serve the function of flagging good informants, so long as we also use *knowledge* for this purpose. His non-vindicatory genealogy would explain why we have *knowledge*, without demonstrating that we have reason to continue engaging in this practice over possible rivals, which may include using other concepts to serve the same purpose. A stronger objection explicitly questions whether Craig's diachronic model gives us reason to think that *knowledge* does, presently and actually, function to flag good informants. We can grant that Craig has constructed a model out of uncontroversial facts about the kinds of creatures we are, the environments we live in, and so on; that he has told a plausible story about why a concept to flag good informants would arise in this model, and why this concept would change when further facts are added into the model in such a way that it would quite closely resemble our intuitive concept *knowledge*; such that he may well have articulated an idealised version of the actual origins of our concept *knowledge*. But even granting all this, it might be that *knowledge* does not, presently and actually, function to flag good informants; rather we now use other concepts for this purpose. So a genetic fallacy-like problem remains for diachronic model-based conceptual reverse-engineering. This problem doesn't arise for synchronic model-based conceptual reverse-engineering, which doesn't make reference to the development, actual or hypothetical, of a conceptual practice.

3.2. Synchronic modelling and contingency

Synchronic model-based conceptual reverse-engineering faces a problem of its own, which is that it struggles to demonstrate the practical necessity of our concepts: they cannot show whether, for some concept C, given the kinds of creatures we are in the kinds of environments in which we live, we must have C, or something very like it. In order to achieve sufficient resemblance between the target system and our actual conceptual practice to be able to apply her findings about the target system to the actual practice, the theorist must build more into her model at the stage of its construction than the diachronic modeller need do. For example, recall from §2.4 that Hannon must build into his model at the point of construction that the agents in the model engage in practices of pooling and sharing information; this is something that arises later in Craig's diachronic model. But the more that has to be built into the model in its construction, the less plausible it is that the needs that arise in the model are needs that any creatures like us would have, and as such, the less plausible it is that

the concept that meets this need is one that would inevitably arise in connection with this need.

Given that the kinds of concepts that theorists have been interested in reverse-engineering, such as *knowledge*, occupy such a central role in our lives (see Hannon's motivation for his project, discussed in Chapter 1, §2.1), that we might only have these concepts contingently (where the relevant domain of possibility here is the same as that for practical necessity) can give rise to what Srinivasan calls "genealogical anxiety": the anxiety that discovering the causal origins of our concepts "will somehow undermine, destabilize, or cast doubt on the legitimacy" of those concepts (2019: 128). If one of our concepts turns out to be such that we could have "so easily" done without it and "cut up the world in terms of rival concepts" (128), our entire conceptual practice (which, in the case of *knowledge*, is far-reaching and central to our lives) begins to look arbitrary.

This problem doesn't arise for diachronic model-based conceptual reverse-engineering. When constructing her diachronic model, the theorist need add to it only facts about the kinds of creatures that we are and the environments in which we live that are uncontroversially true about all humans living on Earth. It is the progression of the model that shows how even this minimal situation gives rise to a conceptual need, which can be met by the concept of interest, or an ancestor of that concept; further needs which are anticipatable from within the model then shape the ancestor concept to look more like the concept of interest. Diachronic model-based conceptual reverse-engineering can thus show that any creatures like us, in environments like ours, would have a need for a concept like the concept of interest. This modal insight can then, as Queloz puts it, "bolster our sense of a [conceptual] practice's necessity and thereby defend the way we go on against … genealogical anxiety" (2021: 58).

3.3. Robust confirmation

As diachronic model-based conceptual reverse-engineering faces a problem that synchronic model-based conceptual reverse-engineering does not face, and *vice versa*, then engaging in only one conceptual reverse-engineering project can only confirm one's original hypothesis about the function of a concept to a certain degree. However if a theorist engages in both kinds of model-based conceptual reverse-engineering, her hypothesis will be more robustly confirmed – provided that both projects do confirm her hypothesis. As such, a theorist interested in model-based conceptual reverse-

engineering would do well to engage in both kinds of projects: diachronic-model based conceptual reverse-engineering and synchronic model-based conceptual reverse-engineering. I do this myself, to test my hypothesis about the concept *doubt*, in the next chapter.

4. Advantages of the inquiry-stopper picture over the informant-flagging picture

In this section, I present three advantages of the inquiry-stopper picture of the function of *knowledge* over the informant-flagging picture. These are, first, that the inquiry-stopper picture is more robustly supported than the informant-flagging picture; second, that the inquiry-stopper picture of *knowledge* better captures the modal structure of *knowledge*; and third, that the informant-flagging function can be explained in terms of the inquiry-stopper function but not *vice versa*.

4.1. Robust support

That *knowledge* functions to flag good informants is made more plausible by Craig's diachronic modelling than by Hannon's synchronic modelling. Although we do, actually and presently, have a conceptual need to flag good informants, we have many concepts that we can use for this purpose, such as *trustworthiness* and *reliability*. As Hannon's synchronic model-based conceptual reverse-engineering project fails to demonstrate the practical necessity of a concept *knowledge* that serves this function (see §2.4 and §3.2), we have no reason to think that *knowledge* is what serves this purpose for us rather than some other concept.

In contrast, that *knowledge* functions to signal when inquiry should come to an end is made plausible by both Kelp's diachronic modelling and Kappel's synchronic modelling. Kelp's diachronic model shows that, from a minimal description of the kinds of creatures we are and the physical environments we inhabit, a need arises for a concept to flag when inquiry should come to an end; given further needs we can predict from inside this model, namely the needs to share and pool information to be used by others and on future occasions, we should expect this concept to change shape to look very much like the concept *knowledge* that we recognise as ours. Meanwhile, Kappel's model shows that, given three observations about the kinds of creatures we are and the physical environments we inhabit, we presently have a need for a concept to flag when inquiry should end; and the intension that such a concept would have closely

resembles the intuitive intension of our concept *knowledge*. Thus the inquiry-stopper picture of the function of *knowledge* is more robustly supported than the informant-flagging picture.

4.2. Captures the modal structure of *knowledge*

A second advantage of the inquiry-stopper picture of *knowledge* over the informant-flagging picture is that the inquiry-stopper picture, but not the informant-flagging picture, neatly explains the modal structure of knowledge. Many theories of *knowledge* have as a necessary condition on knowing that a subject's belief enjoys some kind of modal stability, such as true belief in close worlds (Sainsbury 1997, Sosa 1999, Pritchard 2005) or in some range of relevantly similar worlds (Dretske 1970, Stine 1976, Goldman 1976, Lewis 1996). If *knowledge* functions to signal the legitimate end of inquiry, that knowledge (the object of *knowledge*) has such a modal profile is easily explained. To see this, we must say a bit more about questions.

Recall (from Chapter 1, §2.4) that, on the dominant picture of questions, a question is a partition on possibility space, creating jointly exhaustive and mutually incompatible cells that determine the possible answers to the question (Groenendijk and Stokhof 1984). For example, the question 'Which of the Olsen twins is in Paris?' partitions possibility space thus:

Neither Olsen twin is in Paris	Both Olsen twins are in Paris
Mary-Kate, but not Ashley, is in Paris	Ashley, but not Mary-Kate, is in Paris

A complete answer to a question is one that fully settles the question, by ruling out all but one cell as that which contains the actual world. A partial answer rules out some cells as those which contain the actual world, but leaves open more than one. If we understand Q-inquiry as a process of ruling out various possible answers to Q (as does Kappel; see §2.5), then when an inquirer has done enough to legitimately terminate inquiry, we should expect her to have ruled out all but one cell that makes up her question as that which contains the actual world. But then it is clear that she is in a state that has a particular modal profile: what matters is not just her relationship to the true answer to Q, but her relationship to the various non-actual possibilities she has ruled out in getting to this answer.

Does this picture suggest any particular modal condition on *knowledge*? That is, does this way of understanding what it takes for S to have legitimately terminated inquiry tell us what is the 'particular modal profile' that her resulting knowledge-state

must have? We might think that the answer to this question lies in what counts as a 'possible answer' to a question. At first blush, it seems that what counts as a possible answer to a question Q is a pragmatic issue, to be determined by looking to features of the conversational context in which the question is asked, including the speaker's purposes in asking the question and the background knowledge of the conversational participants. For example, if I ask my partner where the cat is, the "suitable answers" (Groenendijk and Stokhof 1984: 211) to this question will be restricted to propositions about rooms in our flat: the cat is in the living room; the cat is in the bedroom; the cat is in the bathroom; the cat is in the kitchen. This is because we both know, prior to his answering my question, that she must be in one of these rooms. As such, the thought goes, the following propositions are not possible answers to my question: the cat is in my office at university; the cat is at my partner's mother's house in Port Glasgow; the cat is on the moon. For we both know that she is not in any of these places. On the other hand, the following is not a suitable answer to my question, either: the cat is in the flat. This is because we both already know this to be true, prior to his answering my question. Ferene Kiefer puts the point by saying that suitable answers to questions are "pragmatically adequate": they are both informative and useful, relative to the conversational context (1988: 258).

But we must tread carefully here. The matter of what counts as a *suitable* answer to a question is not the same as that of what counts as a *possible* answer to a question. The standard view among linguists and philosophers of language working on questions is that what counts as a suitable answer to a question is a pragmatic issue, which requires looking to features of the conversational context in which the question is asked, including the speaker's purposes in asking the question and the background knowledge of the conversational participants (see van Fraassen 1977; Groenendijk and Stokhof 1984). But what counts as a possible answer to a question is not determined by pragmatic facts; rather, it is standardly understood as a matter of the semantics of the question, such that the same question Q will have the same set of possible answers regardless of pragmatic features of the context in which Q is asked. Schematically, we can put the standard view thus: a possible answer to a question Q is any proposition P such that, were S to know that P, S would thereby know the answer to Q (see Diegan ms. for this heuristic for identifying a question's possible answers, though the idea that the possible answers to a question are all those propositions that could resolve the question goes back to Hamblin (1958, 1973) and has been

developed by Groenendijk and Stokhof (1984) and Ciardelli, Groenendijk and Roelofsen (2019)).

This makes the set of possible answers to a question hugely expansive, for many types of question. Not 'Whether P?' questions, whose possible answers are either P or not P. But for 'where', 'why', 'who', and so on, questions, the set of propositions such that, if S knew that P, S would know the answer to the question will be huge. Consider for example the question 'Who is the President of the U.S.A.?'. Any proposition of the form 'X is the President of the U.S.A.', where 'X' is a singular term, will be such that, if S knew that proposition, S would know the answer to the question. If S knew that Margaret Thatcher is the President of the U.S.A., then S would know the answer to this question, and as such, 'Margaret Thatcher is the President of the U.S.A.' is a possible answer to this question; if S knew that the dirty sock under her bedroom drawers is the President of the U.S.A., then S would know the answer to this question, and as such, 'the dirty sock under S's bedroom drawers is the President of the U.S.A.' is a possible answer to the question.

So, to return to our question: does this way of understanding what it is for S to have legitimately terminated inquiry tell us what is the particular modal profile that her resulting knowledge-state must have? Not really. I said that, in order to have legitimately terminated inquiry, S must have ruled out various possible answers to Q. As such, the state she is in as a result – the state of knowledge – will be related not just to Q's actual answer, but to these various possible answers. Possible answers are possible worlds, thus the state of knowledge has a modal profile: it relates a subject to some set of other possible worlds. But getting clearer on what it takes for a proposition to be a possible answer to a question has not told us anything about which possible answers S must rule out in order to know the answer to Q, nor what it means to 'rule out' a possible answer to Q. As such, it is left open what it takes for some possible world to be a member of the set of worlds to which S must be related if she knows that P as an answer to Q, and exactly what is the relation that S bears to those worlds. The framework is thus compatible with various modal conditions on knowledge, such as safety (Sainsbury 1997, Sosa 1999, Williamson 2000, Pritchard 2005), sensitivity (Nozick 1981) and relevant alternatives theory (Dretske 1970, Stine 1976, Lewis 1996). I take this to be an advantage for the inquiry-stopper picture of the function of *knowledge*, as it can be endorsed by a variety of epistemologists who posit modal conditions on knowledge.

In contrast, the informant-flagging picture does not easily explain why knowledge should have a modal profile. It is not at all clear why an inquirer seeking an informant should bear any particular relationship to non-actual possible worlds; what matters to him is that he actually gets told the truth, and that's that. Craig argues that this is too quick, as in fact an inquirer "cannot help being interested in the contents of possible worlds as well as those of the actual" (1990: 19). For the inquirer's knowledge of the actual world is incomplete, not just regarding the question for which they seek an informant, but for "all sorts of things about himself, the environment, and the potential informant. There are, in other words, enormously many propositions such that he does not know whether A or not-A, whether B or not-B, and so on" (19). If we think of possible worlds as demarcated by the totality of propositions that are true in those worlds, then there are "indefinitely many possible worlds any one of which, so far as he knows, might be the actual world" (19). Therefore, Craig holds, the inquirer will "hope for an informant who will give him the truth about *p* whichever of all these possibilities is realised. Which is to say … he wants an informant who will give him the right answer in a range of possible worlds" (20). In particular, he wants an informant who would give him the right answer in a range of worlds that are both close to the actual world and epistemically possible for him: compatible with what he knows (21-22).

The problem with Craig's argument is that, although there are many close worlds that are epistemically possible for an inquirer, she can know which world the actual world is simply via ostension: the actual world is *this* world, the world she is in. Given that she knows that she is in the actual world, all that matters to her is that the informant actually tells her the truth. Then she can 'help being interested', to use Craig's turn of phrase, in what goes on in merely possible, non-actual worlds. So we have been given no reason to think that knowledge would have a modal structure, on the informant-flagging picture. Another point in favour of the inquiry-stopper function.

4.3. Explanatory priority

Hannon argues that insofar as *knowledge* functions as the inquiry-stopper, this can be explained in terms of the informant-flagging function. He writes:

> … the way to reasonably terminate inquiry is by identifying a sufficiently reliable informant. A reliable informant as to whether *p* is someone from whom we can

> take it that *p*, which is to say that we treat her word on whether *p* as settling the question whether *p*. … This connection explains why the functional role of flagging reliable informants also serves to mark the point at which further inquiry is unnecessary. Spending more time and resources to continue one's inquiry would be impractical: continuing to inquire beyond this point would commit us to paying higher "informational costs" that are not worth the lessened risk of being wrong. (2019: 109)

He concludes that the informant-flagging function of *knowledge* is "more fundamental" to the inquiry-stopper function, as it is "explanatorily prior" to it: "[t]hat we have found a reliable informant explains why we must terminate inquiry at a certain point. Without such an explanation, it is unclear when we should reasonably end inquiry" (109).

However this way of understanding the explanatory connection between the informant-flagging function and the inquiry-stopper function is tenuous at best. Obviously, identifying a reliable informant is not *the* way, the *only* way, to reasonably terminate inquiry. S can reasonably terminate her inquiry into the question 'What's the weather like today?' by going outside and seeing that it is sunny, for example. In this case, S has reasonably terminated inquiry without identifying a reliable informant. Then we cannot explain *knowledge*'s role in terminating inquiry in terms of its role in flagging reliable informants.

Hannon might respond that we understand this case, and others like it, as follows: S identifies herself as a reliable informant concerning a question Q, and this is why she terminates her inquiry into Q. It is part of Craig's objectivisation story that there will be occasions on which an agent recognises herself as a good informant whether P, and thus recommends herself to others (1990: 64-5). Hannon could borrow this idea from Craig to argue that, in some cases, an agent recognises that she herself is a sufficiently good informant regarding whether P that she may terminate her inquiry into this question.

But applied to our example, this response is either clearly false or unnecessarily complicated. It is clearly false if what this means is that the inquirer takes herself to be a reliable informant regarding what the weather is like, asks herself what the weather is like today, tells herself the answer and terminates inquiry upon receiving this answer. Rather, she has to actively engage in inquiry: she must rule out various possibilities (possibilities in which it is raining, snowing, cloudy) by going outside and seeing that it is sunny. But it is unnecessarily complicated if what it means is that, at the point at

which she goes outside and sees that it is sunny, S recognises that she is now a reliable informant regarding what the weather is like and should thus take it from herself, as it were, that it is sunny – that is, she should end her inquiry into this question. A much simpler explanation of what goes on in this case is that S ends her inquiry because she sees that it is sunny. Further, this response unnecessarily attributes higher-order propositional attitudes to the subject: in order to properly terminate her inquiry into the question 'What is the weather like today?', she not only needs to have some propositional attitude towards an answer to this question (for example, to believe that it is sunny), she also needs to have some propositional attitude towards the proposition that she is a reliable informant with respect to this question. Applying Ockham's razor to our propositional attitudes, we should prefer not to posit propositional attitudes beyond necessity. Understanding this case along the lines of this suggestion would mean positing two propositional attitudes where we otherwise would need to posit only one.

Thus I conclude that *knowledge*'s role in terminating inquiry cannot be neatly explained in terms of its role in flagging reliable informants. In contrast, I argue that *knowledge*'s role in flagging reliable informants can be neatly explained in terms of *knowledge*'s role in terminating inquiry. One way – not the only way, by any means – of terminating inquiry into Q is to find a reliable informant on Q. A reliable informant on Q will herself have appropriately terminated inquiry into Q. Then the concept that signals when a subject has appropriately terminated inquiry will apply to the informant, and once the informant has told the inquirer her information, to the inquirer as well. Thus the inquiry-stopper picture of the function of *knowledge* explains why *knowledge* can be applied to flag reliable informants.

5. Recalcitrant data for the inquiry-stopper hypothesis

In this section, I consider some recalcitrant data for the inquiry-stopper picture of *knowledge*. However, the cases I discuss in this section are not potential counterexamples to this picture in the same way that Seal of Confession, Secret Crush and Secret Sect from §2.2 are for the informant-flagging picture. Seal of Confession, Secret Crush and Secret Sect are cases in which our intuitions about whether a subject has knowledge diverge from the verdict issued by the concept generated by Craig's conceptual reverse-engineering project. In Seal of Confession and Secret Crush, the concept that is the output of Craig's project issues the verdict that the subject in the

case does not know, as s/he is not willing to share his/her information with anyone, so does not satisfy the weak channels of communication condition that remains in Craig's objectivised concept. In Secret Sect, the subject satisfies all the conditions for knowing, on Craig's account, but intuitively doesn't know. The following cases are not like this. Recall from §2.3 that the concept that emerges from Kelp's conceptual reverse-engineering project applies to some subject S with respect to the question whether P when S satisfies the following three conditions:

1. S has formed a belief on whether P,
2. S's belief is true,
3. S's belief stems from a highly reliable source.

The following cases are not cases in which we intuit that a subject knows P, but she fails to satisfy all three of these conditions; nor cases in which we intuit that a subject doesn't know that P, but she satisfies all three of these conditions. Rather, they are cases in which a subject intuitively knows that P, but may permissibly continue her inquiry into whether P.

I have already pointed out (in §2.2 of this chapter, though the idea is implicit in the discussion of hammers in Chapter 2, §2.1) that an item X can have a function F without manifesting that function on all occasions of its use, and indeed that X can have F even if, on some occasions, some item Y, rather than X, serves F, where $X \neq Y$. Nevertheless, insofar as these cases are typical of cases in which a subject has knowledge, some pressure is put on the original hypothesis that *knowledge* functions to signal the legitimate end of inquiry. Recall from §2.1 of the previous chapter that we can reverse-engineer the function of some item by looking to how it is used in typical cases. If these are typical cases in which the concept *knowledge* is used – cases that are representative of this general type, namely cases in which the concept *knowledge* is used – but they are such that the subject to whom the concept applies may not terminate her inquiry, this undermines the original hypothesis. As such, these cases constitute recalcitrant data and need to be explained away, even though they are not potential counterexamples to the concept of *knowledge* that emerges from Kelp's conceptual reverse-engineering project in the same way that Seal of Confession, Secret Crush and Secret Sect are potential counterexamples to the concept that emerges from Craig's project. So I will try to reconcile these cases with the inquiry-stopper picture of *knowledge*.

5.1. Brown's Surgeon

Jessica Brown offers the following case as a counterexample to the thesis, endorsed by John Hawthorne and Jason Stanley (2008) and Jeremy Fantl and Matthew McGrath (2009), that if S knows that P, then S is permitted to rely on P in her practical reasoning. But the case also constitutes a recalcitrant datum for the inquiry-stopper picture of *knowledge*. The case is as follows:

> *SURGEON*
>
> A student is spending her day shadowing a surgeon. In the morning he observes her in the clinic examining patient A who has a diseased left kidney. The decision is taken to remove it that afternoon. Later, the student observes the surgeon in theatre where patient A is lying anaesthetised on the operating table. The operation hasn't started as the surgeon is consulting the patient's notes. The student is puzzled and asks one of the nurses what's going on:
>
> Student: I don't understand. Why is she looking at the patient's records? She was in clinic with the patient this morning. Doesn't she even know which kidney it is?
>
> Nurse: Of course, she knows which kidney it is. But, imagine what it would be like if she removed the wrong kidney. She shouldn't operate before checking the patient's records. (Brown 2008: 176)

Though this view was not Brown's target, Surgeon sits uneasily with the inquiry-stopper picture of *knowledge*, as it is a case in which a subject (the surgeon) knows that P (patient A's left kidney is diseased), yet ought to inquire further (by checking A's records). Then in this case, *knowledge* does not manifest the inquiry-stopper function.

I think that Surgeon is easily explainable on the inquiry-stopper picture of *knowledge*. Note that in this case, it is very plausible that other kinds of norms to zetetic (i.e. to do with inquiry, see Friedman 2020) and/or epistemic norms are in play, namely professional norms. Though the surgeon knows that patient A's left kidney is diseased, she should still check whether this is the case before performing the operation. But it isn't at all clear that the normative force of the 'should' in this sentence is zetetic. It is more plausible that the surgeon should double-check because it is a professional norm for surgeons to always double-check what operation is to be performed before beginning surgery. Further, it could be the case that, relative to zetetic normativity, the surgeon should not inquire further, even though, relative to her profession's norms,

she ought to inquire further. Then we would have a case in which two domains of normativity clash: professional and zetetic norms offer different permissions and prohibitions on action. But this wouldn't undermine the hypothesis that *knowledge* functions to signal the legitimate end of inquiry. For it wouldn't be a case in which, relative to the normativity that governs inquiry *qua* inquiry, S both knows that P yet should inquire further into whether P.

5.2. Woodard's Locked Door

Elise Woodard offers another case in which a subject who knows that P may permissibly continue inquiring into whether P. However, in this case, the inquiry seems to be permissible relative to epistemic and/or zetetic normativity, rather than some other domain of normativity. Call this case Locked Door:

> Deming is quite confident that she locked the door behind her when she left for work. *Indeed, she knows that she did*. However, she decides to double-check that she locked the door by walking back to the door and trying to open it, (a) just to be sure. (b) By double-checking, she may also come to know that she knows that the door is locked. At the very least, she gains greater (propositional) justification for believing that she knows. (2021: 8, emphasis in original)

Woodard holds that this is a case in which a subject knows that P, yet epistemically/zetetically permissibly continues inquiring (by double-checking) into whether P past the point of having achieved knowledge.

One response to Locked Door is to argue that Deming, and indeed all those who double-check something that they know, are irrational. This is a bold claim, but it has been endorsed by a number of philosophers. Jane Friedman writes that "[if] one knows the answer to some question at a time then one ought not to be investigating that question, or inquiring into it further … at that time" (2017: 131); Matthew McGrath writes that cases in which a subject knows the answer to Q but continues inquiring "involve peculiarities (such as irrationality or fragmentation)" (2021: 472, n. 37); Hawthorne and Stanley write that there is "something to be said for the claim that the person who knows they have turned the coffee pot off should not be going back to check" (2008: 587). So proponents of the inquiry-stopper picture of *knowledge* could argue that cases of double-checking what one knows don't constitute recalcitrant data, as they are cases in which a subject acts in a way that she ought not to: because she

knows that P, she ought not be inquiring into whether P, which double-checking whether P involves; thus *knowledge*'s applying to her does signal that she should end any ongoing inquiries into the relevant question.

But as noted, it would be very bold to claim that double-checking whether P when you know that P is never permissible. As Woodard says, double-checking is often what "responsible agents" do: such agents "double-check their calculations and reassess their evidence, sometimes seeking out more" (2021: 2). However there are ways that we might be able to sweeten this pill. First, we could follow Keith DeRose (2002: 180) in distinguishing primary from secondary propriety, and hold that, although it is inappropriate in the primary sense to double-check whether P when one knows that P, it may be appropriate in the secondary sense. Primary propriety is a matter of compliance with a norm. Secondary propriety requires only *reasonably believing* that one complies with a norm. Then we might say that agents who double check whether P when they know that P (or know that not-P) act inappropriately in the primary sense, as they fail to conform to the norm 'Don't inquire into Q if you know the answer to Q'; but they may reasonably believe that they don't know that P (or that not-P), and as such, reasonably believe that they are not failing to comply with this norm. Then they would be inquiring appropriately in the secondary sense.

As Woodard sets up Locked Door, it's not clear whether this response will help in this case. Is it plausible that, prior to double-checking the door, Deming believed (with justification) that she didn't know that she had locked the door? Woodard writes that double-checking whether the door is locked might allow Deming to "come to know that she knows that the door is locked" (2021: 8). Her coming to know that she knows that the door is locked upon double-checking is compatible with her having any of the following higher-order attitudes prior to double-checking: believing (without knowing) that she knows that the door is locked; believing that she doesn't know that the door is locked; or not believing anything about whether she knows that the door is locked (i.e., having no higher-order attitude). So we could simply stipulate that Deming believes that she doesn't know that the door is locked, prior to double-checking. Is this belief justified? The case, so described, doesn't settle the matter. But again, we could stipulate that she justifiably believes that she doesn't know that the door is locked (perhaps she often thinks she remembers that she locked the door today, when in fact she is remembering locking the door some other day).

But at this point a different worry arises: if Deming does justifiably believe that she doesn't know that P, can it still be the case that she knows that P? Martin Smith has recently argued that having a justified belief that one doesn't know that P can defeat one's justification for believing that P (2022: §§3-4).[9] If justified belief is necessary for knowledge, this would imply that having a justified belief that one doesn't know that P defeats one's knowledge that P. If this is the case,[10] then if Deming has a justified belief that she doesn't know that P, this would defeat her knowledge that P. So she wouldn't be violating the relevant norm, that one ought not inquire into Q if one knows the answer to Q. Then appealing to primary and secondary propriety will not allow us to say that agents who inquire into Q while knowing the answer to Q behave inappropriately in the primary sense (because they violate this norm) but may behave appropriately in the secondary sense (because they justifiably believe that they don't violate it): by virtue of justifiably believing that they don't know that P, their knowledge that P will be defeated, such that it will become true that they don't know that P. Then they won't be behaving inappropriately in the primary sense: they won't have violated the norm according to which one mustn't inquire into Q and know the answer to Q.

So here's another attempt at sweetening the pill. One could hold that, while it may always be zetetically and/or epistemically impermissible to inquire into Q when one knows the answer to Q, other domains of normativity may issue conflicting instruction. This is one way of understanding Brown's Surgeon case: zetetic/epistemic normativity says the surgeon ought not check (thus inquire into) which kidney to remove, as she already knows the answer; but professional normativity says that she must check what surgery is to be performed immediately before beginning each procedure. Then there can be cases in which agents who inquire into Q while knowing the answer to Q behave permissibly – further, that they behave in the way they must – according to some other domain of normativity, even though they behave impermissibly according to zetetic and/or epistemic normativity.

---

[9] As Smith notes (2022, fn. 17), given the weak assumption that one cannot justifiably believe both a proposition and its negation, it is also a consequence of Kvanvig's principle that if one justifiably believes that P, then one justifiably believes that one knows that P (2009: §3), that if one justifiably believes that one doesn't know that P, then one does not justifiably believe that P.

[10] This is not uncontroversial: see Lasonen-Aarnio (2010, 2014) and Field (2021) for arguments that S can be justified in believing P while being justified in believing that she is not justified in believing P.

But if one still finds this response implausibly strong, another is available. Friedman (as well as Kelp (2021a) and Jared Millson (2021)) endorses the following norm of inquiry: one ought not to inquire into a question Q at a time *t* and believe a complete answer to Q at *t*. Call this 'DBI', for Don't Believe and Inquire (2019b: 303). As knowledge entails belief, Friedman also thinks it is impermissible to know a complete answer to Q and inquire into Q. However she doesn't think that double-checking is always zetetically impermissible. This is because, she argues, when a subject double (or triple, or *n*-tuple)-checks Q, she can "drop her belief" in the complete answer for the duration of her double-checking inquiry (2019b: 304). To double-check Q is to re-open Q; "a fully rational double-checker drops her answer belief before she re-opens the relevant question" (304). We could then understand Locked Door as follows: though Deming does know that she locked her door, in order to rationally double-check this, she must drop her belief until the conclusion of her double-check. If Deming does drop her belief (thus cease knowing) until the double-check has come to an end, she inquires permissibly. If she doesn't drop her belief, she does not inquire permissibly. On this way of understanding Locked Door and similar cases, it isn't that subjects who know P can never permissibly double-check whether P, just that they cannot permissibly know P *while* double-checking whether P.

Thus one who endorses the inquiry-stopper picture of *knowledge* can explain why it can be that *knowledge* functions to signal when inquiry should come to an end, yet there be cases in which a subject intuitively has knowledge but may continue her inquiry. In some cases, this will be because the force of 'may' in 'may continue her inquiry' is not zetetic. In other cases, a subject zetetically may inquire into Q because she knew a complete answer to Q before inquiring into Q, but drops her belief (thus doesn't have knowledge) for the duration of the inquiry. But it is crucial to reiterate that these kinds of cases are not potential counterexamples for the inquiry-stopper picture in the way that Seal of Confession and Secret Crush are for the informant-flagging picture. These are not cases in which a subject intuitively has knowledge but does not satisfy the conditions of the concept that is the output of an inquiry-stopper conceptual reverse-engineering project, or alternatively satisfies the conditions for the objectivised concept but intuitively lacks knowledge. For instance, Brown's surgeon and Woodard's Deming may both be such that they believe that P, their belief is true, and their belief stems from a highly reliable source; as such, both could satisfy Kelp's concept *objectivised protoknowledge*. Rather, these are cases in which *knowledge* is not

serving its hypothesised function. As noted, items can have a function yet not manifest that function in all cases: that we can use a hammer to break ice does not undermine the claim that the function of a hammer is to drive nails into things. Then those who endorse the inquiry-stopper picture of *knowledge* needn't be able to explain away these cases in order for the view to be defendable in the face of alternatives. Nevertheless, it doesn't hurt to be able to do so.

6. Conclusion

In this chapter, I surveyed four conceptual reverse-engineering projects on *knowledge*, offered by Edward Craig, Christoph Kelp, Michael Hannon and Klemens Kappel. I argued that the picture of *knowledge* endorsed by Kelp and Kappel, on which this concept functions to signal when inquiry should come to an end, has a number of advantages over the informant-flagging picture endorsed by Craig and Hannon. First, the hypothesis that *knowledge* functions to signal the legitimate end of inquiry is more robustly confirmed than the hypothesis that it functions to flag good informants, as it is confirmed by both synchronic model-based and diachronic-model based conceptual reverse-engineering, whereas the informant-flagging hypothesis is made plausible only by Craig's diachronic model-based conceptual reverse-engineering. Second, the inquiry-stopper picture, but not the informant-flagging picture, captures the modal structure of *knowledge*: it explains why knowing that P requires bearing some relation to other possible worlds. Third, the informant-flagging function of *knowledge* can be explained in terms of *knowledge*'s inquiry-stopper function, but the opposite is not true. Thus the inquiry-stopper function has explanatory priority over the informant-flagging function. I then considered and explained away some apparently recalcitrant data for the inquiry-stopper picture. I take myself to have shown that the inquiry-stopper picture of the function of *knowledge* is preferable to the informant-flagging picture. In the next chapter, I will argue that similar considerations as those motivating our conceptual need for an inquiry-stopper suggest that we have a distinct conceptual need for an inquiry-starter: a concept that flags when inquiry should begin. I argue that meeting this need is the function of the, or at least a, concept *doubt*.

Chapter 4. The point of *doubt*

1. Introduction

In the previous chapter, I argued that the hypothesis that the concept *knowledge* functions to signal when inquiry should come to an end generates an account of *knowledge* that is more plausible than that generated by the hypothesis that *knowledge* functions to flag reliable informants. In this chapter, I argue (in §2.1) that similar considerations as those that motivate our need for an inquiry-stopper concept also demonstrate a need for a concept to signal when inquiry should begin: an inquiry-starter concept. I test the hypothesis that the concept *doubt*, or at least one concept that we pick out using the word 'doubt', meets this need. I reverse-engineer this concept using two models: one synchronic (§2.2) and one diachronic (§2.3). Both models confirm my hypothesis. The picture of *doubt* that arises from my conceptual reverse-engineering projects is as follows: the concept *doubt* applies to some subject S with respect to some question Q when:

1. S has a questioning attitude to Q,
2. S does not believe any complete answer to Q,
3. S's situation with respect to Q is epistemically risky, or is represented to her as epistemically risky,
4. S is immediately motivated to inquire into Q.

In §3, I raise and respond to some objections to this picture of *doubt*. First, could Jane Friedman's concept of *suspended judgement* as "the most general questioning attitude" (2019b: 299) play the role of the inquiry-starter, instead of *doubt*? Second, do all inquiring creatures doubt? Third, is there a disanalogy between the roles of *knowledge* and *doubt* in inquiry that undermines my claim that *doubt* can be the inquiry-starter to *knowledge*'s inquiry-stopper? Finally, what kind of normativity governs the start of inquiry?

1.1. 'Doubting that' and 'doubting whether'

Before beginning my conceptual reverse-engineering project on *doubt*, I want to suggest that 'doubt' is polysemous. Polysemy is a linguistic phenomenon whereby a word has multiple but related meanings (Sennet 2016; Vincente 2018). I think that 'doubt' has two distinct, but related, meanings. In one sense, to doubt is a propositional

attitude, standardly characterised as something like thinking that a proposition is probably false (see Hookway 1998: 204). This is the sense of 'doubt' found in sentences like 'I doubt that Sally will come to your party.' In another sense, to doubt is a questioning (or 'question-directed' (Friedman 2013, 2017) or 'interrogative' (Archer 2018; Friedman 2019b; Willard-Kyle forthcoming)) attitude: an attitude one takes towards a question. When one doubts whether P, in this sense, one needn't take any propositional attitude to P, and in particular one needn't think that P is probably false. Christopher Hookway writes that to doubt whether P in this sense is to "find that the available evidence is insufficient to warrant either acceptance or rejection of a proposition – or hold that it is unclear whether the evidence is sufficient for that purpose" (1998: 205). I am interested in the concept *doubt* picked out in this second sense of 'doubt'. That is, I am interested in the concept of doubt as a questioning attitude.

This concept *doubt* is often in play in philosophy. For example, Descartes's First Meditation begins with reflection on the "highly doubtful nature of the whole edifice" of his beliefs, motivating him to "demolish everything completely and start again right from the foundations". To do this, he doesn't think it necessary to "show that all [his] opinions are false, which is something [he] could perhaps never manage." Rather, "for the purpose of rejecting all my opinions, it will be enough if I find in each of them at least some reason to doubt" (2017: 15). He notes that everything he has up until now accepted as true he has either acquired "from the senses or through the senses"; however, his senses have in the past issued him with false beliefs (16). As such, he imagines a case in which his sense experience could seem, to him, exactly as it does, and yet impart him with false beliefs: a case in which he is dreaming that he is "here, sitting by the fire, wearing a winter dressing-gown, holding this piece of paper in my hands" (16). In this case, he would be "convinced of just such familiar events – that I am here in my dressing-gown, sitting by the fire – when in fact I am lying undressed in bed" (16). He thus takes himself to have found a reason for doubting whether he is sitting by the fire in his dressing-gown: he might be dreaming in bed, in which case this belief would be false. But note that the attitude that Descartes now has is not that of thinking it is probably false that he is sitting by the fire, wearing a winter dressing-gown, holding a piece of paper in his hands. Rather, it is an attitude of finding that the evidence he has, his sense experience, doesn't rule out the possibility that he is

dreaming in his bed. Then we can understand Descartes here as *doubting whether*, rather than *doubting that*.

C. S. Peirce is similarly interested in the concept *doubt* that picks out a questioning attitude. He writes, for example, that doubt is distinguished phenomenologically from belief in that the "sensation of doubting" arises "when we wish to ask a question" while the sensation of belief arises "when we wish to pronounce a judgement" (1877: 5). For Peirce, "[d]oubt is an uneasy and dissatisfied state from which we struggle to free ourselves and pass into the state of belief; while the latter is a calm and satisfactory state which we do not wish to avoid, or to change into anything else" (5). Doubt and belief also differ in the kind of behaviours they motivate. When we believe that P, we are in "such a condition that we shall behave in a certain way, when the occasion arises" (6). But when we doubt whether P we are immediately motivated to act so as to settle the question: "The irritation of doubt causes a struggle to attain a state of belief" (6). He thus compares the experience of doubt to "the irritation of a nerve and the reflex action produced thereby", while belief is instead analogous "what are called nervous associations – for example, to that habit of the nerves in consequence of which the smell of a peach will make the mouth water" (6). So doubt is distinguished from belief, for Peirce, in three ways:

1. Doubt is related to questions, and belief to judgement;
2. Doubt is affectively unpleasant, and belief is affectively pleasant;
3. Doubt immediately motivates action, while belief creates dispositions to act in certain ways under certain conditions.

For Peirce, the "struggle to attain belief" that doubt motivates is inquiry, thus doubt motivates inquiry. He writes:

> The irritation of doubt is the only immediate motive for the struggle to attain belief. It is certainly best for us that our beliefs should be such as may truly guide our actions so as to satisfy our desires; and this reflection will make us reject any belief which does not seem to have been so formed as to insure this result. But it will only do so by creating a doubt in place of that belief. With the doubt, therefore, the struggle begins, and with the cessation of doubt it ends. (6)

Indeed, Peirce writes as if he thinks genuine inquiry is possible only when a subject truly doubts a question Q:

> Some philosophers have imagined that to start an inquiry it was only necessary to utter a question or to set it down upon paper, and have even recommended us to begin our studies with questioning everything! But the mere putting of a proposition into the interrogative form does not stimulate the mind to any struggle after belief. There must be a real and living doubt, and without this all discussion is idle. (6)

Once the subject stops doubting Q, her inquiry into Q must come to an end:

> Some people love to argue a point after all the world is fully convinced of it. But no further advance can be made. When doubt ceases, mental action on the subject comes to an end; and, if it did go on, it would be without a purpose. (7)

We thus get the following picture of doubt from Peirce: doubt is a questioning attitude; doubt is incompatible with belief; doubt is aversive; doubt motivates inquiry, and this motivation is immediate in that it isn't conditional on particular conditions obtaining, and doesn't require an independent desire to inquire.

In everyday English, we more often find 'doubt' picking out a propositional attitude than we find it picking out a questioning attitude. This might make one worry (or indeed doubt) whether there really is a questioning attitude of doubt, despite how philosophers use the term. But I think this worry is unfounded. Sentences in which 'doubt' takes a question complement are often perfectly intelligible to us. Most obviously, 'doubt' takes the question complement 'whether' with ease. Consider:

A. I doubt whether Maggie will come to the party.

B. I doubt whether your motives were good.

These sentences sound fine. However other sentences where 'doubt' takes a question complement sound bad:

C. # I doubt who will come to the party.

D. # I doubt what your motives were.

But note that these sentences can be drastically improved if we replace 'doubt' with 'have doubts about':

E. I have doubts about who will come to the party.

F. I have doubts about what your motives were.

Further, our 'bad' sentences, C and D, sound much better when 'doubt' is under negation:

G. I don't doubt who will come to the party.

H. I don't doubt what your motives were.

The orthodoxy in semantics has it that adding a negation shouldn't make a difference to whether a verb can take an interrogative clause (see Grimshaw 1979, 1990; Pesetsky 1991), so that 'doubt' takes other question complements under negation without awkwardness suggests that we understand what it would be to doubt questions beyond whether-questions.[11] All this suggests that 'doubt' can take a variety of question complements, which in turn suggests that there is a meaning of 'doubt' on which it picks out a questioning attitude.

I am interested in doubt as a questioning attitude. From now on, when I use the term '*doubt*', I mean to pick out the concept of doubt as a questioning attitude. We can take the characterisation of this concept so far suggested to make up our intuitive concept *doubt*. That is, the intuitive concept *doubt* has it that doubt is a questioning attitude; doubt is incompatible with belief; doubt is aversive; and doubt immediately motivates inquiry. If I have reason to refer to the concept that picks out the propositional attitude of doubt, I will make clear that that is what I am doing. My hypothesis, to be tested in the rest of this chapter, is that *doubt* functions to signal that inquiry ought to begin. In §2.2, I follow Kappel in constructing a synchronic model from three facts about us; in this model, we will see a need for an inquiry-starter concept arise. I will show that the concept that meets this need within the model is relevantly similar to the intuitive concept *doubt*. In §2.3, I follow Kelp in constructing a diachronic model: an epistemological state of nature consisting of creatures like us but who lack the concept *doubt*. We will see that these creatures have a need for an inquiry-starter concept, that the concept that would meet this need for these creatures would be fairly similar to the concept *doubt*, and that this concept would change to look more like *doubt* in response to practical pressures anticipatable from within the model.

2. Reverse-engineering *doubt*

In this section, I argue that similar considerations as those that motivated the inquiry-stopper picture of *knowledge* in the previous chapter suggest a distinct conceptual need for an inquiry-starter: a concept to signal when inquiry should begin (§2.1). I hypothesise that this is the function of the concept *doubt*. I then undertake two

[11] I am grateful to Peter van Elswyk and Thomas Stephen for discussion on this point.

conceptual reverse-engineering projects on *doubt* to test this hypothesis. The first (§2.2) makes use of a synchronic model, inspired by Kappel's conceptual reverse-engineering project on *knowledge* (discussed in Chapter 3, §2.5). The second (§2.3) makes use of a diachronic model, inspired by Kelp's conceptual reverse-engineering project on *knowledge* (discussed in Chapter 3, §2.3). I construct a hypothetical genealogy of *doubt*, akin to Kelp's hypothetical genealogy of *knowledge*. However, my genealogy has a further stage than Kelp's, wherein our ancestors come to regard the concept that functions for them as the inquiry-starter as intrinsically valuable, thus generating its own reasons for action. I develop this stage in §§2.3.2-2.3.3.

2.1. Our need for an inquiry-starter

Recall from §2.5 of the previous chapter Kappel's three "trivial observations" (2010: 74) about creatures like us, which demonstrate our need for an inquiry-stopper:

1. Truth matters: generally, we prudentially benefit from treating a proposition as true iff it is true.
2. We must inquire to get true beliefs, but inquiry is costly in terms of time and cognitive resources.
3. Inquiry has no natural stopping-point: we can, in principle, go on trying to rule out possible answers to our question for as long as we have time and resources.

Note that the first of these observations speaks to the importance of inquiring in general. It is only the second and third that speak to the importance of *stopping* inquiry, so suggest a need for an inquiry-stopper. But the second observation just as much suggests a need for a concept to signal when inquiry should begin: since inquiry is costly in terms of time and cognitive resources, we should only start inquiry when doing so would likely be worth the loss in time and resources. Further, an analogous consideration to (3) also suggests a need for an inquiry-starter: just as inquiry has no natural stopping-point, neither does it have a natural starting-point. For we can inquire into any question whatsoever. We can inquire into uninteresting questions, like 'How many bricks make up McEwan Hall?' We can inquire into unimportant questions, like 'What is the name of Reverend Lovejoy's dog on The Simpsons?' I argue that we can even inquire into nonsensical questions, like 'How blue is the cat's desire?' For I could endeavour to find out to which cat 'the cat' refers. This plausibly counts as taking steps into inquiring into this question. Compare: it certainly does count as inquiring into the

question ‘How blue are the cat’s eyes?’ that I endeavoured to find out to which cat ‘the cat’ refers to here.

All this to say: there are limitless questions into which we could inquire, but we are finite creatures with finite resources, so ought not to inquire into just any question. Rather, we should ration our time and resources to inquiring into just those questions that matter, for one reason or another. As such, we can list three facts from which the need for an inquiry-starter arises:

1. Truth matters: we generally prudentially benefit from treating a proposition as true iff it is true.
2. We have to inquire in order to get true beliefs, but inquiry is costly in terms of time and cognitive resources.
3. Inquiry has no natural starting-point: there are limitless questions into which we could inquire.

(1) to (3) give rise to a need for a concept that signals when inquiry should begin. What concept could play this role?

Patrick Rysiew makes a comment suggesting that he thinks *knowledge* can play the dual roles of inquiry-stopper and inquiry-starter. He writes that the function of *knowledge* is to “fulfil the need, in one’s linguistic social interactions and deliberations, for some way of marking the opening and closing of specific lines of inquiry – of indicating (and/or recommending) just which things may or should be reasonably assumed to be true, and so may (/may not) be open to reasonable further questioning” (2012: 275). That *knowledge*, for Rysiew, functions to mark the opening, as well as the closing, of lines of inquiry suggests that it is for him both the inquiry-starter and the inquiry-stopper. How could *knowledge* play both roles? Perhaps the idea is that in attributing *knowledge*, we signal that inquiry has gone on long enough and should now come to an end, but in denying *knowledge* – or attributing *non-knowledge* – we signal that some matter needs to be inquired into and so inquiry should begin.

But there are two reasons to think that *non-knowledge* could not play the inquiry-starter role. The first is that there’s nothing intrinsically wrong with an agent who fails to know the answer to a question Q without inquiring into Q. Kelly doesn’t know what is the capital of Kosovo, but there needn’t be anything wrong with her failing to inquire into this question. And we can assert that Kelly doesn’t know what is the capital of Kosovo without implying that she ought to inquire into this question. (This isn’t to say that there’s never anything wrong with a situation in which S simultaneously

fails to know Q's answer and fails to inquire into Q. For example, S might be practically irrational for being in such a situation, if she wants to know Q's answer, is capable of finding this out with relative ease, yet doesn't inquire into Q. But in this case, S is practically irrational because she's failing to do what is required to satisfy some practical goal she has. Failing to inquire into Q while failing to know Q's answer isn't irrational or otherwise bad just because of the kind of situation it is.) On the other hand, there is something intrinsically wrong with an agent who knows the answer to Q but continues to inquire into Q. Franny knows what is the capital of France, and as such, there would be something wrong with her inquiring into the question 'What is the capital of France?'. Bradley Armour-Garb writes that "there is a sort of incoherence between taking oneself to know something and going on to investigate further whether it is the case" (2011: 670); while Dennis Whitcomb says, continuing an inquiry past the point of knowing the answer "is like continuing to eat after being nourished" (2010: 640).

The second reason to think that *non-knowledge* can't be the inquiry-starter is that failing to know Q's answer, and even recognising that one fails to know Q's answer, won't motivate one to inquire into Q. When it occurs to Kelly that she doesn't know what is the capital of Kosovo, this needn't motivate her to inquire into this question. Compare this with the relationship between knowing, or maybe between recognising that one knows, and terminating inquiry: a subject who takes herself to know whether P will normally be motivated to end any ongoing inquiries she has into whether P.

Could Rysiew respond that Kelly's case is atypical? If this is so, then *non-knowledge* being unable to serve the inquiry-starter function won't show that it is false that a function of *non-knowledge* is to signal when inquiry should start. Recall that the sense of 'typical' that matters for conceptual reverse-engineering is that which means 'representative of its type'. Kelly's case seems typical of a case in which a subject lacks *knowledge*, and in which we would deny *knowledge* to a subject. For many substantive claims that are widely endorsed about *knowledge*, and its connection to other concepts, are clearly true in this case. For example, it is widely held that knowledge is the norm of assertion, such that a subject who lacks knowledge that P may not assert that P (Williamson 2000, DeRose 2002, Hawthorne 2004, Stanley 2005a, Schaffer 2008, Turri 2010). Intuitively, Kelly may not assert that Pristina is the capital of Kosovo. Similarly, many argue that knowledge is the norm of action, such that a subject who lacks knowledge that P may not act as if P (Williamson 2000, Fantl and McGrath 2002,

Hawthorne and Stanley 2008). Kelly may not act as if Pristina is the capital of Kosovo, for example by betting £1 against her flatmate on this being so. Then this response doesn't look promising.

We can see from *non-knowledge*'s failure to fulfil the role of the inquiry-starter two conditions we want the inquiry-starter concept to meet. The first is that there should be something intrinsically wrong with a situation in which a subject S falls under the concept with respect to some question Q, yet doesn't inquire into Q. As such, we should be able to implicate that S ought to inquire into Q by asserting that the concept applies to S. Second, S's being in the state picked out by the concept should motivate S to inquire into Q. *Non-knowledge* doesn't meet either of these conditions: Kelly falls under the extension of *non-knowledge* regarding the answer to the question 'What is the capital of Kosovo?', but there is nothing wrong, so *a fortiori* nothing intrinsically wrong, with her failing to inquire into this question; and Kelly's being in the state of non-knowledge with respect to this question, and even knowing that she is in this state, needn't motivate her to inquire. Hence *non-knowledge* cannot play the role of the inquiry-starter. Could *doubt* do any better?

2.2. Reverse-engineering *doubt* using a synchronic model

Recall that, on our intuitive concept outlined in §1.1, *doubt* picks out a questioning attitude; to doubt whether P is incompatible with believing that P; doubting is aversive; doubt motivates inquiry; and this motivation is immediate. *Doubt*, on this picture, satisfies the second requirement: one who is in the state that doubt picks out will be motivated to inquire into Q. What about the first requirement? Peirce has more to say about *doubt* that is relevant for this purpose. Peirce holds that doubt is triggered by confrontation with a "surprising phenomenon", which is "some experience which either disappoints an expectation, or breaks in upon some habit of expectation" of the inquirer (CP §6.469). The inquirer finds himself "confronted with a phenomenon unlike what he would have expected under the circumstances" (CP §2.776); he is "more or less placidly *expecting* one result, and suddenly finds something in contrast to that forcing itself upon his recognition" (CP §5.57). What the inquirer expects is largely determined by his body of beliefs. That some new phenomenon contrasts with what he expects therefore suggests that it imparts new information that conflicts, or at least sits uneasily, with his beliefs about the world. Insofar as this information is not misleading, this suggests that he's in a risky situation, epistemically speaking: he's at

risk of holding a false belief. Insofar as being in an epistemically risky situation is bad, and insofar as inquiry can get him out of this risky situation, he then ought to inquire. As such, there is something intrinsically wrong with a situation in which a subject doubts Q without inquiring into Q: the subject is failing to do something he ought to do.

However, being at risk of holding a false belief is not the only way that one can be in an epistemically risky situation. Risk has to do with the potential of some disvaluable event to occur. Epistemic risk, then, can be understood as the potential for some epistemically disvaluable event to occur. A variety of events can be epistemically disvaluable. Holding a false belief is epistemically disvaluable, but so can be missing out on true belief, or failing to understand something, or failing to know. Suppose that some question Q pops into S's head, seemingly out of nowhere. She doesn't presently believe any answer to Q, but nevertheless, the question vexes her: it seems to her in need of an answer. She consequently inquires into Q. In this case, we could say that S's situation is epistemically risky, not because she is at risk of forming a false belief, but because she is at risk of missing out on some worthwhile true belief or piece of knowledge. So let's say that a subject who doubts a question Q is in an epistemically risky situation; or, at least, her experience of doubt represents her situation to her as epistemically risky.

We then end up with the following picture of *doubt* as applying to some subject S regarding some question Q when:

1. S has a questioning attitude to Q.
2. S does not believe any complete answer to Q.
3. S's situation with respect to Q is epistemically risky, or is represented to her as epistemically risky.
4. S is immediately motivated to inquire into Q.

Doubt, so understood, meets both requirements of the inquiry-starter. If one is in an epistemically risky situation, then one ought to do something to reduce that risk. If one's situation is represented to one as epistemically risky, then from one's own perspective, one ought to do something to reduce the relevant risk. Inquiring into a question Q will typically reduce the epistemic risk that attaches to some subject S with respect to Q. So if the concept *doubt* applies to a subject with respect to a question Q, then either she ought to inquire into Q (if her situation is genuinely epistemically risky), or it is the case that, from her perspective, she ought to inquire into Q (if her situation is represented to her as epistemically risky). This 'or' is inclusive: one's situation can

be both epistemically risky and represented to one as epistemically risky. In either situation, there is something intrinsically wrong with a situation in which a subject doubts Q without inquiring into Q: either she is failing to do something that she ought to do, or she is failing to do something that, from her own perspective, she ought to be doing. The former is a case of failing to abide by a norm, and the latter is a case of *akrasia*. Neither is a happy situation. So *doubt* meets the first requirement. And S's being in the state picked out by the concept motivates S to inquire. So *doubt* meets the second requirement. Thus this synchronic model confirms my original hypothesis, that *doubt* functions to signal when inquiry should begin: this explains why *doubt* would have the features that it typically has.

2.3. Reverse-engineering *doubt* using a diachronic model

In this section, I construct a model of a state of nature from a set of facts about us. At the first stage of this model, the creatures in the state of nature do not have the concept *doubt*. We will see that a need arises in this model for an inquiry-starter. I will show that the concept that meets this need in the state of nature model would change over time, in response to various practical pressures anticipatable from within the model, to look like our concept *doubt*. My state of nature story mirrors that offered by Kelp to test his hypothesis that *knowledge* functions as the inquiry-stopper. However, it has a further stage than Kelp's genealogy of *knowledge*, wherein our ancestors come to treat their inquiry-starter concept as intrinsically valuable. This stage of my genealogy (discussed in §§2.3.2-2.3.3) appeals to Bernard Williams's genealogy of *truth*.

2.3.1 A hypothetical genealogy of doubt

Imagine a state of nature, consisting of creatures very much like us, but who lack the concept *doubt*. As in Kelp's model, these creatures – our ancestors – have the same needs as us for food, water, shelter and so on. In order to survive, they need true beliefs about their environments: about what is safe to eat; about where they can find shelter; and so on. To get true beliefs, they must inquire: they need to set about trying to find the answers to these questions. But they can inquire into any number of questions, and their inquiries will be, to different degrees, costly in terms of time and precious resources. Further, questions into which they can inquire are more or less worthwhile: an inquiry into 'Where can I find fresh water?' would be more worthwhile than an inquiry

into ‘Where can I find a place to sit?’ As such, they have a need for a concept that flags when an inquiry would be worthwhile, so they should undertake it.

What would this concept look like? To answer this, imagine that some question Q is open for one of our ancestors. What properties would our ancestor want to have before she will begin an inquiry into Q? Here are some suggestions:

1. She has a questioning attitude to Q.
2. She doesn’t already believe a complete answer to Q.
3. She believes that the complete, true answer to Q is available through inquiry.
4. Settling Q would advance her current practical goals.

The property described by (1) is trivial. To inquire into Q, our ancestor must have Q “open in thought” (Friedman 2019b: 299), she must be asking Q. Regarding (2): if our ancestor believes an answer to Q, she will believe that inquiry into Q would be pointless, as it won’t imbue her with more true beliefs than she currently takes herself to have. This idea is reminiscent of the first of Meno’s pronouncements making up his paradox: “it’s not possible for someone to inquire … into that which he knows … for he wouldn’t inquire into that which he knows (for he knows it, and there’s no need for such a person to inquire)” (translated in Fine 2014: 7). We don’t need to make the strong claim that it is impossible for someone who knows or believes an answer to Q to inquire into Q. That this person would not be motivated to inquire into Q is sufficient to justify including (2) in our list. There are two parts to (3). First, the inquirer must believe that Q has a complete, true answer. Second, she must believe that inquiry into Q would lead her, or at least has a good shot at leading her, to believe Q’s true, complete answer.

The first part of (3) is reminiscent of Meno’s second pronouncement: “it’s not possible for someone to inquire into … that which he doesn’t know (for he doesn’t even know what he’ll inquire into)” (translated in Fine 2014: 7-8). On Gail Fine’s interpretation of Plato’s response to Meno’s paradox, Plato thinks that, in order to inquire into Q, one need not know the answer to Q, but must know enough to have “a target to aim at” or “to specify what it is that one wants to discover” (2014: 73). That is, one cannot be in a “cognitive blank” with respect to the question (71). My idea is similar: an inquirer must have a good enough grasp on a question Q that she believes that Q has a true, complete answer. Having a sufficiently good grasp on Q to believe that Q has a true, complete answer doesn’t, or at least doesn’t always, require grasping what Q’s possible answers are. For some questions, such as ‘Whether P?’ questions, there is what Friedman calls “semantic transparency from questions to answers”:

understanding the question requires having "a good sense of what the possible answers are" (2013: 159). But not all questions are like this. One can understand the question 'Why do beaches on Harbour Island have pink sand?' without having any idea of its candidate answers.

The second aspect of (2) is that the inquirer must believe that inquiring into Q will, or at least probably will, lead her to believe Q's complete true answer. If the inquirer didn't believe this, she'd think inquiry into Q would be a waste of time and resources. For example, an inquirer might believe that the question 'Are there an even number of stars?' has a complete true answer, while also believing that any inquiry that she, or anyone else, could undertake wouldn't get her to that answer. As such, she would not be motivated to inquire into this question.

The fourth property our ancestor would want before inquiring into Q is to be such that settling Q would advance her practical goals. The value of inquiry, thus the value of stopping and starting inquiry, in the state of nature is derivative from the practical value of true belief. As such, for an inquiry into Q to be worthwhile for our ancestor, settling Q must advance her practical goals to some degree.

We thus get the following picture of what the inquiry-starter concept would look like in the state of nature. The concept applies to some subject S regarding a question Q when:

PD-1: She has a questioning attitude to Q,

PD-2: S does not believe any complete answer to Q,

PD-3: S believes that the true, complete answer to Q is available to her through inquiry,

PD-4: Settling Q would advance S's practical goals.

Call this concept '*protodoubt'*, and say that an ancestor in the state picked out by the concept 'protodoubts Q'.

How similar is *protodoubt* to the intuitive concept *doubt* we have outlined? Recall that the concept *doubt* will typically apply to S regarding Q if:

D-1: S has a questioning attitude to Q,

D-2: S does not believe any complete answer to Q,

D-3: S's situation with respect to Q is epistemically risky, or is represented to S as epistemically risky,

D-4: S is immediately motivated to inquire into Q.

Protodoubt and *doubt* share the first two conditions. PD-4 seems relevant to D-4: if settling Q would advance S's practical goals, then S has prudential reason to inquire into Q that ought to bear on her motivation to inquire. But it is not clear what in *protodoubt* is relevant to D-3. Further, our intuitive concept *doubt* doesn't feature an analogue of PD-3. As such, I need to tell a story about why *protodoubt* would change, given practical pressures that are anticipatable from within the state of nature model, to look more like *doubt*. To do this, I will appeal to objectivisation.

Protodoubt applies when inquiring into Q would advance an ancestor's immediate practical goals. However it would be useful for our ancestors if they could inquire now, and store the results of their inquiries for their own use in the future, thus advancing their future practical goals as well. But they won't always be aware of what information they'll need for their future purposes. It would also be useful for our ancestors to inquire on behalf of each other: often, other community members will be better placed than oneself to inquire into Q; developing a practice of inquiring on behalf of each other will mean that one can get information that would not otherwise be accessible, or at least easily accessible, to one. But again, they won't always be aware of the information that others will require. Both of these considerations put pressure on PD-4, because what matters now is not just the advancement of S's current practical goals, but the potential advancement of any community member's practical goals, now or in the future.

Then objectivisation will change PD-4 so that it doesn't make reference to any individual's current practical goals. Rather, it would look something like this:

> **PD-4*:** Settling Q could reasonably be expected to advance some community member's practical goals.

PD-4* gives S prudential reason to inquire into Q. However, it gives S less prudential reason to inquire into Q than PD-4, assuming that S has more prudential reason to advance her own goals than she does to advance some community member's practical goals. (S is, of course, 'some community member'. So there is some community member for whom S has exactly the same prudential reason to advance their practical goals as she does to advance her own goals: namely, herself. But there are other ways that S can advance some community member's goals than by advancing her own goals: by advancing any other community member's goals. She has less prudential reason to do this than to advance her own practical goals.) So our objectivisation story has, at

this point, got us no further to *doubt*, and perhaps even further away. Moreover, we've still not seen why PD-3 would dissolve, nor why D-3 would arise, through objectivisation. So we need more from our objectivisation story.

2.3.2 Deriving intrinsic value from a diachronic model

What I need to do is to explain how we get from the condition PD-4*, 'Settling Q could reasonably be expected to advance some community member's practical goals', to D-3, 'S's situation with respect to Q is epistemically risky, or is represented to S as epistemically risky'. I've said that epistemic risk is the potential for some epistemically disvaluable event to obtain. Then part of what I need to do is to derive a purely epistemic reason for inquiring out of a prudential reason: failing to settle Q must be bad for epistemic reasons, rather than because it fails to advance someone or other's practical goals.

To this end, it will be helpful for me to look to Bernard Williams's (2002) genealogy of *truth*. Something unique to Williams's project, compared to Craig and Kelp's, is his endeavouring to explain why a community would come to value a conceptual practice of theirs intrinsically, even though their ancestors engaged in the practice for purely instrumental reasons. In my genealogy, I don't need to derive intrinsic from instrumental value as such, but I do need to derive epistemic from prudential value. *Doubt* is an epistemic concept, like *knowledge*. It is governed by, and bears on, epistemic normativity. The reasons it gives us – reasons to inquire, suspend, act, and so on – should end up as epistemic reasons. I hope that Williams's project can guide me in my own endeavour. Williams's genealogy is not of the semantic concept *truth*, which he, following Donald Davidson (1990), takes to be so central to our ability to form beliefs and communicate that we can't help but live by it. Rather, Williams aims to vindicate "the value of truth" for us, by which he means "the value of various states and activities associated with the truth (2002: 7). In particular, he is concerned with what he calls the "virtues of truth", which are "qualities of people that are displayed in wanting to know the truth, in finding it out, and in telling it to other people" (7).

Williams's genealogy begins in a state of nature, but unlike Craig's and Kelp's, it is not one in which our ancestors do not possess the concept of interest. Rather, Williams models the "primitive basis" of our actual situation with respect to *truth* (21-

2), which he calls the situation of "primitive openness and immediacy" (94). In this situation, a speaker utters a sentence that:

a. Describes her immediate environment,
b. Is true,
c. Can be clearly seen by its hearer to be true – it is "plainly true" (45).

In the primitive situation, we already see an "internal connection" between belief, assertion and fact (84): the speaker believes P because P is plainly true, and asserts that P because she believes that P. Following Davidson (1990), Williams holds that belief, assertion and fact must be aligned in order for linguistic understanding to get off the ground. This captures a thin sense in which we value *truth*: it constitutes a constitutive norm of our "belief-assertion-communication system" (2002: 84), that we understand beliefs and assertions *as* beliefs and assertions because we take them to be governed by the norm that they should be true.

However once this belief-assertion-communication system is established, the possibility of misalignment between belief, assertion and fact opens up. A deceitful speaker can assert that P without believing that P, to try to get her hearer either to believe that P or to believe that she, the speaker, believes that P (75). This opens up a gap between belief and assertion. And belief can be the result of "wishful thinking, or in other ways [can] become hostage to desire and wishes" (83). Then subjects may believe P for reasons other than that P is true, thus opening up a gap between truth and belief. The thin form of *truth* we get from the situation of primitive openness doesn't give our ancestors reason to align assertion to belief, or belief to truth, in any given case. A thicker form of *truth* is required.

Williams notes that true belief has survival value in the state of nature. This gives our ancestors reason to "do the best [they] can to acquire true beliefs" (11). Further, as Craig's state of nature story shows, our ancestors have an interest in using each other as informants. Some other member of the community might have a "purely positional advantage" over oneself with respect to some information (42); recall Fred up the tree, who has an advantage with respect to the whereabouts of the tiger to Mabel in the cave (Craig 1990: 11). These two facts generate a collective need for a "division of epistemic labour" among our ancestors (Williams 2002: 43), whereby information is shared and pooled, and each ancestor is a contributor to the communal pool of information.

Here, Williams's story diverges from Craig's. Rather than focusing on what makes a good informant in this scenario, Williams focuses on what qualities make some ancestor a good contributor to the pool. He suggests that this requires developing two bundles of dispositions, which close the gaps that arise between belief and truth on the one hand, and assertion and belief on the other. The former bundle he calls dispositions of "Accuracy", and the latter, dispositions of "Sincerity". Accuracy involves dispositions to use truth-acquiring methods in forming one's beliefs (133), and to resist self-deception and wishful thinking (125). Acquiring these dispositions requires the engagement of one's desires: one must "actually want to find out the truth" (133). Sincerity involves dispositions to "make sure that one's assertion expresses what one actually believes" (96), not only by resisting any temptation that may arise to lie (97), but by abiding by something like Grice's cooperative principle (99-100), so that one doesn't convey false or misleading information through implicature. Accuracy and Sincerity are what Williams calls the "virtues of truth" (7). Once our ancestors develop these virtues, they have a thicker form of *truth* than that found in the situation of primitive openness, which goes some way to closing the gaps between belief and truth, and assertion and belief.

However even this thicker form of *truth* has limitations. First, it is not clear that the value of Sincerity can be recognised "from an impersonal or general point of view" (58). Each ancestor has an interest in possessing Accuracy. But Sincerity seems primarily to be "a value for other people": though it is "obviously … useful for an individual to have the benefits of other people's information", it is "not useful to him that they should have the benefit of his", and sometimes "may well be a good idea for him to keep it to himself" (58). As such, a free-rider situation arises, whereby "each participant wants there to be a practice in which most of the others take part, without, if he can get away with it, taking part in it himself" (58). Second, the values of Accuracy and Sincerity are for this community purely instrumental: "they are entirely explained in terms of other goods, and in particular the value of getting what one wants, avoiding danger, mastering the environment, and so on" (58). This will set a limit even to each individual's interest in possessing Accuracy. For an "investigative investment" is required to get true belief: acquiring true beliefs has "a cost, in time, energy, opportunities lost, perhaps dangers run" (87). While Accuracy is valued only instrumentally, this value can be outweighed if such an investigative investment is too

costly. Putting these two limitations together, Williams concludes that "no society can get by … with a purely instrumental conception of the values of truth" (59).

To overcome these limitations, Williams holds that our ancestors must begin to value Accuracy and Sincerity intrinsically. Accuracy must be elaborated into "desire for truth… for its own sake", a "passion for getting it right" (126); and Sincerity into "dispositions to think that telling the truth (to the right people, on the right occasions) is in itself a good thing" (Williams 2014: 408). How would the community come to value Accuracy and Sincerity for their own sake? Williams appeals to social and psychological mechanisms that will inculcate in our ancestors certain habits of treating Accuracy and Sincerity as intrinsically valuable: our ancestors will be "discouraged or encouraged, sanctioned, shamed, or rewarded" into developing and manifesting the virtues of truth (2002: 44).

But that our ancestors treat Accuracy and Sincerity as intrinsically valuable is not enough for them to actually *be* intrinsically valuable. Rather, our ancestors must be able to make sense of Accuracy and Sincerity as intrinsically valuable "from the inside" (91): they require "some insight into these values, some account of their relations to other things which [they] know that [they] need and value, but an insight which does not reduce them to the merely instrumental" (90). That is, our ancestors must be able to relate Accuracy and Sincerity to their broader system of values, without thereby reducing their value to the extent to which they help our ancestors to attain other things they value. Further, Accuracy and Sincerity must come to engage our ancestors "ethical emotions" (92), such as shame, guilt, and regret (115). For Williams, it is sufficient for something's being intrinsically valuable that "first, it is necessary (or nearly necessary) for basic human purposes and needs that human beings should treat it as an intrinsic good; and second, they can coherently treat it as an intrinsic good" in the way just outlined (92).

In order to demonstrate the relationship of *truth* to our ancestors' broader framework of values, Williams starts to bring real history into his genealogy. Although *truth* is "[e]verywhere … related, psychologically, socially, and ethically, to some wider range of values", what those values are "varies from time to time and culture to culture". So at this point, armchair theorising will not show fully how *truth* relates to our broader system of values, and as such, "philosophy needs to make way for history" (93). The details of the historical stage of Williams's genealogy are not relevant for my purposes. Rather, I take this stage as schematic for the final stage in my own genealogy

of *doubt*. Though I don't need to derive intrinsic value from instrumental value in my model, I do need to show how prudential value can turn into epistemic value; and I will do this by showing how the concept that is developing in my model is related to our broader system of epistemic value.

2.3.3 Treating *protodoubt* as intrinsically valuable

Here's the state we're in with *protodoubt* in its current stage of objectivisation. *Protodoubt* applies to a subject S with respect to some question Q when the following conditions are met:

PD-1: S has a questioning attitude to Q,

PD-2: S does not believe any complete answer to Q,

PD-3: S believes that the true, complete answer to Q is available to her through inquiry.

PD-4*: Settling Q could reasonably be expected to advance some community member's practical goals.

The value of *protodoubt* is at this point prudential, and our ancestors treat *protodoubt* as valuable only instrumentally: the value of *protodoubt* is "entirely explained in terms of other goods, and in particular the value of getting what one wants, avoiding danger, mastering the environment, and so on" (58). As such, *protodoubt* will be unstable in the same way that *truth* is in William's genealogy before agents come to treat Accuracy and Sincerity as intrinsically valuable.

Recall Williams's idea that developing and manifesting Accuracy involves an "investigative investment" such that, before an agent values Accuracy for its own sake, she can ask questions like "How much trouble is it worth to find out about this?" (2002: 87). If the agent determines that finding out whether P isn't worth the investigative investment, she won't bother to try to believe the truth with respect to P. Hence agents need to come to regard Accuracy as worth striving for, for its own sake. Only then can *truth* begin to "carry its own weight on the balance of reasons", to use Queloz's phrase (2021: 168): agents will strive to believe the truth just because it is the truth. I want to say something similar, though not the same, for *doubt*. I need *doubt* to carry its own weight on the balance of reasons. That *doubt* applies to S should, in itself, mean that S has reason to inquire. I will borrow Williams's framework for understanding intrinsic value to show why our ancestors would come to understand *protodoubt* as generating its own reasons for inquiry.

Recall that, for Williams, it is sufficient for an item's being intrinsically valuable that, first, it is necessary or nearly necessary for "basic human purposes and needs" that humans treat it as an intrinsic good (2002: 92); and second, that they can coherently treat it as an intrinsic good. Humans coherently treat something as an intrinsic good when they understand it as importantly related to other things that they value, without reducing its value to the extent to which it helps them achieve these other things, and when it engages their ethical emotions. I don't need to commit myself to Williams's view that this is sufficient for something's being intrinsically valuable. But it is plausible that it is sufficient for some group G's treating some item X as intrinsically valuable that these conditions hold, and that is, I think, enough to get me where I need to be.

Protodoubt meets Williams's first condition: it is necessary, or nearly necessary, for basic human purposes and needs that human beings should treat it as an intrinsic good. For insofar as *protodoubt* is only treated as valuable insofar as it is instrumental to advancing some community member's practical goals (as in PD-4*), inquiring into Q will involve an investigative assessment whereby an ancestor will weigh up the prudential value of advancing some community member's practical goals against the prudential disvalue of losing her time and resources in the inquiry. But this will make the community's practice with *protodoubt* unstable in just the same way as Williams's community's practice with Accuracy is unstable before they come to treat Accuracy as intrinsically valuable. For if some agent doesn't care about advancing some community member's practical goals, she will quickly judge that the investigative investment for inquiring into Q is too great, thus will not inquire even though *protodoubt* applies to her with respect to Q. And on those occasions where she does care about advancing some community member's practical goals enough to undertake an inquiry, it isn't the application of *protodoubt* that motivates her, but her independent desire to advance a community member's practical goals. *Protodoubt* would then be, as Queloz writes (of *loyalty*, in a counterfactual situation in which we value *loyalty* only instrumentally), "mere window-dressing where it aligns with individual interest, and irrelevant where it does not" (2021: 56). In this case, *protodoubt* wouldn't be able to function as the inquiry-starter: our ancestors would inquire if and only if they had some independent desire to advance a community member's practical goals. Our need for an inquiry-starter, I have argued, is a basic human need. So insofar as our ancestors value *protodoubt* only instrumentally, some basic need of theirs goes unmet.

Protodoubt meets the first aspect of Williams's second condition for intrinsic value: our ancestors can relate the concept to other things that they value. Or at least, our ancestors can relate whatever concept functions as the inquiry-starter to other things they value. Our ancestors value true belief, they understand that inquiry is often necessary to get true belief, and as such they can relate the inquiry-starter concept to inquiry, thus to true belief, which they value. Does *protodoubt* engage their ethical emotions? That is, will agents feel ashamed, guilty, regretful, and so on, if they are such that *protodoubt* applies to them, but they don't inquire? Perhaps. It is plausible that our ancestors would develop practices of encouraging and discouraging, blaming and shaming each other to partake, to the best of each of their abilities, in inquiry, so that the results of inquiries can be stored and drawn upon by the wider epistemic community to the betterment of all. Insofar as these practices develop, our ancestors would naturally come to experience shame, guilt, and so on, when they fail to inquire when the inquiry-starter concept applies to them. Then it is plausible that *protodoubt* would engage our ancestors' ethical emotions. Then *protodoubt* seems to meet Williams's second condition for something's being intrinsically valuable, too – though recall that I only take these to be conditions for some group G's treating something as intrinsically valuable.

If our ancestors come to treat *protodoubt* as intrinsically valuable, it is no surprise that conditions PD-3 and PD-4* would drop out, and D-4 would arise. PD-3 is in *protodoubt* as something like a permissibility condition from practical rationality. Part of calculating the investigative investment of one's inquiry will be trying to determine whether one's inquiry would settle one's question; if it wouldn't, then inquiry is not worth the investigative investment; if it would, then inquiry may be worth the investigative investment. It is only if inquiry may be worth the investigative investment that it is permissible. So PD-3 partly determines whether one's inquiry is permissible, relative to practical rationality. But now our ancestors intrinsically value *protodoubt*, that the concept applies to S would, in itself, give S reason to inquire, such that an independent permissibility condition is not required: if S ought to inquire into Q, presumably she rationally may inquire into Q. PD-4* is a condition that bears on S's motivation to inquire. But now that our ancestors treat *protodoubt* as intrinsically valuable, it can generate its own reasons for action, such that the concept applying to S is itself enough to motivate S to inquire. Thus PD-3 and PD-4* drop out of the concept, and D-4 appears.

But what about D-3? Where does epistemic risk come into all this? Well, in the rough characterisation I gave in §2.2, epistemic risk is the potential for some epistemically disvaluable event to occur. Examples of epistemically disvaluable events are forming or maintaining a false belief, missing out on true belief, failing to know, and misunderstanding. As S doesn't believe any answer to Q, though she has Q in thought (PD-1), she is in a situation where she is at least at risk of missing out on true belief or of failing to know. Specifically, she is at risk of missing out on a true belief, or a knowledge-constituting belief, in the true, complete answer to Q. Then she is in an epistemically risky situation. So D-3 is true of her.

Thus, the concept at the end of the process of objectivisation would apply to some subject S with respect to some question Q when the following conditions are met:

D-1: S has a questioning attitude to Q,

D-2: S does not believe any complete answer to Q,

D-3: S's situation with respect to Q is epistemically risky, or is represented to her as epistemically risky,

D-4: S is immediately motivated to inquire into Q.

That is to say, my objectivisation story has shown why the concept *protodoubt* that meets the need to signal when inquiry should begin for our ancestors in the state of nature would change over time, given pressures anticipatable from within the model, to resemble our intuitive concept *doubt*. Thus I take my diachronic model of *doubt* to have confirmed my original hypothesis, that the concept *doubt* functions to signal when inquiry should begin.

3. Objections and replies

In the previous section, I tested my hypothesis that *doubt* functions to signal when inquiry should begin via two conceptual reverse-engineering projects, one that makes use of a synchronic model (§2.2) and one that makes use of a diachronic model (§2.3). Both projects confirm my hypothesis. We thus end up with the following picture of *doubt*. If a subject S doubts Q, then typically the following conditions will hold:

D-1: S has a questioning attitude to Q,

D-2: S does not believe any complete answer to Q,

D-3: S's situation with respect to Q is epistemically risky, or is represented to her as epistemically risky,

D-4: S is immediately motivated to inquire into Q.

These shouldn't be read as individually necessary and jointly sufficient conditions for *doubt*. Rather, they are a description of the conditions that hold when S doubts Q in typical cases, akin to Craig and Kelp's descriptions of the conditions that hold when S knows that P in typical cases (see Chapter 3, §§2.2-2.3). We can understand the concept *doubt* as picking out a questioning attitude that we call doubt. Typically, when S doubts Q, S won't believe any answer to Q, her situation will be epistemically risky, and she will be immediately motivated to inquire into Q. In this section, I consider some objections to this picture of *doubt*.

3.1. Why does the inquiry-starter have to be *doubt*?

I have said that we can understand doubt (the object of the concept *doubt*) as a questioning attitude. Jane Friedman offers a unified account of all questioning attitudes as species of suspended judgement. For Friedman (2017, 2019b), suspended judgement is the most general questioning attitude, in the same way that, for Timothy Williamson (2000), knowledge is the most general factive attitude. According to Williamson, if S has any factive attitude to P – if S sees that P, remembers that P, regrets that P, and so on – then S knows that P. All other factive attitudes can be thought of as *ways of knowing* that P. According to Friedman, if S has any questioning attitude towards Q – Friedman gives as examples of questioning attitude being curious about Q, wondering about Q, contemplating Q and deliberating Q (2019b: 299) – then one is suspended on Q.

For Friedman, suspended judgement is intimately related to inquiry. In her earlier work, Friedman holds that it is both necessary and sufficient for inquiring into Q that one suspends on Q (2017: 302). More recently, she's dialled back to just the necessity claim: "Having some IA [interrogative attitude] at *t* is necessary to count as an inquirer at *t*" (2019b: 300). On my account of *doubt*, the function of this concept is to signal that inquiry should begin. But why not think that the general concept *suspended judgement* plays the role of the inquiry-starter? I suspect that Friedman would be sympathetic to such a view. She compares her account of suspended judgement to Peirce on doubt, writing that, for Peirce, "doubt has a kind of suicidal tendency – it prompts us to inquire until it is extinguished. My proposal is that suspension shares this push towards its own demise: in suspending we ask a question and (at least in some minimal sense) seek an answer" (2017: 316). Further, as *knowledge* is the inquiry-stopper and picks out the most general factive mental state,

we would achieve a pleasing symmetry if *suspended judgement* were the inquiry-starter, given that it picks out the most general questioning attitude.

Response. I don't think that *suspended judgement* can play the role of the inquiry-starter. Recall Friedman's examples of questioning attitudes, which are, for her, all ways of suspending judgement on Q: being curious about Q, wondering about Q, contemplating Q, and deliberating Q. Note the variety in kind amongst these attitudes. Curiosity is an emotion, and as such is a passive state: one doesn't have to *do* anything in order to be curious. Contemplation and deliberation, on the other hand, are active: one must do something in order to contemplate or deliberate Q, even if this action is only in one's thought. Wondering seems somewhere between the two. It has an affective aspect, like an emotion, but in order to wonder about Q it is plausible that you have to be doing something – wondering about Q is an activity in which you engage, rather than a state that you instantiate.

Friedman is right to say there is "a minimal sense" in which someone who has any of these questioning attitudes towards Q "seek[s] an answer" to Q (2017: 314). But this really is a minimal sense of 'seeking an answer' to Q. One who is curious about Q needn't take any steps to try to settle Q. Someone who seeks an answer to Q without ever taking steps to settle Q is like someone who seeks a job without ever putting in an application, or seeks a partner without ever asking anyone on a date. Though the job, or the partner, might fall into their lap (so to speak), this isn't likely and certainly shouldn't be counted on. (Another worry for Friedman here is that it is hard to hear 'seeking' in these cases as meaning anything other than 'desiring', but Friedman is very insistent that questioning attitudes should not be understood as metacognitive desires, such as desires to know or desires to have true belief. This ensures that creatures who are not plausibly capable of metacognition could still have questioning attitudes. See Friedman 2013: 155; cf. Carruthers 2018.) A less minimal sense of 'seeking' some goal has it that one must take steps towards achieving that goal in order to count as seeking it.

These more and less minimal senses of 'seeking' correspond to more and less minimal senses of 'inquiry'. One who is curious about Q but doesn't take any steps towards settling Q isn't inquiring in what Friedman calls "the most active sense of 'inquiring'," which involves "acting so as to advance her goal of closing Q" (315). But there is a less demanding sense of 'inquiring' on which being curious about Q is sufficient for inquiring into Q: someone who is curious about Q "roughly wants to know

Q, will be sensitive to information that bears on Q, might act to get information that will help her to close Q when she can and so on" (315).

However, it is only the more active sense of 'inquiring' for which we require an inquiry-starter concept. Recall from §2.1 the three facts from which our need for an inquiry-starter arises:

1. Truth matters: we generally prudentially benefit from treating a proposition as true iff it is true.
2. We have to inquire in order to get true beliefs, but inquiry is costly in terms of time and cognitive resources.
3. Inquiry has no natural starting-point: there are limitless questions into which we could inquire.

It is the active sense of 'inquiry' and its cognates that these facts invoke. We must *actively* inquire in order to get true beliefs; simply being curious about Q won't get us to Q's answer. It is only *active* inquiry that is costly in terms of time and cognitive resources; simply being curious about Q doesn't require that one does anything, spends any time, or uses any cognitive resources in finding out Q. Thus we need a concept that flags when the benefits of active inquiry are worth these trade-offs. *Suspended judgement* isn't up to the task, as it is too general, encompassing modes of suspension that do not immediately motivate active inquiry.

3.2. Do all inquiring creatures doubt?

On a functional view of belief, articulated but not endorsed by Stephen Stich (1979), non-human animals have beliefs about the world. On this view, we attribute belief that P to some creature when this best explains the creature's behaviour. Often, the best explanation for a non-human animal's behaviour attributes belief to the animal, usually in conjunction with an attribution of desire. Consider: we say that the dog ran to his dish because he *desired* to eat and *believed* that his master just put a meaty bone in the dish (1979: 16; cf. Armstrong 1973: 25). If we endorse this picture of belief, we will think, too, that non-human animals can have true or false beliefs. Just as true beliefs have survival value for humans, so too do they have survival value for non-human animals. But just as true belief does not often fall into the laps of humans, neither does it often fall into the laps of non-human animals. Rather, animals often have to *do something* in order to find out the answer to Q. As Peter Carruthers notes, an animal confronted with some unusual object "might look closer at it, move up to

sniff it, walk around it to examine it from the other side, and so on" (2018: 133). We can understand this animal's behaviour as attempting to find out about the object, and, if all goes well, forming a true belief about the kind of object it is – for example, is it a living creature? Is it edible? Is it poisonous?

Given how I've been characterising inquiry throughout this thesis, this looks very much like inquiry, and indeed like inquiry in the 'active' sense discussed in the previous section. If this is so, then non-human animals are capable of inquiry. But are non-human animals capable of doubt? To doubt, one might think, is a metacognitive attitude: one can doubt only if one can think about one's cognitive situation. For example, Christopher Hookway suggests that doubting some proposition involves "becom[ing] anxious about any tendency to accept it that I still possess; I shall also become anxious about any other beliefs I hold which may depend on it" (1998: 218), and as such, he understands doubt as a kind of anxiety, which he calls "epistemic anxiety" (222). But for many animals that we might think of as capable of inquiry – for example, dogs – we might not find it plausible that they are capable of metacognition (see Carruthers 2018: 133). But if some creatures can inquire without being capable of doubting, how can *doubt* be the inquiry-starter?

Response. Hookway is interested in doubt as a propositional attitude, whereas I am interested in doubt as a questioning attitude: Hookway is giving an account of what it is to doubt some proposition P, whereas my account is of what it is to doubt a question Q. Hookway's account of doubt (the propositional attitude) is metacognitive: to doubt P, for Hookway, is to be anxious about one's cognitive relationship to P; for example, if S doubts that P, she will be "anxious about any tendency to accept it that [she] still possess[es]" (218), and anxious "about any inquiry that relies upon a doubted proposition" (221). But on my account of doubt as a questioning attitude, doubt is not metacognitive: it is not an attitude about one's cognitive situation with respect to a question Q. Rather, doubt is a first-order questioning attitude: it is an attitude towards Q itself. Then that an animal is not capable of metacognition does not mean that they could not have doubt, the questioning attitude.

Carruthers, along with Dennis Whitcomb (2010) and Friedman (2013), take it to be advantages of their accounts of curiosity, wondering, and other questioning attitudes that they are first-order attitudes that could be possessed by animals incapable of metacognition. I can similarly claim this advantage for my account of doubt. Though it might sound strange to say, for example, that Fido doubts whether

we're going to the park or to the vets, insofar as an animal's behaviour is best explained by its doubting Q, my account of doubt as a questioning attitude can make sense of that animal having a doubt, even if it is not plausibly capable of metacognition.

But another thing to say in response to this worry is that, on my account, it is the concept *doubt* that plays the role of the inquiry-starter, not the questioning attitude doubt. The inquiry-starter role played by *doubt* is that of signalling that inquiry ought to begin. It is no part of my account of *doubt* that one cannot inquire into Q without doubting Q. All I've said about what is required in order to be inquiring into Q, in the sense I'm interested in ('active inquiry', recall from §3.1), is that one needs to take steps to settle Q. One needn't doubt Q in order to do this.

3.3. Disanalogy between the roles of *knowledge* and *doubt* in inquiry

Alessandra Tanesini raised the following objection to me in conversation. There is something genuinely defective about an inquiry that doesn't end in knowledge. This explains why the concept *knowledge* can play the role of the inquiry-stopper: because an inquiry into Q is not defective only if it ends with the inquirer's knowing Q's answer, we can apply the concept *knowledge* to signal that inquiry should now come to an end. But there is nothing defective about an inquiry that doesn't start with doubt. Then why should applying the concept *doubt* signal that inquiry should begin? That is, why should *doubt* be able to play the role of the inquiry-starter?

To make this vivid, consider some examples. The following case is from Kelp:

> *The Hire.* You are a private detective. I have contracted you to find out whether (MURDER =) someone in my family is the murderer of my wife. The first suspect you investigate is my uncle who, let us suppose, has a particularly strong motive for the deed. Fortunately for you, my uncle credibly admits to having committed the crime upon questioning and even signs a confession in writing. On the basis of this evidence, you come to believe that MURDER, pack your bags, including the confession, and get on the next flight to the Caribbean where you intend to take a holiday for the rest of the month. Meanwhile, it becomes widely known that my uncle's claim to have murdered my wife is false. In fact, he has a watertight alibi for the time of the deed and was protecting the perpetrator. At the same time, MURDER is true: someone in my family did indeed murder my wife. It is just that it wasn't my uncle but, say, my cousin. You are currently

> sipping cocktails in the sun and are entirely unaware of the news about my uncle's confession. (2021b: 361)

In this case, you ought not have terminated your inquiry into the question whether MURDER. This is despite your having a justified true belief to a complete answer to this question (namely, that MURDER). Kelp argues, and Tanesini would agree, that this is precisely because you don't know this answer. So The Hire supports Tanesini's claim that, in order to be non-defective, an inquiry must end in knowledge. But compare The Hire to the following case:

> *Mitchell Brothers*. You wonder who is the older of the Mitchell brothers: Phil or Grant. This question doesn't strike you as in need of an answer, and you briefly deliberate about whether or not to inquire into it. Both inquiring and not inquiring seem fine choices – you have nothing to do for the next 5 minutes – so you Google the question. You learn that Phil is the older Mitchell brother.

There is nothing at all defective about your inquiry. But it doesn't begin with doubt, but with wonder. Then it isn't required for an inquiry to be non-defective that it begins with doubt.

Response. I said that the function of the concept *doubt* is to signal that inquiry s*hould* begin, not that inquiry *may* begin. It is compatible with this that there are non-defective inquiries that begin without doubt. It can be true that if the concept *doubt* applies to S with respect to some question Q, then S ought to inquire into Q, without it being true that if *doubt* does not apply to S with respect to Q, then S ought not to inquire into Q. Compare: it might be true that if you are wealthy, you ought to give money to charity. This doesn't entail that if you are not wealthy, you ought not give money to charity.

A worry one might have with this response is that I motivate our need for an inquiry-starter concept by appealing to the idea that inquiry has no natural starting-point (§2.1). By this, I mean that we can inquire into pretty much any question we can conceive of. But given that inquiry is costly in terms of time and resources, we ought not inquire into just any question. Rather, we should ration our time and resources to just those questions that matter, for one reason or another. This suggests that the role of the inquiry-starter concept isn't just to signal when inquiry should begin, but also when it shouldn't begin. That is, instead of licensing the conditional instruction, 'If the concept applies, then begin inquiry', it should license the biconditional instruction:

'Begin inquiry if and only if the concept applies'. That is, S should begin inquiry into Q if and only if S satisfies *doubt* with respect to Q.

However this worry is misguided. I do motivate the need for an inquiry-starter concept in this way. But this doesn't mean that the inquiry-starter concept needs to license the biconditional instruction: 'Begin inquiry if and only if the concept applies'. *Doubt* can serve the function of the inquiry-starter if the following norms hold:

> **Permitted to Inquire if Questioning (PIQ):** S is permitted to inquire into a question Q if S has some questioning attitude to Q.
>
> **Required to Inquire if Doubting (RID):** S is required to inquire into a question Q if S doubts Q.

This will be enough to ensure that subjects are normatively bound to ration their time and cognitive resources in the right way. When a subject doubts Q, then she ought to settle Q. This might require abandoning, or at least postponing, other inquiries she is currently undertaking; or not inquiring into questions for which she has some questioning attitude other than doubt. But this wouldn't mean violating PIQ. One abides by a permissive norm – a norm that says one is permitted, but not required, to φ – when one doesn't φ. So in a situation in which a subject S doubts a question Q1, and is curious about, but doesn't doubt, a distinct question Q2, the set of norms PIQ and RID would tell her that she ought to inquire into Q1 rather than Q2. Thus we can see that, if RID is a norm, this will be enough to ensure that subjects are normatively bound to ration their time and cognitive resources to those questions that matter, and if PIQ is a norm, then there can be non-defective inquiries that do not begin with doubt, and these two norms are compatible.

I take myself to have established that it needn't be the case, on my picture of *doubt* as the inquiry-starter, that an inquiry that begins without doubt is defective. Nevertheless, I want to make a more general point about the consequences of violating norms that apply to subjects at the point at which they first take some questioning attitude, such as doubt, to a question. This is that, given a natural way of understanding what it is to start an inquiry, these norms will not be constitutive norms of inquiry. That is, they will not be norms that are such that inquiry is the kind of activity it is because it is governed by these norms. As such, a subject's violating these norms, whatever they might be, needn't make any of her subsequent inquiries defective. In contrast, if a properly settled inquiry is one that ends in knowledge, then it will be a constitutive

norm of inquiry that it should end in knowledge. That is, that this norm applies to the activity of inquiry is partly constitutive of what it is for that activity to be the activity of inquiry, rather than some other activity. Then if this norm is violated, the inquiry is defective *qua* the kind of activity it is. That is, it is defective *qua* inquiry. Therefore, we should expect there to be a disanalogy between the normative profiles of the concepts that act as inquiry-stopper and inquiry-starter: norms involving the inquiry-stopper will be constitutive norms of inquiry, while norms involving the inquiry-starter will not be; as such, violating the former will make one's inquiry defective *qua* inquiry, while violating the latter will not.

Let me unpack this. As Kelp argues, many activities we engage in are activities with constitutive aims or norms, or "ACANs" (2021b: 366). What this means is that the aims and norms that govern an ACAN are essential to it: "Anything that does not have these aims and norms will not qualify as a token of this ACAN" (366). Many, if not most, if not all, games are like this. Kelp offers chess as an example. The constitutive aim of chess is to checkmate one's opponent. Among the constitutive norms of chess are norms that specify the starting points of the chess pieces, as well as what are moves in chess. Failing to abide by these norms means failing to play chess: throwing chess pieces around willy-nilly does not qualify as playing chess, for example.

Kelp argues that inquiry, like chess, is an ACAN. The aim of Q-inquiry is to settle the question Q. There are different views about what it is to settle Q, but for the sake of argument, let's assume (with Tanesini and Kelp) that S settles Q iff S knows the answer to Q. Then the aim of Q-inquiry is to know the answer to Q. This aim is partly constitutive of what inquiry is. If one were engaged in an activity with a different aim, one wouldn't be engaged in the activity of inquiry, just like if one were engaged in an activity that didn't have checkmating one's opponent as its aim, one wouldn't be engaged in the activity of playing chess. Thus, the constitutive aim of Q-inquiry is to know the answer to Q.

Unlike for the earlier Friedman, it isn't sufficient to be inquiring into Q, on this picture, that one has a questioning attitude to Q. It is necessary for being engaged in the activity of inquiry that one is engaged in an activity that has knowing Q's answer as its aim. If one is curious about Q, but has resolved to avoid finding out the answer to Q, then one is not engaged in an activity that has knowing Q's answer as its aim, so is not engaged in inquiry. For example, I am curious about whether my ex-partner is in a relationship. However, I think that finding this out wouldn't do me any good, and might

make me sad. So I resolve to avoid finding out the answer to my question: I block my ex on social media, I ask my friends not to update me on their life, and so on. Nevertheless, I remain curious about whether they are in a relationship. On the earlier Friedman's picture, I am inquiring into whether my ex is in a relationship. On Kelp's picture, I'm not: I'm not engaged in an activity aimed at coming to know the answer to this question, and therefore I am not engaged in the activity of inquiry.

The constitutive aim of inquiry generates norms which are themselves partly constitutive of what it is to be engaged in the activity of inquiry. There will be norms that specify what count as "moves" in inquiry, for example belief will be specified as "the type of move that closes inquiry into whether p for one in the affirmative or negative" (368). But the only way to "properly close" inquiry is to come to know Q's answer (360). As such, the only *permissible* closing move in inquiry is to know Q's answer. Then an inquiry that terminates in anything other than knowledge is defective in that a constitutive norm of inquiry has been violated. As such, the activity is defective *qua* the kind of activity it is. It is defective *qua* inquiry. The same isn't true, I will now argue, for an inquiry that begins after a subject has violated some norm like PIQ or RID.

There are a few candidate ways of specifying the 'opening' move in inquiry. The first is to say that the opening move in inquiry is to take some questioning attitude to Q. Of course, this won't be sufficient for starting inquiry. In the case just considered, I am curious whether my ex-partner is in a new relationship, but I never embark on an inquiry into that question. But this doesn't show that taking a questioning attitude to Q cannot be the opening move in inquiry. Compare: the opening moves in chess are to either move some pawn one or two squares forward on the board, or to move a knight two squares forward and one to the side. But it isn't sufficient for starting a game of chess that one moves one of these pieces in one of these ways. For one might be doing something other than beginning a chess game by so moving the piece, for example, practising for a future game. Alternatively, we could say that the opening move in inquiry is to decide or intend to settle Q. Then we wouldn't have to demarcate cases where S has a questioning attitude to Q and thereby starts inquiry from those cases, like the ex-partner case, where S has a questioning attitude to Q without starting inquiry into Q. A third option is to say that the opening move in inquiry is to take the first step towards settling Q: to search for one's first piece of evidence, say, or to survey one's memory for Q's answer for the first time.

This third option seems to me the most plausible for inquiry in the 'active' sense in which I am interested. In contrast, the first way is more plausible for inquiry in Friedman's weak sense. One reason to prefer the third to the second option for specifying the opening move in (active) inquiry is that it makes better sense of cases like the following:

> *Forgetful Freya.* While Freya is at the market, she sees on sale a kind of cheese she's never heard of, called Neufchâtel. She is curious where this cheese is made, and decides to look this up when she gets home. But a few minutes later, she has forgotten all about this cheese, and her decision to look it up.
>
> *Unfortunate Ursula.* Ursula is on a walking holiday along the Dover coast. One afternoon, while she is walking the White Cliffs of Dover, she sees a beautiful bird, the size of a sparrow, greyish in colour with brown and buff mottling and a dark band running down its back. She wonders what kind of bird it is. She knows very little about birds, but her grandfather is an avid birdwatcher. She forms the intention to phone him when she gets back to her hotel to ask him to identify the bird from her description. Unfortunately, while distracted by the beautiful bird, she steps off the cliff edge and falls to her death.

I think that neither Freya nor Ursula begins inquiring into their respective questions, 'Where is Neufchâtel made?' and 'What kind of bird is this?' This is so even though they decide and intend, respectively, to try to settle their questions. (This is also reason to prefer the third way of specifying the opening move of inquiry to the first: Freya and Ursula both have questioning attitudes to their questions.)

What matters for my purposes is that, on either the second or third ways of specifying the opening move in inquiry, having a questioning attitude to Q will come prior to the start of inquiry into Q. Any norms that govern what happens prior to the start of the activity of inquiry cannot be constitutive norms of the activity of inquiry: they can't be norms such that, if one is not engaged in an activity governed by those norms, then one isn't engaged in the relevant activity, as they are explicitly norms that apply to one *before one is engaged in* that activity, and therefore *when one is not engaged in the activity*. So whatever are the norms that govern what kind of questioning attitude one must have to Q in order for any ensuing inquiry into Q to be permissible or required, be they PIQ and RID, or some other norms, these norms will not be constitutive norms

of inquiry. As such, violating these norms won't make any ensuing inquiry defective *qua* the kind of activity it is – that is, it won't make that inquiry defective *qua* inquiry.

To make this clearer, let's return to the analogy with chess. It isn't a constitutive norm of chess that play starts only when a buzzer is sounded. But suppose that Magnus and Ian are in a chess competition and are supposed to begin their match only when the buzzer sounds. Suppose that the buzzer doesn't sound when it should, because it's broken; but at the time at which the buzzer was supposed to sound, an audience member's phone vibrates, and everyone mistakes it for the buzzer. Magnus and Ian begin their chess match. In doing so, they violate a norm for participating in the chess competition: begin your match only when the buzzer sounds. But they don't violate any constitutive norms of chess. Then their violating this norm about when to begin their game of chess does not make the ensuing match defective *qua* chess match.

Note that this is entirely compatible with the concept *doubt* functioning as the inquiry-starter, given what I mean by this term. The sense in which *doubt* is the 'inquiry-starter', for me, is not that application of the concept starts inquiry, nor that being in the state picked out by the concept is to begin inquiry. Rather, application of the concept signals that inquiry *ought* to start. It is perfectly compatible with this that inquiry does not start with doubt (so with the proper application of *doubt*), but that doubt (so the application of *doubt*) comes prior to the start of inquiry.

3.4. What kind of normativity governs the start of inquiry?

I have said that any norms that govern the start of inquiry are not constitutive norms of inquiry. Then what are they? In particular, if PIQ and RID from the previous section are norms, from what domain of normativity do they derive their force? The default view in epistemology says any reasons we have to inquire lie outwith the purview of epistemic normativity; rather, reasons to inquire are moral or prudential.[12] For

[12] A notable opponent of the default view is Gilbert Harman, whose principles for theoretical reasoning include prescriptive norms for when we may and ought to inquire, such as the following "interest condition": "One is to add a new proposition P to one's beliefs only if one is interested in whether P is true" (1986: 55). This is a permissibility condition on the activity of *coming to form* a belief, rather than an evaluative norm about a belief already formed. This is not intended as an instrumental norm aimed at maximising the prudential value of belief, but as a norm for how to reason well. Whether one reasons well is, for Harman, essentially tied to whether one ends up with knowledge or justified belief (48). So it

example, Richard Feldman writes that although it is "surely true that there are times when one would be best off finding new evidence … this always turns on what options one has, what one cares about, and other non-epistemic factors" and "these are prudential or moral matters, not strictly epistemic matters" (2000: 689). More generally, "questions of how to conduct inquiry over periods of time … are moral or prudential questions rather than epistemic questions" (689). Many epistemologists endorse this more general view that epistemic normativity includes only norms on synchronic states, not diachronic actions (Foley 1987, Feldman 2000, Dougherty 2014, Hedden 2015). On this picture, any norms governing the diachronic activity of inquiry would necessarily lie outside the bounds of epistemic normativity.

Friedman pushes back against the default view, arguing that a construal of epistemic normativity that excludes norms of inquiry from its purview is "parochial" (2020: 527). For inquiry is an activity aimed at achieving epistemic ends, such as knowledge. Norms of inquiry are norms that tell us "how to properly engage in the activity so that we end up in the sort of epistemic state we want or need to end up in – they tell us how to come to have knowledge" (527). As such, there is an "important sense in which the norms that guide and constrain us in our efforts to acquire knowledge … are epistemic" (527). To deny this is to place undue importance on the exact moment of belief formation, ignoring the process of evidence-gathering, reasoning, and so on, that led up to it. Further, this picture of epistemic normativity can only tell subjects what they should do with information they have, not "how to get and manage the information they want and need" (527). But why should epistemology care only about what we do with information that we happen to get, but not about whether and how we get the information we need and want? This pushes Friedman towards an "expansive picture of the epistemic", on which epistemic normativity includes not just norms determining what makes a belief good (in the attributive sense of 'good'), but also norms about how to inquire.

So Friedman attempts to articulate some such norms of inquiry. Her first offering is the following, which she calls the "zetetic instrumental principle" or ZIP:

---

is reasonable to understand Harman is as positing principles for theoretical reasoning as deriving their normative force from the domain of epistemic normativity.

> **ZIP** If one wants to figure out [a question] *Q?*, then one ought to take the necessary means to figuring out *Q?* (2020: 503).

However Friedman argues that a problem arises when we try to incorporate ZIP into the realm of epistemic normativity: ZIP conflicts with widely accepted, "traditional" epistemic norms. In particular, ZIP conflicts with the following:

> **KP** If one is in a position to know *p* at [a time] *t*, then one is permitted to come to know *p* at *t* (504).

She offers a case to demonstrate this conflict. I need to find out how many windows the Chrysler Building has (I'm a window cleaner who gets paid by the window). Suppose that the best way for me to find this out is to count the windows myself. During the time in which I do my counting, say 14:00-15:00, I am in a position to know all kinds of things: about the people who walk past me, what they're doing, what they're wearing; about the logical consequences of things I already know; and so on. But, given my goal of finding out how many windows the Chrysler Building has, I better not do any of those things. I must stay focused on my inquiry, which requires ignoring everything else going on around me. As Friedman puts it, I will succeed in my inquiry only "by failing to respect my evidence for some stretch of time" (503). But then ZIP and KP stand in conflict. ZIP tells me that I mustn't come to know various things that I am in a position to know (for doing so will prevent me from figuring out my question). But KP tells me that I am always permitted to come to know that which I am in a position to know.

Insofar as we think one and the same domain of normativity cannot issue contradictory guidance, this is reason to think that one of ZIP or KP cannot be an epistemic norm. One might then think that, as KP is the more 'traditional', so less controversial and more widely endorsed, of the two, we'd likely be better off excluding ZIP from the domain of epistemic normativity (though this is not Friedman's own view).

In response to this worry, recall from the previous section that an agent can abide by a permissive norm according to which one is permitted to φ in circumstances C by either φ-ing in C or failing to φ in C. Then in the Chrysler Building case, ZIP and KP do not issue incompatible instructions about what I should do. KP says that we *may*, not that we *must*, come to know that P when we are in a position to know that P; we abide by this norm when we either come to know those propositions that we are in positions to know, or we fail to come to know those propositions. ZIP says that, if we want to find out the answer to Q, then we must take the necessary steps to figuring out

Q. We abide by this norm by being such that, whenever we want to find out the answer to some question Q, we take the necessary steps to figuring out Q.[13] In the Chrysler Building case, I abide by both KP and ZIP when I do what is necessary to figure out how many windows the Chrysler Building has, which includes not coming to know many propositions that I am in a position to know. So the 'tension' between KP and ZIP doesn't amount to their issuing incompatible instructions about how to act in this, or any, case; and this worry needn't undermine the justification for Friedman's move to a more expansive picture of epistemic normativity, one which includes norms of inquiry.

But how justified is Friedman's move, anyway? Her comments about 'parochialism' are suggestive, but not entirely convincing: even if the goal of inquiry is to be in some epistemic state, such as knowledge, this doesn't mean that any norms governing the activity of inquiry must likewise be epistemic. Recall that Friedman writes that norms of inquiry will tell us "how to properly engage in the activity [of inquiry] so that we end up in the sort of epistemic state we want or need to end up in"; they are "norms that guide and constrain us in our efforts to acquire knowledge" (2020: 527). Friedman argues that whether we conform to norms of inquiry, such as ZIP, is "highly relevant to whether or not we come to know what we want to know", and as such, "ZIP is a norm that a rational subject trying to know more and understand better will conform to" (511). It may be true that subjects who want to achieve knowledge will conform to ZIP. But this doesn't seem sufficient to make a norm epistemic. As David Thorstad notes, abiding by the following norm might help one to achieve knowledge:

> Sandwich: If you are inquiring for many hours, you ought to pause and eat a sandwich. (2022: 410)

This is because, as Nomi Arpaly (2017) points out (attributing the idea to Sophie Horowitz), eating a sandwich raises blood sugar, and having enough blood sugar improves one's cognitive capacities. Then subjects who want to achieve knowledge have reason to conform to Sandwich. But we don't want to say that Sandwich is an epistemic norm. So it isn't sufficient for a norm's being epistemic that abiding by that norm helps one to achieve one's epistemic goals.

---

[13] Friedman intends ZIP to be read as having narrow scope (2020: 508). A wide-scope equivalent to ZIP would say that we ought to be such that: if we want to figure out Q, then we take the necessary steps to figuring out Q. We can abide by the wide-scope equivalent of ZIP by either taking the necessary steps to figuring out Q, or by dropping our desire to figure out Q.

What would be sufficient for a norm's being epistemic? Carolina Flores and Elise Woodard (forthcoming) have recently suggested the following test for a norm's being epistemic: if we criticise someone for failing to abide by the norm, and this criticism takes a distinctively epistemic form, then we have good reason to think that the norm is epistemic. They use this test to argue that there are epistemic norms on evidence-gathering. Gathering evidence is one way that we inquire, so if Flores and Woodard can show that there are epistemic norms on evidence-gathering, they will have shown that there are at least some epistemic norms on inquiry. This won't yet show that PIQ or RID are epistemic norms on inquiry, but this would at least give us a framework for making an argument to that conclusion.

Flores and Woodard's argument runs as follows.

1. If there is a legitimate practice of epistemically criticising agents for whether and how they φ, then we have reason to think that there are epistemic norms on φ-ing.
2. There is a legitimate practice of epistemically criticising agents for their evidence-gathering.
3. Therefore, we have reason to think there are epistemic norms on evidence-gathering. (forthcoming: 11)

In support of (1), Flores and Woodard appeal to the roles that norms play in our lives. One such role is that of setting standards for assessing other agents and their actions. When an agent violates a norm, we assess them negatively, at least if they don't have any excuse for violating the norm (for example, that they reasonably believed themselves to be following the norm; see DeRose 2002: 180). Then a good indicator that there exists a norm telling agents to φ in circumstances C is that we criticise agents for failing to φ in C. For example, evidentialists take the following to be a norm: believe P iff your evidence supports that P. It is evidence that this norm exists that we criticise agents for failing to abide by it; for example, if S believes that P when P is not supported by her evidence, we label her irrational (Flores and Woodard forthcoming: 12). Further, the way that we criticise agents for breaking a norm indicates the kind of normativity at work in the norm. To criticise an agent as irrational is to criticise them *qua* epistemic agent. It would be inappropriate to morally criticise someone for believing against their evidence. As Flores and Woodard put it, "the type of criticism match[es] its target" (12).

There are many ways we criticise each other, or otherwise hold each other accountable, that are distinctly epistemic. Alongside labelling people as 'irrational',

'ignorant', or as exhibiting some other epistemic failing, we might "explicitly encourag[e] compliance with the norms" (12); for example, we might instruct some agent who has failed to believe in accordance with her evidence that she must so believe in the future. Further, some epistemologists have argued that we engage in distinctly epistemic practices of blaming, and blaming someone for φ-ing, or for failing to φ, can be understood as a way of criticising or otherwise holding them accountable. Cameron Boult argues that we epistemically blame one another for epistemic failings by "suspend[ing] a presumption of epistemic trust, at least within some restricted domain, or on some specific matter", where epistemic trust in some agent is "confidence that they are a reliable source of information" (2020: 525). Antti Kauppinen also argues that we hold agents epistemically accountable by reducing epistemic trust in them, where reducing epistemic trust in some agent consists in "giv[ing] a lower credence to the rest of what he says, [being] less willing to regard him as a potential partner in co-operative inquiry, and [having] less attitudinal confidence in him" (2018: 1). As Flores and Woodard note, it won't be sufficient for epistemically criticising someone that we reduce epistemic trust in them: we might reduce epistemic trust in someone "simply because we have more information than them" (forthcoming: 13). Rather, to count as epistemic criticism, the reduction of epistemic trust must be accompanied by a judgement of blameworthiness (see also Boult 2020: 519).

In support of their premise (2), Flores and Woodard offer cases in which we reduce epistemic trust in an agent, and judge them to be blameworthy, for whether and how they gather evidence. We criticise agents in epistemic bubbles, where an epistemic bubble is a "social epistemic structure which has inadequate coverage through a process of exclusion by omission" (Nguyen 2020: 143). People who get all their news from a single news channel, or who form all their beliefs about some topic from one Facebook group, for example, are in epistemic bubbles. Flores and Woodard construct a case that demonstrates the phenomenon:

> *Cloistered Claire:* Claire gets all of her nutrition news from Guup, which tends to endorse fad diets that are not always scientifically backed. For example, this month, it encourages its readers to add 1 tbsp of coconut oil to their coffee each day. As it turns out, this is actually a scientifically backed suggestion, but Guup does not offer good evidence for it. Claire believes Guup's claim, and she feels no need to check additional sources. (forthcoming: 16)

Flores and Woodard hold that Claire is criticisable for failing to gather more evidence before forming her belief that she should add 1 tbsp of coconut oil to her coffee. Further, the ways that we would criticise her are distinctly epistemic. Upon finding out that she gets all her nutritional information from Guup, we would lower our trust in her on issues of nutrition; we might also explicitly recommend that she gets her nutritional information from more diverse sources in the future. This suggests to Flores and Woodard that Claire has violated an epistemic norm, specifically a norm that tells subjects to gather evidence from diverse sources before forming certain beliefs.

Flores and Woodard also argue that we criticise subjects who are "lazy in their evidence-gathering", in that, when they are offered additional relevant evidence, they don't take it (17). They offer one such case:

> *Lazy Larry:* Larry is a chemistry major, who forms his beliefs about the structure of the atomic nucleus based on over-simplifying and idealizing diagrams, depicting electrons as marble-like entities that orbit the nucleus in precise tracks. However, this is misleading: electrons actually are spread out diffusely within a massive region. The textbook includes this information, but Larry limits his efforts to just looking at the pictures. (17)

Larry is behaving badly, epistemically speaking: he is being "negligent" (17). There is information readily available to him that is such that, had he gathered this information, he would have known that electrons are massively diffused. He is thus criticisable. But it may well be that the belief he has formed is justified, given the evidence he has: the diagrams. Thus Flores and Woodard hold that he is criticisable not because he has failed to respond appropriately to his evidence, but "for his failure to adequately gather evidence bearing on his beliefs" (17). Further, the kind of criticism we would level at Larry is distinctly epistemic: if we were classmates of Larry's, we would cease to uptake any testimony he offered on the structure of the atom, or about chemistry in general. That is, we would lower our epistemic trust in him, at least relative to the domain of chemistry, upon learning that he is lazy in gathering evidence in this way.

Flores and Woodard have given us reason to think there are some epistemic norms on evidence-gathering. As evidence-gathering is one way of inquiring, they have given us reason to think that there are epistemic norms on inquiry. But they have not given us reason to think that either PIQ or RID from the previous section are norms. I think, however, that subjects who violate RID are epistemically criticisable in much the

same way that Lazy Larry is. (PIQ is a permissive norm, so it cannot be violated: an agent abides by a norm that says they are permitted to φ in circumstances C by either φ-ing in C or not φ-ing in C.) Recall that RID says that S is required to inquire into Q if S doubts Q. Now consider the following case:

> *Avoidant Avery*. Avery's girlfriend is a video game journalist. Avery loves to read her articles, and learns a lot about video games in the process. In her most recent article, Avery's girlfriend has written that Nintendo was founded in 1889. Avery finds this very surprising – incredible, even – and so doubts whether Nintendo was founded in 1889. But she is so distressed at the thought of her girlfriend making such a silly mistake in such a public venue that she doesn't inquire into whether Nintendo was founded in 1889, for example by looking up the Wikipedia article about Nintendo.

Avery doubts whether P, but fails to inquire into this question. As such, she violates RID.

If we found out that Avery is the kind of epistemic agent who, when someone she loves states that P but she doubts whether P, will not inquire into whether P for fear of her loved one having made a mistake, we would criticise her. Specifically, we would criticise her in her capacity as an epistemic agent, rather than as a moral agent, or as an agent bound by prudence. (Indeed, she may be morally praiseworthy for not wanting to find out that her loved one has made such a mistake; this demonstrates the moral virtue of loyalty. Further, it would certainly be prudentially prudent, if you will, for her to avoid inquiring into this question, given that she knows she would be more distressed if she found out her loved one was mistaken.) Avery is intellectually cowardly, where Larry was intellectually negligent, and both cowardice and negligence are standardly taken to be intellectual vices (see Zagzebski 1996: 152). Upon discovering her intellectual cowardice, we would reduce epistemic trust in her in one respect that Kauppinen highlights: we would be "less willing to regard [her] as a potential partner in co-operative inquiry" (2018: 1). For Avery's inquiring habits don't seem to be motivated by proper concern with attaining knowledge; she actively avoids trying to attain knowledge when doing so would be emotionally difficult for her.

Thus, I hold that we have reason to think that RID is an epistemic norm: we epistemically criticise agents for failing to abide by it. However, I don't think that RID is an epistemic norm. This is because doubting Q doesn't imply that one's situation with

respect to Q is, in fact, epistemically risky. In cases in which one's situation with respect to some question isn't in fact epistemically risky, it isn't the case that one ought to inquire into that question. But this is a topic to be explored in the next, and final, chapter.

4. Conclusion

In this chapter, I argued that similar considerations as those motivating the inquiry-stopper picture of *knowledge* suggest a distinct conceptual need for an inquiry-starter: a concept that functions to signal when inquiry ought to begin. I hypothesised that this is the function of the concept *doubt*, or rather the concept that picks out the questioning attitude of doubt. I tested this hypothesis via two conceptual reverse-engineering projects, one using a synchronic model and one using a diachronic model. Both confirmed my hypothesis. The picture of *doubt* that emerged from these conceptual reverse-engineering projects has it that, typically, a subject S doubts some question Q if:

1. S has a questioning attitude to Q,
2. S doesn't believe any complete answer to Q,
3. S's situation with respect to Q is epistemically risky, or is represented to her as epistemically risky,
4. S is immediately motivated to inquire into Q.

I then raised and responded to a number of objections to this picture of *doubt* as the inquiry-starter.

Chapter 5. From doubt to epistemic anxiety

1. Introduction

In the previous chapter, I reverse-engineered the concept *doubt*, ending up with a characterisation of the concept as typically applying to some subject S with respect to a question Q if:

1. S has a questioning attitude to Q,
2. S doesn't believe any complete answer to Q,
3. S's situation with respect to Q is epistemically risky, or is represented to her as epistemically risky,
4. S is immediately motivated to inquire into Q.

Is there anything in the world that this concept picks out? In this chapter, I argue that there is: the epistemic emotion of epistemic anxiety. Epistemic anxiety is a phenomenon that has been posited to undermine the motivation for stakes-sensitive theories of knowledge, according to which what a subject knows is partly determined by what is at stake for her in the context. Jennifer Nagel argues that knowledge is not sensitive to stakes, rather our reluctance to attribute knowledge in high-stakes contexts is due to our expectation that subjects will think adaptively: they will invest greater cognitive resources into forming beliefs when error would be particularly costly, or true belief particularly beneficial. Nagel posits epistemic anxiety as a "force" (2010a, p. 408) that triggers subjects to gather information and reason more carefully in high-stakes contexts. However she does not have much to say about the nature of epistemic anxiety: what it is, and how it serves its function. One of my aims in this chapter is to provide such an account of epistemic anxiety. I argue that epistemic anxiety is an emotional response to epistemic risk.

The motivation for my account of epistemic anxiety is threefold. First, it makes epistemic anxiety a species of anxiety, thus rendering psychologically respectable a notion that has heretofore been taken seriously only by epistemologists. Second, my account of epistemic anxiety contributes to recent philosophical work on risk, by specifying to which philosophical kinds of risk anxiety can be a response. Anxiety, the broad emotion of which I will argue epistemic anxiety is a species, is understood by psychologists to be an emotional response to risk. But psychologists, very reasonably, have little to say about risk itself, as opposed to risk judgement. Philosophy can aid

psychology on this matter. Three accounts of risk have gained prominence in the philosophical literature: the probabilistic account, on which the risk of a negative event is determined by the likelihood of its obtaining; Duncan Pritchard's (2015) modal account, on which risk is determined by the closeness of worlds in which a negative event obtains; and Philip Ebert, Martin Smith and Ian Durbach's (2020) normic account, on which risk is determined by the most normal worlds in which a negative event obtains. I argue that anxiety is a response to normic and probabilistic risk. Third, my account improves on extant accounts of epistemic anxiety. It is more fleshed out than Jennifer Nagel's (2010a), which is largely agnostic about the nature of epistemic anxiety, focusing instead on what work it does in our epistemic lives. In offering an account of epistemic anxiety as an emotion, my account explains how it is able to do the epistemological work to which Nagel puts it. My account is also more plausible than Juliette Vazard's (2018, 2021), on which epistemic anxiety is an emotional response to potential threat to one's practical interests. Vazard's account cannot distinguish epistemic anxiety from anxiety in general, and also fails to capture all instances of what we want to call epistemic anxiety. My account does better on both counts.

The chapter proceeds as follows. In §2, I present an account of anxiety as an emotional response to risk, on which anxiety is triggered in the presence of normic and probabilistic risk without having either notion encoded in its representational content. In §3, I apply this picture to the epistemic realm, arguing that epistemic anxiety is an emotional response to epistemic risk, triggered in the presence of normic or probabilistic epistemic risk, but not modal epistemic risk. I introduce an objection from anti-risk epistemology: modal epistemic risk is the kind of epistemic risk that threatens knowledge, so if epistemic anxiety doesn't track modal epistemic risk, can it be epistemically valuable? I respond to this objection by appealing to the idea of normative coincidence: I argue that one cannot aim to reduce normic risk without aiming to reduce modal risk, so insofar as epistemic anxiety motivates us to reduce normic epistemic risk, it motivates us to reduce modal epistemic risk. In §4, I compare my account of epistemic anxiety to Nagel's and Vazard's. I then argue that epistemic anxiety should be understood as identical to doubt, and in §5 demonstrate an advantage of making this identification: we can evaluate doubts as we evaluate anxieties in general, as doubts are a species of anxiety. This allows us to shed new light on very old debates in epistemology over the value of sceptical doubts, as well as making clear why there isn't a norm on inquiry that tells us to inquire if we doubt: some

doubts are improper. That is, we will see why the RID norm, articulated in the previous chapter, is not a genuine epistemic norm. In §6, I explicate the concept *doubt* that emerged from the previous chapter, replacing it with the more exact concept *epistemic anxiety*.

1.1. Disambiguating 'epistemic risk'

Before embarking on the primary task of this chapter, that of developing my account of epistemic anxiety as an emotional response to epistemic risk, I must note a discrepancy in how 'epistemic risk' is used in the literature, and justify my using this expression in the way that I do throughout this chapter. There are different ways that 'epistemic risk' is used within epistemology. On one use, 'epistemic risk' means simply risk in the epistemic realm. This is how I characterised epistemic risk in the previous chapter (§2.2). There, I said that risk has to do with the potential of some disvaluable event to occur; as such, epistemic risk, in this sense, has to do with the potential of some epistemically disvaluable event to occur. A variety of events can be epistemically disvaluable, for example holding a false belief, missing out on true belief or knowledge, misunderstanding, and failing to understand. Therefore, there are a variety of ways that a situation can involve epistemic risk, in this sense. Epistemologists who use 'epistemic risk' in this broad sense are Jesús Navarro (2019) and the more recent Duncan Pritchard (2021). On a second use, 'epistemic risk' picks out the specific risk of forming or holding a false belief. This narrow use of 'epistemic risk' is found more often than the broad sense in the literature (see Collins 1996, Wright 2004, Lasonen-Aarnio 2008, Smith 2012, Pritchard 2016, Pettigrew 2019; Vazard 2021: 6921).

There are further uses of 'epistemic risk' within philosophy than these. Boris Babic (2019), a decision theorist, uses 'epistemic risk' to pick out two kinds of epistemic risk: risk of false belief and risk of inaccurate credence. This use is broader than the very narrow use of 'epistemic risk' as picking out just the risk of false belief, but is narrower than the broad use of 'epistemic risk' as picking out the risk of all kinds of disvalue in the epistemic realm. Justin Biddle and Rebecca Kukla (2017), working within the philosophy of science, use the term 'epistemic risk' to name only the risks of accepting a false hypothesis or rejecting a true hypothesis in scientific inquiry. This kind of risk is standardly called 'inductive risk', hence I shan't worry about what kind of risk Biddle and Kukla are interested in, whether it is a risk of the same kind that these other philosophers call 'epistemic risk', and so on.

What is important for my purposes is that 'epistemic risk' is used within epistemology in more and less broad senses. 'Epistemic risk' is used sometimes to pick out a wide variety of risks in the epistemic realm, sometimes to pick out a narrow variety of risks in the epistemic realm (as in Babic's use), and sometimes to pick out only one kind of risk in the epistemic realm: the risk of false belief. As noted, this third and most narrow use of 'epistemic risk' is the use most often found in the literature. Given the variety of disvalue in the epistemic realm, this use might strike one as unduly narrow. But there is a clear reason why this kind of risk is that with which epistemologists have typically been concerned. This is because a primary task of epistemologists is to specify what is required in order for a subject to know some proposition P. It is near-universally held that true belief that P is necessary for knowledge that P (though see Radford 1966 for a(n in)famous denial of the necessity of belief for knowledge); the task then becomes that of specifying what is required to turn true belief into knowledge. The kind of epistemic risk that undermines a true belief's claim to knowledge is the risk of that belief's being false, given the way it was formed. It doesn't undermine S's true belief's claim to knowledge that S was at risk of failing to form that belief, or that she was at risk of missing out on very good evidence that she actually has (cf. Pritchard 2005: 133-140, for the point made in terms of luck, rather than risk). But it does undermine S's true belief's claim to knowledge that she was at (a sufficiently high) risk of believing falsely, given the way she formed her belief.

Given that it is the risk of false belief that is relevant to whether a true belief amounts to knowledge, it is obvious why epistemologists have been preoccupied with this kind of epistemic risk. Even the recent Pritchard, who acknowledges the varieties of epistemic risks that are out there, treats the risk of false belief as the "fundamental epistemic risk" (2022: 14). He writes that, since truth is "fundamental to the epistemic domain of evaluation, so the avoidance of error (and thus false belief) is also fundamental, albeit in a negative sense (i.e., to be avoided rather than promoted)"; as such, the risk of false belief is the "fundamental" or "core epistemic risk" (14).

These epistemologists occupy what Bernard Williams calls the "examiner situation" (1973: 146): the situation in which one knows that P, knows that some subject S believes that P, knows all the relevant facts about S's situation, and asks whether S knows P. But when we switch perspectives to what we might call the 'inquirer situation', whereby we are inquiring (or deliberating about whether to inquire) into some question Q and don't know Q's answer, the idea that the risk of false belief is the

most fundamental epistemic risk seems unwarranted. Supposing that inquiry aims at knowledge (see Chapter 4, §3.3), to falsely believe an answer to Q takes one no further from achieving the aim of inquiry than to fail to believe any answer to Q. From the perspective of an inquirer seeking knowledge of Q's answer, failing to believe the true answer is as bad as falsely believing. So, from the perspective of the inquirer, there is no reason to treat the risk of false belief as more fundamental than the risk of failing to truly believe, or failing to achieve knowledge.

Or so one might think. However there is a sense in which one who falsely believes an answer to Q is, after all, further from achieving the aim of inquiry than one who simply fails to believe any answer to Q. As belief is the move that closes inquiry (see Kelp 2021b, and §3.3 of the previous chapter), an inquirer who falsely believes some answer to Q will close her inquiry into Q. Having closed her inquiry into Q, she typically won't re-open inquiry without receiving new information that suggests that her belief is false. In contrast, an inquirer who fails to form a belief in any answer to Q will be more likely to re-open her inquiry into Q in future situations in which the question is made salient, even without receiving new information that bears on the truth of any possible answer to Q. If all goes well, this further inquiry will lead her to believe the true answer to Q. But even if it doesn't, and she remains in suspension after further inquiry, she will still be disposed to re-open this inquiry on yet other future occasions in which Q is made salient, even without receiving new information that bears on the truth of any answer to Q. Therefore the 'type I' error of failing to believe the true answer to Q is more likely to be rectified than the 'type II' error of falsely believing an answer to Q. Then far from undermining the treatment of the risk of false belief as the fundamental epistemic risk, shifting perspective from the examiner situation to the inquirer situation gives us new justification for so treating the risk of false belief. And throughout this chapter, this is what I will do.

Still, my focus on the risk of false belief in what follows should not be taken as an endorsement of a narrow conception of epistemic risk, on which the risk of false belief is the *only* epistemic risk. There are a variety of epistemically disvaluable events, and therefore a variety of epistemic risks. As such, my use of 'epistemic risk' throughout this chapter should be understood as referring to epistemic risk in the broad sense. When I want to talk specifically about the risk of false belief, I will follow Pritchard (2016: 566) in using the term 'veritic epistemic risk', or I will explicitly say 'the risk of false belief'.

2. Anxiety and risk

In this section, I give an account of anxiety as a response to risk. I consider in §2.2 which of three philosophical accounts of risk – the probabilistic, modal and normic accounts – articulate a notion of risk to which anxiety can be a response. I argue in §2.3 that anxiety is triggered in response to normic and probabilistic, but not modal, risk.

2.1. Anxiety as an emotional response to risk

Psychologists conceive of anxiety as an emotional response to *threat* or *risk* (see Lader and Marks 1973, Butler and Matthews 1987, Öhman 1993, Barlow 2001, Kemeny and Shestyuk 2008). It is standard to distinguish between *trait* and *state* anxiety. 'State anxiety' refers to emotional episodes of anxiety. These are short-lived affective responses to specific stimuli. 'Trait anxiety' refers to an individual's disposition to experience state anxiety. An anxious person, someone with high trait anxiety, will be disposed to experience state anxiety more often than other people, or in response to a greater variety of stimuli. It is state anxiety with which I am concerned in this chapter. Henceforth when I mean to refer to state anxiety, I will just use 'anxiety'.

Risk is commonly characterised in terms of potential *unwanted* events. Pritchard defines risk events as "potential unwanted events" (2015: 436); Adam Bricker writes that "a risk is an unwanted possible event" (2018: 201); and Sven Ove Hansson notes that risk has been variously defined as "an unwanted event that may or may not occur", "the cause of an unwanted event that may or may not occur" and "the probability of an unwanted event that may not occur" (2018).[14] This is problematic, because what people want can fail to line up with what is good or valuable. A depressed person may want to die; then her continued survival would count as a risk for her, and her death, at least in some manners, will not. But it is extremely counterintuitive to describe her continued survival as a risk, and her death (at least in some ways) as not constituting a risk.

[14] We may assume that 'threat' and 'risk' are synonymous. Compare these definitions of 'risk' to the Cambridge English Dictionary (2013) definition of 'threat' as "unwanted possibility … the possibility that something unwanted will happen". Further, the Merriam-Webster dictionary (2020), the Collins thesaurus (2013) and the Macmillan thesaurus (2018) all list 'risk' and 'threat' as synonyms.

Pritchard notes this problem, writing that "whether an event is unwanted will be a subjective matter; one might actively want the plane to crash, for example, because one is suicidal" (2015: 437). However, he "set[s] this complication to one side and take[s] it as given" that risk events are potential unwanted events (437). I do not wish to set this problem aside, so, taking a hint from the risk analysis literature, will use the terminology of *negative* events (see Jensen 2012: 436-7). Negative events are events whose obtaining would be disvaluable in some way, but which need not be unwanted: they might be harms (Möller 2012: 74), or events that involve loss of moral, aesthetic, or other kinds of value. I intend for 'negative' to be a placeholder for whatever way risk events are disvaluable. In what follows, I use 'risk events' to name potential negative events, and 'risk-possibilities' to name possibilities in which such events obtain.

Anxiety is an emotional response to risk. It functions to direct the experiencer's attention towards some risk-possibility and motivate her to take steps to avoid or reduce the relevant risk. There are three elements to anxiety that enable it to function in this way. First, anxiety has representational content. It represents some event as possible, in a robust sense. That is, when one experiences anxiety about an event E, E is represented to one not just as merely metaphysically possible – as something that happens in some possible world, however bizarre is that world, or however incompatible that world is with how one knows the actual world to be – but as something that, for all one knows, might happen in the actual world. We can say, then, that one's anxiety represents E to one as *epistemically*, not merely metaphysically, possible: possible relative to what one (or perhaps some wider group of subjects) knows.[15] Pritchard argues that the relevant sense of 'possibility' is even more restrictive than this: not only must a risk-possibility be epistemically possible, it must be "realistic", where this means it is "something that could credibly occur" (2015: 439; see also Grimm 2015: 132). In what follows, I assume that anxiety represents an event to its experiencer as *at least* epistemically possible. One's anxiety also represents that

[15] Some philosophers reject characterisations of epistemic possibility in terms of what some subject or group of subjects know. For example, Dougherty and Rysiew argue that P is epistemically possible for a subject S only if P is compatible with S's evidence, which is *not* identical to what S knows: "what is epistemically possible for a subject is those things which his *evidence*, rather than what he *knows*, does not rule out" (2009: 127). However it is standard to define epistemic possibility in terms of what is known by a subject or group of subjects, as I have done here; see Hacking (1967: 149), DeRose (1991: 593-4), Anderson (2014: 597).

event to one as negative: disvaluable in some way. Thus, anxiety represents an event to one as a potential negative event – that is, *as* a risk event. We can put this point by saying that the formal object of anxiety is risk: risk is what anxiety is about.

Emotions have intentionality, or aboutness, in two ways. First, they are about particular objects or states of affairs. If I feel affection towards my cat, my affection is *about* him. This sense of aboutness is that of having a material object. Second, emotions represent their material objects as (dis)valuable in particular ways. My affection towards my cat represents him as worthy of affection. This sense of aboutness is that of having a formal object. While the material object of an emotion is different in different cases, the formal object of an emotion cannot vary between different token experiences of that emotion. This is because an emotion's formal object determines just what emotion it is: emotions are (at least partly) individuated by their formal objects. In the case of anxiety, though the material object of anxiety can be any number of things, its formal object is always risk. For example, one can be anxious about missing one's train, about responding to an email, about the supermarket being out of goat's cheese. In all these cases, one's anxiety represents the event in question to one as a risk event: a potential negative event. If one's emotional experience did not represent a situation as involving risk, it would not be an experience of anxiety.

The second element of anxiety that enables its function is its unpleasant affective aspect. Anxiety is experienced as aversive: as "tension, unease and concern" (Vazard 2018: 142). The representational and affective elements combine to give anxiety its third element: motivation to risk-reduction behaviours. To see the three elements of anxiety in action, consider an example. If I am anxious about catching SARS-CoV-2, the virus that causes COVID-19, my catching SARS-CoV-2 is represented to me as negative and as robustly possible in the way described above. My anxiety has an unpleasant affective aspect: when I think about situations in which I could catch SARS-CoV-2, I feel uneasy and tense. I thus experience my anxiety as aversive, and consequently am motivated to engage in risk-reduction behaviours: to avoid gathering with other people indoors; to wear a mask when I cannot avoid this; and so on. This motivation is immediate: I do not need to have any independent desire to reduce the relevant risk in order to be motivated to do so; rather, my anxiety provides the motivational power required for me to try to reduce the relevant risk.

From the picture taking shape, it should be clear that, perhaps contrary to folk thinking about anxiety, anxiety is a very valuable emotion in our emotional toolbox.

Appropriately experienced anxiety brings to one's attention possibilities whose obtaining would be negative, and immediately motivates one to take steps to guard against those possibilities' obtaining. Consider Charlie Kurth's discussion of the anxiety experienced by neurosurgeon Henry Marsh. Marsh "sees [his anxiety] as the manifestation of his accumulated surgical expertise: when determining whether to remove more of a tumor – at the risk of damaging healthy brain tissue – he is guided by his anxiety" (2018: 3). When he starts feeling anxious, he stops operating (see Marsh 2004). Marsh's anxiety here focuses his attention on a risk event, that of damaging healthy brain tissue, and motivates him to avoid the risk by stopping the surgery. His anxiety is thus very valuable, given his goal of removing tumours without damaging healthy brain tissue.

More generally, anxiety can be positively evaluated both in terms of its fittingness and its utility. As the formal object of anxiety is risk, anxiety will be fitting when it is a response to genuine risk, and unfitting when there is no risk present. For example, Marsh's anxiety is fitting: there is a genuine risk of damaging healthy brain tissue, and his anxiety is a response to that risk. In contrast, the anxiety experienced as part of generalised anxiety disorder (GAD) is unfitting: it is either undirected, in the case of "free-floating anxiety", or it is directed at everyday events that do not involve risk (ICD-11, World Health Organisation 2018). The utility of anxiety is a matter of whether it helps one to avoid or reduce risk, and the extent to which it does this. Marsh's anxiety is useful because, in motivating him to end neurosurgery at a certain point, it reduces the risk of him damaging healthy brain tissue. Again, in contrast, the GAD sufferer's anxiety is useless, as there is no genuine risk to avoid or reduce.

2.2. The nature of risk

If anxiety is fitting only in response to genuine risk, we must get clearer on what exactly risk is. I have so far said that risk events are 'potential negative events': negative events that obtain in some epistemically possible world. Each of the three philosophical accounts of risk that are prominent in the literature – the probabilistic, modal and normic accounts (§1) – can accept this claim. Where the three accounts diverge is over what it is that determines the level of risk of some risk event.

The "standard" account of risk (so called by Pritchard 2015: 436, Bricker 2018: 200, and Ebert, Smith and Durbach 2020: 432) is the probabilistic account, which has it that the level of risk of some risk event is a matter of the probability of that event's

obtaining. On the probabilistic account, an event E is a risk event iff E is a negative event with a non-zero probability of obtaining; high-risk events are negative events with a high probability of obtaining and low-risk events are negative events with a low probability of obtaining, with a continuum of riskiness between these extremes; and a negative event E1 is higher risk than a negative event E2 if the probability of E's occurring is higher than the probability of E2's occurring.

The relevant notion of probability here is *evidential* or *epistemic* probability: probability relative to a body of evidence (Ebert, Smith and Durbach 2020: 433). Evidential probability is to be contrasted with physical probability, which can be thought of as the brute frequency with which tokens of some event-type obtain, not relative to any subject's thinking nor to any body of evidence; and subjective probability or credence, which measures a subject's degree of confidence in a given proposition (Mellor 2005: 8). Evidential probability is not wholly objective, like physical probability, because it is always relative to a body of evidence. But neither is it subjective, like credence. Evidential probability is objective to the extent that, given a body of evidence, there is a fact of the matter about what is the evidential probability of E's obtaining. Then interpreting the probabilistic account in terms of evidential probability "reflects the fact that this is an account of risk as an objective phenomenon", to use Pritchard's (2015: 440) words, without making risk something wholly beyond our ken.

Pritchard (2015) and Ebert, Smith and Durbach (2020) challenge the standard probabilistic account of risk. They argue that risk is not determined solely by the probability of a negative event's obtaining, but rather is (at least sometimes) a matter of how easily a negative event could obtain (Pritchard), or the extent to which that event's obtaining would call out for explanation given a body of evidence (Ebert, Smith and Durbach). Pritchard calls his picture of risk the *modal* account, and Ebert, Smith and Durbach call theirs the *normic* account.

Pritchard motivates his account of risk by appeal to a pair of cases in which it is stipulated that the probability of a negative event's obtaining is equal in both, but in which the event could obtain more easily in one case than in the other. These are his bomb cases:

> Case 1: An evil scientist has rigged up a large bomb, which he has hidden in a populated area. If the bomb explodes, many people will die. There is no way of discovering the bomb before the time it is set to detonate. The bomb will only

> detonate, however, if a certain set of numbers comes up on the next national lottery draw. The odds of these numbers appearing is fourteen million to one.
>
> Case 2: [All is the same as in Case 1, but] The bomb will only detonate if a series of three unlikely events obtains. First, the weakest horse in the field at the Grand National, Lucky Loser, must win the race by at least ten furlongs. Second, the worst team remaining in the FA Cup draw, Accrington Stanley, must beat the best team remaining, Manchester United, by at least ten goals. And third, the queen of England must spontaneously choose to speak a complete sentence of Polish during her next public speech. The odds of this chain of events occurring are fourteen million to one. (2015: 441)

Pritchard argues that there is a much higher risk of the bomb going off in Case 1 than in Case 2, despite the identical odds, because in Case 1 the bomb blast "could very easily occur. All it would take for the bomb to go off … is for a few coloured lottery balls to fall in a certain configuration" (442). In Case 2, in contrast, the bomb blast couldn't easily occur. For the bomb to go off, not one, but three, incredibly far-fetched events must take place. The probabilistic account cannot explain why the risk of the bomb going off in Case 1 is, at least according to Pritchard's intuitions, higher than in Case 2, as it is stipulated that the probability of the bomb blast is the same in both cases.

Ebert, Smith and Durbach dispute the probabilities Pritchard offers for Case 2 as "unreasonably high" (2020: 436). They note that they were offered odds of 5000 to one on a bet similar to Pritchard's first condition from a bookmaker, and were denied quotes on the second and third conditions due to their "extremely improbable nature" (437, n. 8). Supposing that each of the three conditions has the same probability of 5000 to one, and treating the conditions as mutually independent, Ebert, Smith and Durbach derive a probability of 1 in 125 billion for the bomb going off in Case 2. With these more realistic probabilities in place, the probabilistic account straightforwardly predicts that the bomb going off in Case 1, with odds of 14 million to one, is much riskier than its going off in Case 2, with odds of 125 billion to one. Pritchard pre-empts this objection, arguing that "even though one can always dispute the assignment of probabilities in a particular case, it ought to be clear that there will inevitably be pairs of cases of this general type" (2015: 443) – that is, cases where the probability of two risk events is the same, but their modal closeness differs – so "even if they manage to make this claim stick in this particular pair of cases, it ought to be clear that the

underlying problem is not thereby solved" (444). Pritchard concludes that the probabilistic account is "highly problematic", indeed "fundamentally misguided" (436), as it cannot capture "our natural judgements … about risk" (447).

As such, Pritchard argues that we should abandon the probabilistic account of risk in favour of his modal account of risk, on which the risk of a negative event E is determined by the closest worlds in which E obtains. A world is close to the actual world if it is similar to the actual world, and worlds become more distant to actuality as they become more dissimilar, with similarity being a matter of how much needs to change to get from the actual world to a given possible world. On Pritchard's modal account of risk, the closer is the closest world(s) in which a negative event E obtains, the riskier is E (447). E is high-risk if, keeping relevant initial conditions fixed, E obtains in a close possible world (2016: 562). (That we must 'keep relevant initial conditions fixed' is a metaphysical condition: it means that the relevant possible worlds are identical to the actual world in certain respects up until E's obtaining; see Lewis 1973: 566-7.) As the closest world in which E obtains becomes more distant, the riskiness of E lessens, until eventually E is so remote as to not constitute a risk event.

Ebert, Smith and Durbach agree with Pritchard that there is more to risk than the probabilistic account has it. However, they disagree with Pritchard in two respects. First, they don't accept that Pritchard has shown the probabilistic account to be "fundamentally misguided" (Pritchard 2015: 436). They argue that there are cases for which the probabilistic account can deliver the intuitively correct verdict, but the modal account cannot. As example, they offer a new bomb case:

> An evil scientist has rigged up a large bomb, which he has hidden in a populated area. If the bomb explodes, many people will die. There is no way of discovering the bomb before the time it is set to detonate. The bomb will only detonate, however, if six specific numbers between 1 and 99 come up on the next national lottery draw. The odds of these six numbers appearing are roughly one billion to one. (2020: 446)

Call this 'Case 3'. The odds of the bomb going off in Case 3 are much lower than in Case 1: one in one billion, compared to one in 14 million. The probabilistic account judges that the risk of the bomb going off is much higher in Case 1 than in Case 3. However, the closest world in which the bomb goes off in Case 3 is equally close as in Case 1: just as in Case 1, all that is required for the bomb to go off in Case 3 is that a

few coloured balls fall in a particular configuration. Thus the modal account judges that the risk of the bomb going off is equal in Cases 1 and 3. This is at least as hard to swallow as the probabilistic account's verdict that the risk of the bomb blast is equal between Cases 1 and 2. This suggests to Ebert, Smith and Durbach that the probabilistic notion of risk still has an important place in our dealings with risk. They therefore recommend that we endorse pluralism about risk, according to which different philosophical accounts of risk are understood as different, but equally legitimate, precisifications of our intuitive notion of risk (2020: 449). However they argue that Pritchard's modal account is problematic in other ways, and that their own normic account can capture the benefits of Pritchard's account without these untoward elements.[16] So their risk pluralism does not involve endorsing the modal account. This is their second point of disagreement with Pritchard.

On Ebert, Smith and Durbach's normic account, the risk of a negative event E isn't determined by the closeness of worlds in which E obtains, but by the normalcy of those worlds. The notion of normalcy appealed to is that developed by Smith (2016) in terms of calling out for explanation. Normal conditions don't call out for any special explanation, relative to a body of evidence, while abnormal conditions do. For example, if I say "Iain would normally be home by six," part of what I mean by this is that, if Iain failed to be home by six, some explanation would be required (2016: 39). Perhaps his car broke down, perhaps there was traffic, perhaps he stopped for ice cream. In any case, if Iain is normally home by six, his failing to be home by six calls out for explanation. Possible worlds can be ordered by their normalcy. The most normal worlds are those worlds whose obtaining would call out for the least explanation, given a body of evidence. Worlds become more abnormal as their obtaining calls out for more explanation. On the normic account of risk, the risk of an event E is determined by the most normal worlds in which E obtains. The more normal are these worlds, the higher the risk of E (2020: 444).

The normic account of risk issues the same verdicts as the modal account for the bomb cases. Case 1 is judged to be riskier than Case 2, as it would not be abnormal

[16] Ebert, Smith and Durbach's primary objections to the modal account centre on its consequence that any actually obtaining risk event is maximally risky, and the corollary that we cannot calculate the risk of an event without taking a stand on whether it will actually obtain. This leaves the modal account of risk ill-suited to play important roles we expect of a notion of risk (2020: 441-2).

for the detonation-triggering numbers to come up in the lottery, whereas it would be very abnormal for even one, let alone all three, of the triggering conditions to obtain in Case 2. Then the most normal worlds in which the bomb goes off in Case 1 are very normal, whereas the most normal worlds in which the bomb goes off in Case 2 are very abnormal. So the risk of the bomb going off is much higher in Case 1 than in Case 2. Case 1 and Case 3 are again judged to be equally risky. Just as the bomb blast worlds are equally close in Case 1 and Case 3, these worlds are equally normal. Given our evidence, the bomb blast in Case 3 would require no more explanation than it would in Case 1: in both cases, the detonation-triggering numbers might as well come up as any other series of numbers in the lottery. So there is still a place for the probabilistic account of risk in Ebert, Smith and Durbach's risk pluralism.

An important difference between the normic and the probabilistic accounts of risk, on the one hand, and the modal account on the other, is that only the former two have it that risk is always relative to a body of evidence. This difference can be thought of in terms of objectivity. The modal account of risk is fully objective, in that it has it that risk is a brute fact about the world, neither relative to any subject's beliefs or feelings about risk, nor to any body of evidence. Though a body of evidence might suggest that some worlds are close and others further away, which worlds are close is not determined by evidence. The normic and probabilistic accounts deliver notions of risk that are less than fully objective, though not subjective either. There is a fact of the matter about what is the normic or the probabilistic risk of an event E's obtaining, but only relative to a body of evidence. In contrast, there is a fact of the matter about what is the modal risk of E's obtaining, and this is not relative to any body of evidence (except, perhaps, all the evidence in the world).

2.3. Tracking risk

If anxiety is an emotional response to risk, does that mean that it tracks risk of all different kinds? I argue that it does not. Rather, I argue that anxiety can track normic and probabilistic risk, but because modal risk is fully objective – it is neither relative to any subject's beliefs or feelings about risk, nor to any body of evidence – it is not something that anxiety can track. This might surprise Pritchard, who appeals to the affective situation of the subjects in his bomb cases to support the verdicts issued by his modal account. Regarding Case 1, in which the bomb blast would be triggered by lottery, Pritchard writes that "[nobody] who knew about the bomb plot would be sitting

comfortably while watching the next lottery draw" (2015: 442). However in Case 2, there would be no "corresponding cause for alarm" (441). These affective responses can be understood in terms of anxiety: the subjects in Case 1 are anxious about the bomb going off, while the subjects in Case 2 are not. This seems to suggest that anxiety can track modal risk. But note that in these cases, normic risk and modal risk correspond: in Case 1, both normic and modal risk are high, while in Case 2, both are low. What we need in order to see whether anxiety tracks modal risk are cases where modal and normic risk diverge: where modal risk is high and normic risk low, or *vice versa*. I will now offer two such cases, and argue that in these cases, anxiety tracks normic and not modal risk.

Recall Pritchard's Case 2, in which the bomb will go off only if a) Lucky Loser wins the Grand National by at least ten furlongs; b) Accrington Stanley beats Manchester United by at least ten goals; c) the queen of England spontaneously chooses to speak a complete Polish sentence in her next public speech. Suppose that I am in a case like Case 2, but in which, unbeknownst to me, Lucky Loser has been given performance enhancing drugs, while all the other horses have been given tranquilisers; all of Manchester United's defenders have broken their toes; and the queen of England has been watching lots of Polish films and keeps bursting out with quotes from her favourites at inopportune moments. In this new case – call it 'Case 4' – although the normic risk of the bomb going off, relative to my evidence, is the same as it was in Case 2, the modal risk is much higher. But this would make no difference to the level of anxiety I feel about the bomb blast. I would remain as sanguine about the bomb blast in Case 4 as I would be in Case 2. Thus in Case 4, in which normic risk is low and modal risk is higher, anxiety seems responsive to normic and not modal risk.

Now imagine a case like Case 2, but in which I have received misleading information from an ordinarily trustworthy friend that things are as described in Case 4: that is, she has told me that Lucky Loser has been given performance enhancing drugs, that Manchester United's defenders have all broken their toes, and so on. However none of what she tells me is true.[17] In this case, given my evidence (which now includes that an ordinarily reliable friend has told me that Lucky Loser has been

[17] It doesn't matter why my ordinarily trustworthy friend has come out with false testimony on this occasion. Perhaps a brain lesion has formed overnight and caused her to have these odd beliefs (cf. Plantinga 1993).

given performance drugs, and so on), the normic risk of the bomb going off is higher than in Case 2. The modal risk, however, is the same: the same amount would have to change in this case – call it 'Case 5' – to get from the actual world to the bomb blast world as in Case 2. But in Case 5, I would be more anxious about the bomb blast than I would in Case 2. Again, my anxiety seems responsive to normic and not modal risk.

Because the modal account of risk has it that risk is determined by what goes on in other possible worlds, and subjects can lack evidence, or indeed have misleading evidence, about what goes on in those worlds, the modal account makes it the case that what determines risk is not something that is entirely epistemically accessible to subjects. Our epistemic access to closeness orderings on worlds is incomplete. As we have just seen, sometimes our evidence will suggest that a world is close, when it is in fact distant; and sometimes our evidence suggests that a world is distant, when it is in fact close. Where a risk-possibility is close but one's evidence does not suggest this, one doesn't experience anxiety; and where a risk-possibility is distant but one's evidence suggests that it is close, one does experience anxiety. Then anxiety doesn't track modal risk. The kinds of risk to which anxiety can be a response cannot be fully objective in the way that the modal account has it. Rather, the risk that anxiety tracks must be evidence-relative.

This doesn't mean that the experience of anxiety must be mediated by conscious reflection on one's evidence. It is more plausible that, in the cases just discussed, anxiety is triggered by stimuli in one's environment to which one has epistemic access, and facts about these stimuli also feature in one's body of evidence, which determines normic risk. Then when one's situation involves high normic risk, one also feels anxious. We could say that anxiety tracks high normic risk in that both are common effects of a single cause: some feature(s) of one's environment to which one has epistemic access both triggers anxiety and leads to certain facts being in one's body of evidence, which determines that the normic risk is high. In some cases, however, I do think anxiety will be generated by conscious reflection on one's evidence; for example, when probabilistic risk is high. A (non-pathological) subject will feel anxious when she plays a death lottery of the kind described in Shirley Jackson's *The Lottery*, because this situation is normically very risky. But she would feel more anxious if she knew there were only 100 participants in the lottery than if she knew there were 100 million participants. In this case, her reflection on her evidence alerts her to the

higher probabilistic risk of her death in the case of the smaller lottery, and her heightened anxiety is a response to this greater risk.

Psychologists working on risk have suggested that two different systems are involved in our risk judgements: an affective system and a cognitive system (Loewenstein et al. 2001; Shulman and Cauffman 2014). Ebert, Smith and Durbach suggest that their risk pluralism could be merged with this 'dual system' approach to risk judgement, generating a picture on which normic risk is tracked by the affective system, and probabilistic risk by the cognitive system (2020: 449). If I am right, however, this is too quick. For anxiety is responsive to both normic and probabilistic risk, just in different ways: in the face of significant probabilistic risk, anxiety can be generated by reflection on one's evidence; whereas in the face of high normic risk, anxiety is triggered directly by features of one's environment. Then the affective system may still be involved in a subject's making a probabilistic risk judgement; that a subject's risk judgement is informed by emotional experience does not mean that the kind of risk present in the subject's environment is normic, rather than probabilistic.

Note that claiming that anxiety tracks high normic and probabilistic risk, but not modal risk, is not to claim that anxiety represents a situation as involving risk of the former two kinds. On my picture, anxiety simply represents a situation as involving risk: one's situation is represented to one as implying the robust possibility of a negative event obtaining. This feels affectively unpleasant, and one is thereby motivated to try to take steps to guard against this event's obtaining. But it is not part of the representational content of one's emotional experience that the event's obtaining would be normal, or probable. What kind of risk is present when anxiety is generated need not be part of the content of one's emotional experience for anxiety to function in the way that it does.

3. Epistemic anxiety and epistemic risk

In this section, I apply this picture of anxiety and risk to the epistemic realm, arguing that epistemic anxiety is an emotional response to epistemic risk. Like anxiety in general, epistemic anxiety tracks normic and probabilistic (epistemic) risk, not modal (epistemic) risk. I raise an objection (§3.2) to my account of epistemic anxiety from anti-risk epistemology: modal risk of false belief, i.e. modal veritic epistemic risk, is the only epistemic risk that undermines knowledge; so if epistemic anxiety doesn't track modal epistemic risk, and in particular, modal veritic epistemic risk, how can it be

epistemically valuable? My answer is that epistemic anxiety is valuable because the goals of reducing normic epistemic risk, on the one hand, and reducing modal epistemic risk, on the other, are normatively coincident: one cannot aim for one without aiming for the other. Then insofar as epistemic anxiety motivates one to reduce normic epistemic veritic risk, it motivates one to reduce modal epistemic veritic risk. As such, epistemic anxiety is valuable from the perspective of anti-risk epistemology.

3.1. Epistemic anxiety as a subspecies of anxiety

As noted in §1.1, 'epistemic risk' is used in different ways in the epistemological literature. On one sense, 'epistemic risk' picks out risk in the epistemic realm, and can take many forms: risk of false belief, risk of failing to form a valuable true belief or a knowledge-constituting belief, risk of missing out on a key piece of evidence, and so on. But the term is more commonly used just to pick out the risk of false belief (see Collins 1996: 208; Wright 2004; Lasonen-Aarnio 2008; Smith 2012; Pritchard 2016; Pettigrew 2019; Vazard 2021: 6921). I am using 'epistemic risk' in the former, broad sense, on which it picks out risk in the epistemic realm; following Pritchard, I use 'veritic epistemic risk' to talk about the risk of false belief. Epistemic risk can be understood as a subspecies of risk in general, where the relevant negative event is epistemically disvaluable. Like risk in general, the notion of epistemic risk can be precisified along probabilistic, modal or normic lines. One's situation involves probabilistic epistemic risk if it is a situation in which some epistemic risk-event is probabilistically likely to obtain; one's situation involves modal epistemic risk if it is a situation in which some epistemic risk-event could easily obtain; and one's situation involves normic epistemic risk if it is a situation in which some epistemic risk-event's obtaining would be normal, in that it would not call out for explanation given the relevant body of evidence.

As such, I propose that epistemic anxiety is to epistemic risk as anxiety in general is to risk in general. That is, I propose that we understand epistemic anxiety as an emotional response to epistemic risk. Epistemic anxiety will then have the following representational, affective and motivational profile: it will represent one's cognitive situation as involving epistemic risk; it will be affectively unpleasant; and experiences will motivate subjects to engage in behaviours aimed at reducing the relevant epistemic risk. The formal object of epistemic anxiety will be epistemic risk, as the formal object of anxiety is risk. The material object of some instance of epistemic

anxiety will be some particular epistemic risk-event, for example, some particular instance of forming a false belief, or missing out on true belief.

Here is an example of epistemic anxiety in action. Suppose Lottie holds a ticket in a fair lottery with ten million tickets. Lottie considers the possibility that her ticket is a loser, and reasons that this is so overwhelmingly likely that it would be rational for her to believe it to be so. However, a niggling unease prevents her from forming the belief that her ticket is a loser. After all, one ticket will win, and it might as well be hers as any other. This is an experience of epistemic anxiety. The event of forming a false belief (because she believes that her ticket is a loser when it is, in fact, the winner) is *represented* to Lottie as an epistemic risk event; she has an *aversive affective experience* (niggling unease); and these representational and affective aspects of her emotional experience combine to *motivate* her to take steps to reduce the risk of forming a false belief, by suspending judgment on whether her ticket is a loser. In this case, the formal object of her emotional experience is epistemic risk, and its material object is the event of forming a false belief that her ticket is a loser.

Like anxiety in general, epistemic anxiety will track normic and probabilistic epistemic risk, though its representational content won't be so precise: when one experiences epistemic anxiety, one's cognitive situation is simply represented to one as epistemically risky. In the lottery example just discussed, the probabilistic risk is very low, but the normic (and indeed the modal) risk is high. The emotional experience the Lottie has which prevents her from forming the belief that her lottery ticket is a loser does not have as part of its representational content that the normic or modal risk of such a belief is high. It simply represents forming the belief to her as epistemically risky: as a potential, negative epistemic event.

Again, like anxiety more generally, epistemic anxiety is not responsive to modal epistemic risk. If Lottie receives a ticket for an upcoming ten million-ticket lottery, and her ordinarily trustworthy friend tells her that it is a ticket for last week's lottery and that it lost, then she will believe that this ticket is a loser. This belief is modally very risky. Given that this ticket is not, in fact, a ticket for last week's lottery, but for an upcoming lottery, her belief could very easily be false (because this ticket could easily be the winner). However, Lottie doesn't experience epistemic anxiety regarding this belief. This is because the relevant epistemic risk event, forming a false belief that her ticket is a loser, is not normically or probabilistically high-risk given her evidence, which includes what her ordinarily trustworthy friend has told her.

3.2. An objection from anti-risk epistemology

Here, a worry arises. If epistemic anxiety is not responsive to modal epistemic risk, then is it redundant from the perspective of anti-risk epistemology? It is widely held amongst epistemologists that knowledge is incompatible with significant veritic epistemic risk. This is the explicit thesis of Pritchard's (2015, 2016) anti-risk epistemology. But safety theorists in general can be understood as anti-risk epistemologists, since they hold that a belief that could easily have been false, given how it was formed, cannot constitute knowledge (Sainsbury 1997: 913; Sosa 1999; Williamson 2000). But it is crucial that anti-risk epistemology be formulated in terms of *modal* veritic epistemic risk. Consider again the example from the previous section. Lottie's belief that her ticket is a loser (because her friend told her it is a ticket for last week's lottery) cannot constitute knowledge: it too easily could be false. That is, it is too epistemically risky. But as noted, it is neither normically nor probabilistically epistemically risky. It is only modally epistemically risky. It is modal (veritic) epistemic risk that is knowledge-undermining. But if epistemic anxiety does not track modal epistemic risk, as I argue, then it seems that it cannot help us to reduce the kind of epistemic risk that matters when it comes to having knowledge. From the perspective of anti-risk epistemology, then, epistemic anxiety is redundant.

However I argue that epistemic anxiety is not redundant from the perspective of anti-risk epistemology. Rather, experiencing epistemic anxiety is valuable, given our concern with achieving knowledge, because reducing normic epistemic risk and reducing modal epistemic risk are *normatively coincident* goals. The terminology of 'normative coincidence' is originally due to Crispin Wright (1992: 18). Two goals are normatively coincident if a subject cannot aim for one without aiming for the other. This doesn't mean that one cannot achieve one goal without achieving the other, but that one cannot aim to bring about a situation in which one achieves one and not the other. Smith (2016: 9) gives an example to illuminate the notion. Suppose I am a member of a running club, and hold some of the best times in the club, but not the best. I am due to run a race. Two goals I might have for the race are (1) to set a new club record, and (2) to set a new personal best. Smith argues that (1) and (2) are normatively coincident, because the things I need to do to try to achieve (1) – keep fit, eat healthily, train hard – are just the things I need to do to try to achieve (2). I could end up achieving (2)

without achieving (1). But I couldn't aim to achieve (2) without aiming to achieve (1), and *vice versa*.

Smith argues that aiming for knowledge and aiming for justified belief are likewise normatively coincident goals (2016: 11). The things one must do to aim for knowledge – believe based on good evidence, suspend judgement in the face of defeaters, and so on – are exactly the kinds of things one must do to aim for (*ultima facie*) justification. One might achieve justified belief without achieving knowledge – this is how the predicament of the Gettiered subject is standardly understood. But this will be due to factors beyond one's control. One cannot aim to bring about a situation in which one has a justified belief that P, but doesn't know that P.[18]

Similarly, I argue that the goals of reducing normic epistemic risk and of reducing modal epistemic risk are normatively coincident. In particular, the goals of reducing normic *veritic* epistemic risk and modal *veritic* epistemic risk are normatively coincident. Which worlds one takes to be close will depend on how one's evidence suggests the actual world to be. The worlds that one's evidence suggests are close are those worlds in which there are few differences to how one's evidence suggests the actual world to be. The steps one would take to try to eliminate close error-possibilities, so to reduce modal veritic epistemic risk, would thus be to gather evidence about what the actual world is like; to reason about what are the ways that P could most easily be false, given this evidence (that is, to imagine worlds in which P is false, but there are as few other differences to how one takes the actual world to be as possible); to gather evidence that rules out these error-possibilities; and to suspend belief when some of these error-possibilities remain compatible with one's evidence. But these are exactly the steps that one would take to try to eliminate normal error-possibilities, so to reduce normic veritic epistemic risk. Normal worlds are worlds whose obtaining calls out for the least explanation, given a body of evidence. Generally, the obtaining of worlds in which more things are different to the most normal worlds (the worlds whose obtaining

[18] One can aim to bring about a situation in which *someone else* has a justified belief but does not know that P, for example by planting misleading evidence for them to find. Smith considers the "somewhat contrived" case in which one is about to have one's memory wiped, and plants misleading evidence for one's future self to find (2016: 16). He argues that cases like this should be thought of as cases where one attempts to manipulate the beliefs of another: one's future self is treated as a distinct epistemic subject.

would call out for no explanation) would require more explanation than worlds in which fewer things are different. Then a subject who is trying to eliminate normal worlds would have a body of evidence that includes propositions about how she takes the actual world to be; she would then reason about which error-possibilities could "just so happen" to obtain, given this body of evidence (Smith 2016: 39) – that is, whose obtaining would call out for the least explanation; she would gather evidence that rules out these error-possibilities; and she would suspend judgement while some of these error-possibilities remain compatible with her evidence.

Which possibilities one's evidence suggests are close will roughly align with which possibilities would not call for special explanation on one's evidence. Abnormal events sometimes obtain in the actual world, so the set of close worlds is not always identical to the set of normal worlds. However, where it is part of one's evidence that an event that *was* abnormal – that is, whose occurring *would have* called out for explanation, given one's prior evidence – has occurred, worlds in which that event obtains are now more normal than worlds in which it does not. Given that one's body of evidence now includes that X event obtained – call this proposition 'P' – the situation in which one's body of evidence E, which includes P, is true but P is false would call out for more explanation than the situation in which E is true and P is true, because only the former requires that a contradiction be true, which is impossible. So even though abnormal events obtain in the actual world, when it is part of one's evidence that such an event has obtained, worlds in which the event obtains are, in the end, (relatively) normal worlds, on one's evidence. Thus I hold that trying to rule out normal worlds will be the same activity as trying to rule out worlds that one's evidence suggests are close. Then to try to reduce normic veritic epistemic risk, as epistemic anxiety motivates us to do, is to try to reduce modal veritic epistemic risk, as anti-risk epistemology demands.

Again, that these two goals are normatively coincident doesn't mean that one cannot achieve one goal without achieving the other. One might succeed in reducing the normic veritic epistemic risk of a belief by gathering evidence that is incompatible with all the most normal ways one's belief would be false. But if some abnormal way in which one's belief would be false is nevertheless actual, or very close to it, then one will have failed to rule out close error-possibilities and thus failed to reduce the modal veritic epistemic risk of one's belief (at least to the same extent as one has reduced normic veritic epistemic risk). Suppose I believe that Mr Bond is not on the plane

because I know that he missed its take-off, and I know that the plane hasn't made any stops since then at which he could have got on. I have thus ruled out the most normal ways my belief could be false, and consequently it is not normically veritically epistemically risky. But suppose that Mr Bond is, in fact, on the plane – he parachuted onto it, then climbed inside through the luggage hatch, shortly after take-off. Then I didn't manage to reduce the modal veritic epistemic risk of my belief. I didn't rule out the closest way in which it might have been false – the way that actually obtains, so is maximally close.

In this case, and in many others, ruling out normal error-possibilities does not guarantee knowledge. Nevertheless, if I am right that the goal of reducing normic veritic epistemic risk and the goal of reducing modal veritic epistemic risk are normatively coincident, then in many situations in which you succeed in ruling out all the most normal error-possibilities, you will also rule out all the closest, so be able to know by the lights of anti-risk epistemology. If you have ruled out the most normal error-possibilities, and the world is obliging, then you will have ruled out the closest error-possibilities, too. So experiencing epistemic anxiety is valuable from the perspective of anti-risk epistemology, even though epistemic anxiety doesn't track modal epistemic risk. Epistemic anxiety alerts subjects to normal error-possibilities, and motivates them to take steps to eliminate these possibilities. Once you have eliminated these possibilities, you have done your part, epistemically speaking; now it is up to the world to be obliging (or not). Epistemic anxiety is valuable from the perspective of anti-risk epistemology because it motivates one to take the steps that, when the world is obliging, are the steps required to rule out the error-possibilities that prevent one from being in a position to know. By ruling out those possibilities, again, if the world is obliging, one puts oneself in a position to know the relevant proposition.

4. Other accounts of epistemic anxiety

In this section, I compare my account of epistemic anxiety to two extant accounts in the literature: Jennifer Nagel's (2010a) and Juliette Vazard's (2018, 2021). I argue that my account has an explanatory advantage over Nagel's, because it explains how epistemic anxiety is able to do the epistemological work to which Nagel puts it, and that it is preferable to Vazard's account for two reasons: my account, but not Vazard's, distinguishes epistemic anxiety from anxiety more broadly; and my account captures all instances of emotional episodes of epistemic anxiety as such, while Vazard's cannot

get a grip in cases where nothing is at stake for the subject. I then argue that Vazard's overall picture of the relationships between epistemic anxiety, doubt and inquiry is problematic, and argue that a better picture is one on which doubt is identified as epistemic anxiety, and doubt motivates inquiry.

4.1. Nagel's account

Jennifer Nagel offers an influential account of epistemic anxiety as a "force" (2010a, p. 408) that motivates subjects to gather evidence and reason carefully in certain contexts, such as those in which the practical costs of false belief would be very high. She appeals to epistemic anxiety to undermine the motivation for stakes-sensitive theories of knowledge, on which what is at stake for a subject can make a difference to what she can know. Nagel proposes a view she calls *adaptive invariantism*, according to which the standards for knowledge are invariant, and our reluctance to ascribe knowledge to subjects in high-stakes contexts arises from "an invariant expectation that subjects will think adaptively" (409): that they will invest more cognitive resources into making judgements when there is greater anticipated cost in inaccuracy, or greater anticipated reward in accuracy. Nagel uses 'epistemic anxiety' to name the "heightened need for greater evidence and more thorough processing that is characteristic of high-stakes situations" (414). However, she does not have much to say about the nature of epistemic anxiety, and in particular, does not conceive of it as an emotion, as Vazard notes (2021: 6922).

Because of this, Nagel's account of epistemic anxiety does not explain how epistemic anxiety can do the epistemological work to which she puts it. My account, on which epistemic anxiety is an emotional response to epistemic risk, can. In certain contexts, such as high-stakes contexts, uneliminated error-possibilities are salient; in high-stakes contexts, salient because their obtaining would be practically costly for the subject.[19] One's epistemic anxiety represents the relevant error-possibility to one as robustly possible and as negative; that is, as a risk-possibility, with the relevant risk event being one's forming a false belief. One's epistemic anxiety is experienced as aversive. These representational and affective aspects of one's emotional experience combine to generate motivational power: one is immediately motivated to engage in

[19] As I will argue later in this section, high-stakes contexts are not the only contexts in which epistemic anxiety can be generated by the salience of uneliminated error-possibilities.

behaviours aimed at reducing the risk of forming a false belief. Such behaviours include just the behaviours Nagel cites: evidence gathering and more careful reasoning. My account of epistemic anxiety as an emotional response to epistemic risk can thus be understood as an elaboration of Nagel's account, which explains how epistemic anxiety is able to function in the way that Nagel posits.

4.2. Vazard's account

Vazard too offers an account of epistemic anxiety as an emotion, which she likewise takes to be an elaboration of Nagel's account. However, for Vazard, epistemic anxiety is not an emotional response to epistemic risk. Rather, it is an emotional response to practical risk: risk regarding one's practical interests. Vazard posits epistemic anxiety as the emotion that gives motivational power to what she, inspired by C. S. Peirce, calls "real doubt", a kind of doubt that is "motivated by practical interests and which acts as a reason for mental and physical action" (2021: 6922). To illuminate the notion of real doubt, Vazard gives an example. Suppose that I doubt that it will rain tomorrow. As I am a philosopher, whether it will rain tomorrow has very little, if any, bearing on my practical stakes. Thus, Vazard holds, my doubting that it will rain tomorrow is not a "source of preoccupation" for me; it is not "accompanied by any specific phenomenology and it won't motivate me to launch any specific action plan" (6919). My doubt is thus not a *real* doubt. Compare a farmer who doubts that it will rain tomorrow, while her crops are threatened by drought. She would be "much more likely to experience negative feelings with respect to the situation, to see it as a problem which needs to be solved, and to be moved to action as a result" (6919). The farmer's doubt is thus a real doubt.

For Vazard, epistemic anxiety plays a role in generating real doubt, but cannot do this by itself. Rather, epistemic anxiety generates real doubt only in combination with another epistemic emotion: the feeling of uncertainty. This is a "metacognitive experience … aimed at monitoring the safety of a belief by tracking the fact that the method used to reach that belief produces true belief also in nearby possible worlds" (6930). The feeling of uncertainty is triggered when one's belief is unsafe: when one fails to form a true belief in close possible worlds in which one forms a belief via the same method. Real doubt is generated, on Vazard's account, "when epistemic anxiety appraises the matter as implying a possible threat, and feelings of uncertainty signal that a belief is unreliable" (6931). In such a situation, subjects are immediately

motivated to deploy the "costly cognitive strategies constitutive of doubt – deliberation, reasoning, etc. about whether *p*" (6931). Real doubt is thus a "two-step model involving the intervention of two affective states: (1) an emotional episode of epistemic anxiety signaling that the proposition involved implies a possible threat, or possible negative outcomes and (2) a feeling of uncertainty signaling the lack of epistemic safety of a belief in a proposition" (6931).

Vazard argues that real doubt is so cognitively costly that it is adaptive for us to experience it only in the face of high practical stakes (6921). Thus it is crucial that the threats that trigger epistemic anxiety are not themselves epistemic, but are threats to one's practical interests. In the rain example, it is the practical threat of drought that triggers the farmer's epistemic anxiety, and thus makes her doubt 'real' where mine is only 'paper'. Then Vazard's epistemic anxiety does *not* have epistemic risk as its formal object. Vazard does not, in fact, have much to say about what is the formal object of epistemic anxiety. Following Charlie Kurth (2015: 5), she holds that the formal object of anxiety in general is "problematic uncertainty", which in turn she defines as "potential negative outcomes (implied by some particular event or situation) over which we lack information" (6922). This is similar to how I understand risk: one faces a risk where one faces the robust possibility of a negative event obtaining. However Vazard explicitly adds that one must lack information about whether the possibility will obtain. We can thus say that, for Vazard, the formal object of anxiety is risk *plus* uncertainty.

But Vazard does not specify a distinct formal object for epistemic anxiety. She gives some examples of possible material objects of epistemic anxiety, writing that instances of epistemic anxiety "have as object a certain state of affairs which can be expressed by a proposition such as 'the bank will not be open on Saturday morning' or 'the train does not stop at Foxboro', where this possibility is evaluated as implying a possible threat" (6922). A 'possible threat' is defined as "potential practical costs" (6923). But note that the risk-possibilities denoted in these propositions are not epistemic risk-possibilities: they are possibilities in which, for example, *the bank is not open on Saturday morning*, rather than in which *S has a false belief that the bank is open on Saturday morning*. In the rain example, the farmer's epistemic anxiety is about the practical risk event of drought; this is its material object. Then for Vazard, the material objects of episodes of epistemic anxiety are bog-standard risk events, not epistemic risk events. This suggests that, for Vazard, the formal object of epistemic anxiety is just the same as the formal object of anxiety: risk plus uncertainty.

If this is so, then it is unclear in what sense Vazard's epistemic anxiety is a distinctly epistemic form of anxiety. The anxiety the farmer feels when her crops are threatened by drought is simply anxiety about some non-epistemic event obtaining. Though this anxiety has an epistemic element, this is only the epistemic element shared by all emotional episodes of anxiety, on Vazard's account: one lacks information about whether the risk-possibility will obtain, thus it is represented as uncertain. And note too that the kinds of behaviours that the farmer's anxiety would motivate are not the kinds of epistemic risk-reduction behaviours that epistemic anxiety is supposed to motivate: the farmer would be motivated to, for example, gather water so that she can water her plants, or invest in a better irrigation system. It is not plausible that she would be more motivated than I would to engage in *epistemic* risk-reduction behaviours: behaviours aimed at reducing the relevant epistemic risk. In this case, the relevant epistemic risk is the risk of false belief about whether it will rain tomorrow. I suggest that I would be just as motivated as the farmer to engage in behaviours aimed at reducing this risk: I, just like the farmer, would suspend judgement about whether it will rain tomorrow until I had gathered more evidence, for example by checking the weather forecast. It is therefore unclear what work epistemic anxiety, on Vazard's understanding, could do that anxiety couldn't do just as well. This provides us with one reason for preferring my account of epistemic anxiety to Vazard's: Vazard's account is insufficiently specific to distinguish it from anxiety in general, thus to show what is distinctly valuable about epistemic anxiety.

Vazard's picture of epistemic anxiety is in another way insufficiently general. As epistemic anxiety is, for Vazard, an emotional response to threats to one's practical interest, it cannot get an explanatory grip in cases where one's practical interests are not threatened. But there are cases in which it is plausible that a subject has the same kind of emotional experience, and is motivated to engage in the same kind of epistemic risk-reduction behaviours, as in cases of 'real doubt', but where there is no threat to her practical interest. Nagel discusses a case like this, in which a subject is considering the possibility that what looks like a red table to her may in fact be a white table illuminated by red trick lighting. Nagel writes that this subject "would glance up to check [the lighting] prior to making a judgement about the colour of the table" (2010b:

303).[20] Here, the same behaviour is triggered as in a high-stakes case: the subject feels the need to gather more evidence before forming her belief. But in this case, there isn't anything practically at stake for the subject regarding the colour of the table. She is motivated to check the lighting simply because considering the possibility that it is illuminated by red trick lighting has "supplied [her] with some concerns about error" (301). Vazard cannot allow that these concerns are the manifestation of epistemic anxiety. This is a shame, as the emotional experience the subject has in this case seems the same in all important respects as does Vazard's subject undergoing 'real doubt': both subjects feel the need to, and consequently are motivated to, gather more evidence before forming a belief.

My account of epistemic anxiety does better than Vazard's on both counts. It is specific enough to distinguish epistemic anxiety from anxiety in general. On my account, epistemic anxiety is a subspecies of anxiety whose formal object is epistemic risk: the risk of an epistemically disvaluable event obtaining. (The material object of an emotional episode of epistemic anxiety will be the risk of some particular epistemically disvaluable event obtaining, for example a particular event of a subject forming a false belief, or a particular event of failing to form a true belief, and so on.) And it is general enough to allow that subjects can experience epistemic anxiety in the absence of practical risk. For epistemic anxiety is triggered in the presence of epistemic risk, which needn't correspond to risk of other kinds. Indeed, in some cases, an epistemic risk event's obtaining would be practically beneficial for one. Consider for example a belief that my ticket in a fair, ten million ticket lottery is a loser: my epistemic anxiety about the risk of this belief being false prevents me from forming such a belief, though of course it would be practically very good for me if this epistemic risk event obtained – if I believed that my lottery ticket is a loser, and this belief turned out to be false.

Vazard's overall picture of epistemic anxiety and its relationship to doubt is also problematic. Recall that for Vazard, real doubt is generated by the combination of two

---

[20] See also Hawthorne's discussion of "anxiety-provoking inferences", where S knows that P and tries to infer from P that some far-fetched error-possibility does not obtain, but instead of coming to know that this possibility does not obtain, S loses her belief that P (2004, pp. 160-1). The error-possibilities needn't be such that S would be practically worse-off if the possibility obtained. They are simply possibilities in which S holds a false belief. Note that Nagel cites Hawthorne as giving anxiety a similar role in his epistemology as she does (2010a, p. 429, n. 2).

affective states: epistemic anxiety and the feeling of uncertainty. She writes that "these combined affective states mediate the deployment of costly cognitive strategies *constitutive of doubt* – deliberation, reasoning, etc. about whether *p*" (2021: 6931, my emphasis). Here Vazard seems to identify doubt with epistemic behaviours aimed at reducing epistemic risk: deliberation, reasoning, and so on. This entails that someone who does not undertake these behaviours does not have a (real) doubt. But this is implausible. Someone who feels uneasy about the possibility of a belief she has, or is about to form, being false might not engage in these epistemic risk-reduction behaviours for a number of reasons. She might not have the time to engage in further deliberation before she must make a decision,[21] or she might drop down dead from a heart attack before she is able to reason further.

I suggest that it is more plausible to hold that doubt comes before epistemic risk-reduction behaviour, and motivates this behaviour. (Christopher Hookway (1998), from whom Nagel takes the terminology of 'epistemic anxiety' (see Hookway 1998: 222), thinks the same; though he develops an account of doubt, i.e. epistemic anxiety, as a propositional attitude, not a questioning attitude.) Then a better picture of the relationship between doubt and anxiety is one on which doubt is *identified* with epistemic anxiety: doubt is the epistemic emotion that draws a subject's attention to an epistemic risk-possibility, and motivates her to reduce epistemic risk. The epistemic risk-reduction behaviours in which she consequently engages – evidence-gathering, careful reasoning, and so on – can then be thought of as constitutive of *inquiry*. Two advantages to this conception of the relationship between epistemic anxiety and doubt immediately present themselves. First, it shows that what might seem an exotic theoretical notion, that of epistemic anxiety, has common currency in everyday life under the name of 'doubt'. Second, this picture of doubt is more ontologically parsimonious than is Vazard's: to endorse my account, one only need countenance one epistemic emotion, rather than two distinct epistemic emotions working in tandem to create some new state.

The previous chapter ended by raising a question: is there anything in the world that is picked out by the concept *doubt* that emerges from my conceptual reverse-

[21] It is important to Nagel that someone who experiences epistemic anxiety can weigh up the importance of satisfying epistemic anxiety by gathering evidence, reasoning carefully and so on, against other concerns, for example the importance of coming to a decision quickly (2010a, p. 415).

engineering projects? We can now answer this question: there is. The concept *doubt* picks out the epistemic emotion of epistemic anxiety. Epistemic anxiety is an emotional response to epistemic risk. So the concept *doubt* picks out an emotion that is a response to epistemic risk. The object of *doubt*, i.e. doubt, is epistemic anxiety. Instances of doubt, i.e. doubts, are instances of epistemic anxiety.

At this point, the reader may have the following worry. Doubt – the object of *doubt* – is a questioning attitude, and epistemic anxiety is an emotion. If doubt is to be identified with epistemic anxiety, they must be the same kind of thing. Can emotions be questioning attitudes? Yes, they can. Curiosity is standardly understood as an emotion (see for example Brun, Doğuoğlu, & Kuenzle 2008: 3; Tanesini 2008: 77; Thagard 2008: 168; Morton 2010: 386), as is wonder (Prinz 2004: 85; Thagard 2008: 168; Morton 2010: 386; Ekman & Cordaro 2011: 365; Brady 2013: 123). Both curiosity and wonder are questioning attitudes, too (see §3.1 of the previous chapter). So emotions can be questioning attitudes. Further, recall that I, following Groenendijk and Stokhof (1984), understand questions as partitions on possibility space. Then to take a questioning attitude is to take an attitude to possibilities. Epistemic anxiety is likewise an attitude we take to possibilities: epistemic risk-possibilities. So doubt and epistemic anxiety are the same kind of thing, thus it is not problematic to identify one with the other.

5. Evaluating anxiety, evaluating doubt

A theoretical advantage of identifying doubt with epistemic anxiety is that it allows us to evaluate doubts as we evaluate anxieties in general. This, in turn, can shed light on debates over the propriety of certain kinds of doubts. In this section, I spell out how this framework allows us to evaluate doubts, then apply this method for evaluating doubts to two kinds of doubts: sceptical doubts and (un)reasonable doubts in the law.

5.1. Evaluating doubts

As noted in §2.1, anxiety can be evaluated in terms of its fittingness and its utility. Anxiety is fitting when it is a response to genuine risk. Anxiety is useful insofar as it helps one to avoid or reduce the relevant risk, and the extent to which it does this. Recall the neurosurgeon Henry Marsh, who is guided by his anxiety when performing surgery: when he starts getting anxious, he stops the surgery. Marsh's anxiety is both

fitting and useful: it is fitting because it is a response to a genuine risk of damaging healthy brain tissue; and it is useful because it helps him to reduce this risk by motivating him to stop surgery. In contrast, the anxiety suffered as part of generalised anxiety disorder is both unfitting and useless: it is unfitting because it is either undirected or directed at everyday situations that do not carry with them genuine risk; and it is consequently useless, too, as it doesn't enable one to reduce or avoid the relevant risk, since there is no genuine risk present.

Applying this picture to the epistemic realm, we can say that doubt (i.e. epistemic anxiety) is fitting when it is a response to genuine epistemic risk, and unfitting otherwise; and that it is useful insofar as it helps its experiencer to reduce or avoid the relevant epistemic risk, and the extent to which it does this. Call doubts that are both fitting and useful 'proper' doubts, and call doubts that are both unfitting and unhelpful 'improper' doubts. Recall Lottie from §3.1, whose epistemic anxiety prevents her from forming the belief that her ticket in a fair lottery is a loser, based only on the odds. Lottie's experience of epistemic anxiety constitutes a proper doubt. For her doubt (i.e. her epistemic anxiety) is fitting: a belief that her ticket is a loser is very veritically epistemically risky, in terms of both modal and normic risk, as there is both a close world and a normal world in which this belief is false. And her doubt is useful, because it prevents her from forming this epistemically risky belief, thus enabling her to avoid the relevant epistemic risk. Lottie's doubt is therefore proper.

As an example of an improper doubt, consider another subject, Nottie, who also has a ticket for a fair lottery with ten million tickets. However, this lottery has already been drawn, and Nottie's ticket was a loser. Nottie watched the draw live on TV, and saw a different set of numbers to hers get drawn; the next day, she read the newspaper, which confirmed those numbers were drawn. Nevertheless, she doesn't form the belief that her ticket is a loser. For she can't help but take seriously the possibility that her hated co-worker hacked her TV to show a fake lottery draw, then mocked up several copies of the newspaper and hid them in Nottie's local newsagent's, to make her believe that her ticket is a loser. Nottie has no evidence that her co-worker did this, nor even that he hates her as she hates him. In fact, Nottie's co-worker did no such thing, doesn't know that Nottie hates him, and rarely thinks about her.

Nottie's doubt is an instance of epistemic anxiety that is both unfitting and useless. It is unfitting because the possibility in which Nottie falsely believes that her ticket is a loser because her neighbour hacked her TV and placed fake newspapers in

her newsagent's is not genuinely risky: this possibility couldn't easily obtain; it would call out for a huge amount of explanation, given Nottie's evidence, which doesn't include that her co-worker hates her, is plotting against her, and so on; and it is probabilistically unlikely both that her co-worker would carry out this plot against her, and that her lottery ticket would be the winner. As such, there is no genuine veritic epistemic risk in either the modal, normic or probabilistic sense. So Nottie's doubt is useless, too: it prevents her from forming a belief that would not be veritically risky, so could constitute knowledge. Thus Nottie's doubt is improper.

5.2. Sceptical doubts

On this way of evaluating doubts, the sceptic's doubts do badly along both axes of evaluation: they are both unfitting and useless. Suppose that a sceptic doubts whether she has hands, because she is very troubled by the possibility that she is a brain in a vat. This doubt prevents her from being able to form the belief that she has hands. Her doubt here is an instance of epistemic anxiety that is both unfitting and useless. It is unfitting because it isn't a response to genuine epistemic risk, on either the modal, normic or probabilistic account. The relevant epistemic risk is the risk of her falsely believing that she has hands, when in fact she is a brain in a vat. This epistemic risk-possibility is very modally distant, given that the sceptic is not a brain in a vat, but a full-bodied human. This possibility is also very abnormal: the obtaining of this possibility would call out for a huge amount of explanation, given the sceptic's evidence (who put her in the vat? For what reason? And so on). It is hard to imagine how we might go about calculating the evidential probability that the sceptic is a brain in a vat, given her evidence; but presumably it is very low. As such, her doubt is not fitting: there is no genuine epistemic risk in this case. So her doubt is not useful, either: because there is no genuine epistemic risk regarding the question whether she has hands, in suspending judgement on this question, she fails to reduce or avoid epistemic risk. In fact, her doubt is worse than useless: it prevents her from forming beliefs that would be knowledge-constituting. Her doubts make her epistemically worse-off than non-sceptics: she fails to have knowledge that less – or, rather, more appropriately – epistemically anxious subjects can have.

This gives us a way of precisifying the charge that has been levied against sceptics throughout the ages: that they are mad. For example, Descartes writes that if he were to "deny that these hands and this body is mine", we could appropriately

compare him to "certain persons, devoid of sense, whose cerebella are so troubled and clouded by the violent vapours of black bile, that they constantly assure us that they think they are kings when they are really quite poor, or that they are clothed in purpose when they are really without covering". In short, if Descartes were to deny this, he would be "insane" (2017: 16). Schopenhauer writes that any "genuine conviction" that our representations of the external world are unreal "can only be found in a madhouse: accordingly, it should be treated with medication, not refutation" (2010: 129). More recently, Williamson has written that "[s]cepticism is a disease in which healthy mental processes run pathologically unchecked. Our cognitive immunity system, designed to protect our conception of the world from harmful errors, turns destructively on that conception itself" (2005: 681). The common charge in these statements is that sceptics are mentally unwell, and their doubts are consequently pathological.

We can now make the objection more precise: sceptical doubts are pathological doubts because they are unfitting and useless – they are improper doubts. Further, given just how modally distant, abnormal and improbable are the error-possibilities that trouble the sceptic, her doubts are extremely unfitting; and, as noted, they are beyond useless, instead being positively damaging to her, epistemically speaking. So her doubts are not just a little improper, they are radically improper. Because of their radical impropriety, her doubts are unable to guide her practices of inquiring in the way that proper doubts can. When genuine epistemic risk is present in her environment, she won't be motivated to inquire in order to reduce that risk, as her habits of doubting won't flag the relevant epistemic risk-possibility to her as more pressing than the not-at-all-pressing sceptical risk-possibilities that disturb her. Her epistemic anxiety, in its over-responsiveness to insignificant epistemic-risk possibilities, does not kick in as it ought to when genuine epistemic risk is present. To continue with Williamson's metaphor (and having now done enough to precisify this objection, I see no harm in doing so): her cognitive immunity system is chronically inflamed, triggering costly defence mechanisms to unreal threats, and leaving her worse off than she would have been had her system not mounted the attack.

5.3. Reasonable doubts in the law

Another application of this way of evaluating doubts is to the notion of reasonable and unreasonable doubts in the law. One project in the philosophy of law attempts to clarify

legal notions, for example legal standards of proof, in terms of epistemological notions like sensitivity (Enoch, Spectre and Fisher 2012), normalcy (Smith 2018), safety (Pritchard 2018) and relevant alternatives (Gardiner 2019). The criminal standard of proof operative in the U.K. and U.S., as well as most other anglophone countries, Germany, Sweden, Italy and Israel, requires for conviction to be permissible that the defendant's guilt is proved beyond reasonable doubt. Epistemologists interested in using epistemological notions to clarify legal standards of proof have tended to interpret the 'beyond reasonable doubt' standard in such a way that does not make reference to doubts. For example, Smith argues that some legal standard is met with respect to some proposition P only if P is "normically supported by the evidence – only if the evidence makes the falsity of that proposition less normal, in the sense of calling for more explanation, than its truth" (2018: 1209-1210). The 'beyond reasonable doubt' standard requires that this condition is met, plus that not-P is either very probabilistically unlikely or very abnormal (1211). In either case, reference to 'doubt' drops out of the 'beyond reasonable doubt' standard. Similarly, Pritchard holds that the criminal standard of proof is met with respect to P only if, given the evidence, "it wouldn't be an easy possibility" that not-P (2018: 117). Again, this standard does not mention 'doubt'.

In essence, these epistemologists attempt to clarify what it is the criminal standard of proof requires by offering a roughly functionally equivalent standard of proof from the domain of epistemology: rather than saying that the criminal standard of proof is met with respect to some proposition P when P is established *beyond reasonable doubt*, these epistemologists say that the criminal standard of proof is met with respect to P when *not-P would be very abnormal, given the evidence*, or *not-P is not easily possible*. These standards are roughly functionally equivalent insofar as they issue roughly the same verdicts. But an alternative way that epistemology can clarify legal notions, such as legal standards of proof, is not by wholesale replacing those notions with epistemological ones. Instead, epistemologists can simply offer clear accounts of the notions with which legal theorists and practitioners are concerned. My account of doubt as epistemic anxiety provides us with the resources to do this for the 'beyond reasonable doubt' standard. Instead of replacing the 'beyond reasonable doubt' standard with an alternative standard that we argue to be roughly functionally equivalent, and which makes reference only to epistemological notions, we keep the

'beyond reasonable doubt' standard as it is, and offer an account of what it is for a doubt to be reasonable or unreasonable.

Here is an example of an unreasonable doubt in the law, from a first-hand account of jury deliberation in a shoplifting case:

> … in the jury room, one man reacted to the concept of 'reasonable doubt' as if it were a challenge to his ingenuity. It meant, he insisted, that we were to see if we could think of any possible alternative explanation of events, and he could – somebody had 'planted' the garment in the girl's bag. It was pointed out to him that even the defence had not put forward this explanation. This didn't matter, he said, perhaps they hadn't thought of it. There *must* be reasonable doubt if you could construct another theory, after all, it wasn't physically impossible, was it? It was now pointed out to him that although it wasn't physically impossible, his explanation was not based on a single scrap of evidence. Who did he think had done the 'planting', the store detective? 'A person or persons unknown,' said the odd man out, proudly. (Barber and Gordon 1976: 76, quoted in Ho 2008: 153)

The juror relaying this story clearly takes this doubt to be one that needn't be eliminated before the 'beyond reasonable doubt' standard of proof can be met. Legal theorists and practitioners would surely agree. We can appeal to my framework for evaluating doubts to explain why this is. This doubt is unreasonable because it is unfitting: it is not a response to genuine epistemic risk.

Consider first normic risk. If the possibility that the doubting juror describes were to obtain, substantial explanation would be required: who planted the garment? Why wasn't this brought up by the defence? Then the juror's doubt is about some abnormal epistemic risk-possibility. As such, there is a low normic risk that a verdict that the defendant stole the garment would be false for the reason that the garment was, in fact, planted in her bag. The juror's doubt is thus unfitting. Consider next probabilistic risk. The evidence that the jurors have does not support that the garment was planted in the defendant's bag (the author writes that his fellow juror's "explanation was not based on a single scrap of evidence"). So it is not likely, relative to the jurors' evidence, that a verdict that the defendant is guilty would be false, for the reason given by the doubting juror. As such, there is a low probabilistic risk that the verdict would be false. Again, the juror's doubt is unfitting. Finally, consider modal risk.

Supposing that the garment was not planted in the defendant's bag, this is not something that could very easily happen: it requires that someone is conspiring against the defendant, that this conspirator is skilled at slipping things into other people's bags without them noticing, and so on. It is unlike the easy possibility of a certain set of balls falling out of a lottery machine (recall Pritchard's first bomb case (2015: 441), discussed in §2.2). So there is not a high modal risk that a guilty verdict would be false for the reason that the garment was planted in her bag. Once again, the juror's doubt is unfitting.

More generally, we can say that reasonable doubts are fitting doubts, and unreasonable doubts are unfitting doubts. The doubts that the fact-finder (the jury in a trial by jury and the judge in a bench trial) must rule out through deliberation on the admissible evidence in order for some proposition P to be established 'beyond reasonable doubt' are those not-P possibilities that are sufficiently close, normal, or probabilistically likely. The doubts the fact-finder needn't rule out are those doubts that are sufficiently distant, abnormal, or unlikely. Though this way of understanding '(un)reasonable doubts' appeals to technical notions from philosophy, it does not supplant notions from legal theory and practice with these notions. Rather, these philosophical notions are used to explain the notions already in use in legal theory and practice. In this way, legal theorists and practitioners may be more amenable to understanding the 'beyond reasonable doubt' standard in the way that I suggest, which illuminates rather than replaces the existing standard, than they are to other epistemologists' attempts to clarify legal notions by appeal to epistemological notions.

5.4 Norms on starting inquiry

In §3.4 of Chapter 4, I said that I don't think that the following is an epistemic norm:

> **Required to Inquire if Doubting (RID):** S is required to inquire into a question Q if S doubts Q.

We are now in a position to see why. Not all doubts are proper doubts: some are unfitting and some are useless. We ought not inquire if we have improper doubts. So if there is an epistemic norm governing when we should start inquiring, it should be as follows:

> **Required to Inquire if Properly Doubting (RIPD):** S is required to inquire into a question Q if S properly doubts Q.

RIPD, like RID, doesn't involve a biconditional. That is, RIPD doesn't say that S is required to inquire into Q if and only if S properly doubts Q (for the biconditional equivalent of RID, remove 'proper'). RIPD tells us that if S has a proper doubt regarding a question Q, then S ought to inquire into Q. It doesn't tell us that if S doesn't have a proper doubt regarding Q, then S ought not to inquire into Q. As I argued in the previous chapter (§3.3), a conditional norm is all that is needed to ensure that *doubt* can play the role of the inquiry-starter.

6. Explicating *doubt*

In the previous section, I extolled the virtues of identifying doubt with epistemic anxiety. In this section, I argue that further virtues are found if we explicate the concept *doubt* by replacing it with the concept *epistemic anxiety*. In particular, *epistemic anxiety* is more exact than *doubt*, in that there are cases in which it is unclear whether *doubt* applies, but clear whether *epistemic anxiety* applies. I argue (§6.3) that the explication of *doubt* that replaces this concept with *epistemic anxiety* satisfies all relevant criteria of adequacy articulated by Carnap (1950). I end (§6.4) by considering what name we should give to the explicated concept, and by addressing the topic-preservation challenge to conceptual engineering as it applies to my explication of *doubt*.

6.4 *Doubt* as an inexact concept

In the previous chapter, I articulated the concept *doubt* that emerged from my conceptual reverse-engineering projects as follows:

> If a subject S doubts a question Q, then typically the following conditions will hold:
>
> 1. S has a questioning attitude to Q,
> 2. S does not believe any complete answer to Q,
> 3. S's situation with respect to Q is epistemically risky, or is represented to her as epistemically risky,
> 4. S is immediately motivated to inquire into Q.

This is a characterisation of the typical case in which the concept *doubt* applies to a subject. Conditions (1) to (4) are not intended as individually necessary and jointly sufficient conditions for the application of *doubt*. As such, we can expect there to be cases where the concept *doubt* intuitively applies to some subject S with respect to a

question Q, but one or more of these conditions don't hold. For example, we naturally describe the sceptic who is genuinely troubled by the possibility that she falsely believes that she has hands when she is in fact a brain in a vat as doubting whether she has hands. But we wouldn't expect her to inquire into this question – how would she go about such an inquiry, anyway? In this case, we have the intuition that the concept *doubt* should apply to the sceptic with respect to the question whether she has hands, yet she doesn't satisfy (4).

Similarly, there may be cases in which all of (1) to (4) hold of a subject S with respect to a question Q without *doubt* intuitively applying to S. Imagine a cat that is looking at some object, say, a toy bird. The cat moves closer to the bird, sniffs it, walks around it. We could naturally describe this cat as *curious* about what the object is (cf. Carruthers 2018: 133). We could also, perhaps less naturally, describe the cat as inquiring into what the object is: for the cat is engaged in an activity whose aim is to settle the question, *what is this object?* Further, we can suppose that the cat is engaged in this activity – that it is inquiring – precisely because it is curious about this question. At present, the cat doesn't believe any answer to this question. But on the functional view of belief discussed in §3.2. of the previous chapter, the cat's inquiry might lead it to believe some answer to the question: for example, that the object is a dead bird. Since the object is not a dead bird, however, but a toy bird, the cat would thereby form a false belief. We might suppose that there is a high risk of the cat forming the belief that the object is a dead bird (for example, we can suppose that there is a close world in which the cat forms this belief). Then the cat's situation with respect to the question 'What is this object?' is epistemically risky. Then all of (1) to (4) are true of the cat with respect to this question:

1. The cat has a questioning attitude towards the question 'What is this object?', namely curiosity,
2. The cat doesn't believe any answer to this question,
3. The cat's situation with respect to the question is epistemically risky,
4. The cat is immediately motivated to inquire into this question (that is, it inquires into the question because it is curious, not because it has an independent desire – a desire independent of this experience of curiosity – to inquire into the question).

But intuitively, it seems odd to describe the cat as doubting this question. This isn't just because the word 'doubt' doesn't prefix to 'what' questions with ease (see §1.1 of

Chapter 4), as the following sounds equally odd: The cat has doubts about what the object is.

In response to this case, one could simply insist that the cat does have the questioning attitude of doubt towards the question 'What is this object?'; after all, it satisfies (1) to (4) with respect to this question. This sounds weird, admittedly, but perhaps this is because our intuitions about whether and under what conditions creatures doubt are muddied by the fact that 'doubt' is polysemous in the way discussed in §1.1 of the previous chapter: it picks out both a propositional attitude and a questioning attitude, and what it takes to have each of these attitudes is different. Alternatively, we could draw attention to the atypicality of this case, for present purposes. I wanted to reverse-engineer *our* concept *doubt* – to understand this concept in order to what it does for *us*. Cats are not 'creatures like us'. Creatures like us are human. Then we cannot infer anything about our concept *doubt* by looking to the activities of cats. Cases in which the epistemic subject under discussion is a cat are not typical cases of our dealings with the concept *doubt*. As such, it needn't worry us if it is the case that conditions (1) to (4) are true of the cat, yet it is inappropriate to apply *doubt* to the cat; this simply isn't a typical case, and so not one from which we can infer anything about *doubt*.

But one thing this discussion reveals is that the concept *doubt* that emerged from the conceptual reverse-engineering projects of the previous chapter is inexact in just the ways that concerned Carnap. To begin with, the word picking out the concept is ambiguous, because the word 'doubt' is polysemous, and polysemy is a kind of ambiguity. Then simply in distinguishing the concept *doubt* that picks out a questioning attitude and the concept *doubt* that picks out a propositional attitude, as I did in §1.1 of the previous chapter, I embarked on the first stage of explication as identified by Brun: identifying the *explicandum* concept, the concept to be explicated, as clearly as possible, by disambiguating the word used for the concept (2016: 1215). Even having so disambiguated 'doubt', the concept *doubt* that picks out the questioning attitude remains inexact, in that it has borderline cases: cases in which the concept neither clearly applies nor clearly fails to apply, such as the cat case.

The method of conceptual reverse-engineering does not give us clear rules for what to do in these cases. As the cat case is arguably a "freakish" one (Craig 1990: 14), perhaps we needn't worry about our intuitions diverging from the verdict issued by the explicit concept in this case. If such a divergence took place over a typical case, we

ought to be very worried indeed, as we'll have been given reason to think the original hypothesis about the function of the concept was false (see Chapter 3, §2.2). But what are we to do in cases that are neither freakish nor typical, but in which the intuitive concept and the concept output by the conceptual reverse-engineering process issue different verdicts? No guidance is given.

We can expect there to be many such cases: cases that are neither freakish nor typical, in which our intuitions about whether *doubt* applies diverge from the verdict issued by conditions (1) to (4). Perhaps the case of the sceptic just discussed is one such case: intuitively, she falls under the concept *doubt* with respect to the question whether she has hands, but condition (4) is not true of her. Here is another case. I can't remember whether Phil or Grant is the older Mitchell brother, but I am inclined to think that it is Grant. Still, I'm not sure, so I don't form a belief that Grant is the older Mitchell brother – though if I had felt a smidgen more confident, I would have formed the (false) belief that Grant is the older Mitchell brother. It is natural to describe me as doubting whether Grant or Phil is the older Mitchell brother. (Note that this doubt can't be rephrased as a propositional doubt: I don't doubt that Grant or Phil is the older Mitchell brother. I both believe and know that Grant or Phil is the older Mitchell brother.) But I don't care to inquire into the matter. So it seems that the concept *doubt* applies to me with respect to the question whether Grant or Phil is the older Mitchell brother, yet (4) is not true of me. This is thus a borderline case for the concept *doubt*: a case where it is not clear whether the concept should apply. And conceptual reverse-engineering gives us no guidance about which way to go.

6.5 Replacing doubt with epistemic anxiety

I suspect that the concept *epistemic anxiety* will have fewer borderline cases than does *doubt* – fewer cases in which it is unclear whether the concept applies – because *epistemic anxiety* can be more exactly characterised than *doubt*. *Epistemic anxiety* can be characterised thus:

A subject S is epistemically anxious about some possibility iff S is having an emotional experience with the following profile:

(a) The emotional experience represents this possibility to S as epistemically risky (i.e., as epistemically disvaluable and either modally close, normal, or probable),

(b) The emotional experience is affectively unpleasant,

(c) The emotional experience motivates S to take steps to reduce or avoid the risk of this possibility's obtaining.

In each of the cases just considered, we get a clear verdict about whether or not the subject is epistemically anxious. We can say that the sceptic is epistemically anxious about the possibility that she is a brain in a vat, so falsely believes that she has hands, if and only if (a) she is having an emotional experience that represents this possibility to her as epistemically risky, (b) which is affectively unpleasant, and (c) which motivates her to take steps to reduce or avoid the risk of the possibility's obtaining.

I said that the sceptic is "genuinely troubled" by the possibility that she falsely believes that she has hands when she is a brain in a vat; then (a) and (b) are true of her: she is having an unpleasant affective experience (being troubled) about some possibility in which an epistemic risk-event obtains (false belief). And she is consequently motivated to suspend judgement on whether she has hands, thereby avoiding this risk. So she is epistemically anxious regarding this possibility. The cat isn't epistemically anxious regarding the question what the object is, as it isn't having an unpleasant affective experience. Rather, it is curious about what the object is, and curiosity is standardly understood as affectively pleasant (Day 1971; Kashdan, Rose and Fincham 2004; Kang, Hsu, Krajbich, Loewenstein, McClure, Wang, Camerer 2009). So too for me and the Mitchell brothers: I'm not epistemically anxious because I don't have an unpleasant affective experience that represents any possibility to me as epistemically risky, for example the possibility in which I fail to form any belief on the matter (i.e., the possibility that actually obtains).

We thus have reason to explicate the concept *doubt* by replacing it with the concept *epistemic anxiety*: namely, that *epistemic anxiety* is more exact than *doubt*, in that there are fewer cases in which it is unclear whether the concept applies. In the next section, I will argue that this replacement would satisfy the first three of Carnap's criteria of adequacy for explication (first discussed in §3.2 of Chapter 2), with the fourth being inapplicable:

1. Similarity,
2. Exactness,
3. Fruitfulness,
4. Simplicity.

I will discuss each criterion in turn, and show how this explication of *doubt* as *epistemic anxiety* satisfies the criterion.

6.6 Carnap's criteria of adequacy for explication

The first criterion requires that the *explicandum* (the concept to be explicated) and the *explicatum* (the concept to replace the *explicandum*) must be similar, in the sense that the *explicatum* can be used in place of the *explicandum* in relevant contexts (Carnap 1950: 5). I will argue that *epistemic anxiety* can be used in place of *doubt* in the context most relevant for our purposes: it can play the role of the inquiry-starter. Recall from §2.1 of the previous chapter the two requirements an inquiry-starter concept should meet. First, there should be something intrinsically wrong with a situation in which a subject S falls under the concept with respect to some question Q, yet fails to inquire into Q. Second, S's being in the state picked out by the concept with respect to Q should motivate S to inquire into Q.

Immediately, a worry arises: there are cases in which a subject is epistemically anxious regarding some question Q, but doesn't inquire into Q, and there seems nothing wrong with her failing to inquire into Q. Consider the sceptic who is experiencing epistemic anxiety regarding the question whether she has hands (because she is troubled by the possibility that she is a brain in a vat). She is not motivated to inquire into the question whether she has hands, but to suspend judgement on the question; and suspending on this question seems more appropriate than inquiring into it. Lottie's case (§3.1) is also like this: Lottie's epistemic anxiety regarding the question whether her lottery ticket is a loser prevents her from forming a belief that her ticket is a loser; but she isn't motivated to inquire into, but rather to suspend judgement on, this question; and again, suspension seems like the most appropriate course of action.

Here's what I want to say in response to this worry. The first requirement of the inquiry-starter concept is that there should be something wrong with a situation in which the concept applies to some subject with respect to a question, but the subject doesn't inquire into the question. There are two ways to rectify this wrongness: either the subject can inquire into Q, or she can cease to be in the situation. Suspending judgement on one's question will sometimes bring an end to one's epistemic anxiety, by virtue of allowing one to avoid the relevant epistemic risk. In the lottery case, once Lottie suspends on whether her ticket is a loser, she is no longer at risk of forming a false belief – she isn't going to form any belief in this situation. This new situation is not one in which Lottie is epistemically anxious yet failing to inquire: once Lottie has

suspended on her question Q, she will cease being epistemically anxious about Q, and rightly so.

Epistemic anxiety doesn't meet the second requirement of the inquiry-starter concept, that S's being in the state picked out by the concept with respect to some question Q motivates S to inquire into Q. This is because epistemically anxious subjects are motivated to either inquire into Q or to suspend judgement on Q. Recall Lottie, whose epistemic anxiety motivates her to suspend judgement on the question whether her ticket is a loser, and the sceptic, whose epistemic anxiety motivates her to suspend judgement on whether she has hands. But suppose that we slightly tweak the second requirement. Instead of saying that the inquiry-starter concept must be such that S's being in the state picked out by the concept with respect to Q motivates S to inquire into Q, we can say that S's being in the state picked out by the concept with respect to Q motivates S *either* to inquire into Q *or* to cease being in the state picked out by the concept.

Epistemic anxiety, I argue, will meet the second requirement, so tweaked. For epistemic anxiety (the state picked out by *epistemic anxiety*) is an emotion, and emotional experiences are generally understood to be passive states: it is out of our direct control whether or not we are in them (James 1884: 189-190; Helm 2001: 34, 66, 74; Goldie 2004: 58; Prinz 2004: 71-2; Tappolet 2016; though of course this picture is not universally endorsed: see Solomon 1993 and Slaby and Wüschner 2014 for a more active view of emotions). Emotions are responses to stimuli. If we are presented with the relevant stimulus, and our emotional apparatus is functioning normally, we will experience the corresponding emotion. Emotions are, however, under our *indirect* control, in that we can voluntarily bring about situations in which we will be presented with stimuli that we can expect to trigger this emotion or that, or bring an end to an ongoing situation in which we are presented with such a stimulus. Regarding epistemic anxiety: we cannot directly control whether we experience epistemic anxiety in response to epistemic risk. But we can indirectly control whether we experience epistemic anxiety, by either bringing about, or bringing an end to, an epistemically risky situation. Suspending judgement on a question Q is a way of bringing an end to an epistemically risky situation: if one's situation with respect to Q involves veritic epistemic risk, then suspending judgement on Q eliminates the risk, as one is not at risk of forming a false belief if one is not going to form any belief.

To tweak the second requirement in this way makes it sit better with the first requirement, given how we are now understanding it. The first requirement of the inquiry-starter concept is that there should be something intrinsically wrong with a situation in which the concept applies to some subject S with respect to Q, but S doesn't inquire into Q. We saw two ways that S can rectify this wrongness: either S can inquire into Q, or she can cease to be in the situation. The second requirement, tweaked as above, says that being in the state picked out by the concept must motivate S to do exactly that: to inquire into Q, or to cease to be in that state (by, for example, suspending judgement).

But does tweaking the second requirement in this way undermine the efficacy of the inquiry-starter concept? It looks like what I'm now saying is that all that we require of the inquiry-starter concept is that being in the state picked out by the concept must motivate one to either: inquire or not inquire. But that's not right. For one thing, there is still the normative requirement on the inquiry-starter: for as long as the concept applies to S with respect to a question Q, something is wrong with S failing to inquire into Q. For another thing, ceasing to be in the state picked out by the inquiry-starter – that is, ceasing to be epistemically anxious – is not as easy as deciding not to inquire into some question. Much of the time, this will not destroy the relevant epistemic risk. For example, if the relevant epistemic risk is failing to form a belief in the true answer to Q, then deciding not to inquire into Q will not enable one to avoid this epistemic risk. As such, one will still experience epistemic anxiety, thus will still be in the state picked out by the inquiry-starter concept. In order to cease being epistemically anxious, one must do something that avoids the relevant epistemic risk. This will destroy the stimulus for one's epistemic anxiety, thus allowing one to cease being epistemically anxious.

In the previous chapter (§2.1), I motivated our need for an inquiry-starter concept by noting that we need to inquire in order to get true beliefs, which have survival value; that inquiry has no natural starting-point, in that we can inquire into pretty much any question we like; but we ought not inquire into any question we like, given our finite time and resources; thus we need to prioritise our time and resources into inquiring into those questions that matter, for one reason or another. Thinking in terms of epistemic risk gives us a way of pinning down what are those questions that matter: questions regarding which subjects are in epistemically risky situations. For example, questions about which some subject wants to know the answer, so is at risk

of missing out on knowledge of its answer; or questions about which a subject is at risk of falsely believing an answer. But now we can see that, regarding veritic epistemic risk, inquiring is not the only way to avoid the risk. Suspending judgement works too. So the tweaked second requirement, I hold, is all we need of the inquiry-starter. As such, *epistemic anxiety* can play the role of the inquiry-starter, as it meets both requirements as we are now understanding them. Thus my explication meets Carnap's adequacy condition of similarity.

The second criterion of adequacy for explication is exactness. This criterion says that the *explicatum* must be exact, and moreso than the *explicandum*. There are two aspects to this criterion. The first is that rules for the use of the *explicatum* must be "given in an *exact* form", for example, "in the form of a definition", "so as to introduce the *explicatum* into a well-connected system of scientific concepts" (Carnap 1950: 7). Remember that Carnap's use of 'scientific' is best understood as meaning 'theoretical'. The second aspect is that the *explicatum* must be less vague than the *explicandum*, in the sense of their being fewer cases in which it is unclear whether the concept applies (1950: 5).

Epistemic anxiety meets both aspects of this criterion. The concept *epistemic anxiety* is defined as a concept that picks out epistemic anxiety, an emotional response to epistemic risk. The concept is thereby introduced into a network of theoretical concepts from philosophy: *epistemic risk*, *modal closeness*, *normalcy*, *probability*. Second, as we saw in §6.2, there are cases in which it is unclear whether *doubt* applies, but clear whether *epistemic anxiety* applies. In §6.1, I introduced three such cases: the case of the sceptic and the question whether she has hands; the case of the cat and the question what this object is; the case of me and the question whether Grant or Phil is the older Mitchell brother. In each case, our intuitions about whether *doubt* applies come apart from the verdict issued by the characterisation of the concept that emerged from the previous chapter, whereby a subject S who doubts a question Q will typically be such that:

1. S has a questioning attitude to Q,
2. S does not believe any complete answer to Q,
3. S's situation with respect to Q is epistemically risky, or is represented to her as epistemically risky,
4. S is immediately motivated to inquire into Q.

The sceptic and I both intuitively doubt our respective questions, but don't satisfy (4); the cat satisfies all of (1) to (4), but intuitively doesn't doubt its question. In each case, it is unclear whether we should allow our intuitions to override the verdict issued by conditions (1) to (4), so it is unclear whether *doubt* applies. But for each case, the concept *epistemic anxiety* either clearly does or clearly does not apply. The concept applies to the sceptic, because she is epistemically anxious regarding her question: she has an emotional experience that represents the possibility that she is a brain in a vat falsely believing that it has hands to her as an epistemic risk-possibility; her emotional experience is affectively unpleasant; and she is motivated to avoid this risk by suspending judgement on the question whether she has hands. The concept doesn't apply to the cat, because the cat is not having an affectively unpleasant emotional experience, so is not experiencing epistemic anxiety. And the concept doesn't apply to me, because I'm not having an emotional experience that is either affectively unpleasant, or which represents some possibility to me as an epistemic risk-possibility. So *epistemic anxiety* satisfies the adequacy criterion of exactness: the concept is defined into a well-connected system of theoretical concepts; and there are fewer cases in which it is unclear whether the concept applies than there were for the *explicandum* concept *doubt*.

Carnap's third adequacy criterion is fruitfulness, which he understands as the requirement that the *explicatum* concept must allow the theorist to formulate many laws and generalisations (1950: 6). This understanding of fruitfulness is better suited to science than to philosophy. But we have already seen that the concept *epistemic anxiety* is fruitful in the sense that it can be applied to make progress on debates within philosophy, for example, to debates over the value of sceptical doubts (§5.2), and to the philosophy of law (§5.3). So I take it that the concept *epistemic anxiety* is fruitful for philosophy.

Carnap's final criterion of adequacy is simplicity. This criterion states that if two candidate *explicata* for some *explicandum* concept meet the above three criteria of adequacy to an equal degree, then considerations of simplicity can be used to choose between them. Such considerations concern both how simple are the rules for using each *explicatum* concept, for example how simple is its definition; as well as how simple are the forms of the laws that can be formulated using the *explicatum* (Carnap 1950: 7; Brun 2016: 1215). As we are only considering one candidate *explicatum* for the *explicandum* concept *doubt*, namely *epistemic anxiety*, we don't need to concern

ourselves with this criterion of adequacy. Thus we see that the explication of *doubt* that replaces *doubt* with *epistemic anxiety* will be adequate, according to Carnap's criteria of adequacy, as it meets all applicable criteria.

6.7 Terminological ethics

If we explicate *doubt* as I have suggested, replacing this inexact concept with the more exact concept *epistemic anxiety* that I developed in this chapter, we must make a choice about what name to use for the explicated concept. We could call the explicated concept '*epistemic anxiety*', as we have been doing. Or we could retain the name '*doubt*' for the explicated concept: that is, we could henceforth call *epistemic anxiety* '*doubt*'. In this section, I will present one reason for calling the explicated concept '*epistemic anxiety*', and one for calling it '*doubt*'. This debate lies within what Peirce calls "the ethics of terminology" (CP, Book II, Chapter I): what normative reasons we have for using which expressions; in particular, names. Which words and expressions conceptual engineers are justified in using for their engineered concepts is something that concerns a number of those writing on conceptual engineering (for example Cappelen 2018; Sterken 2020; Koch 2021). I will then address the 'topic-preservation' challenge articulated in §4.1 of Chapter 2, applied to my explication of *doubt*.

A reason for using the name '*epistemic anxiety*' for the explicated concept is that 'doubt' is polysemous: we use the word for both a concept that picks out a propositional attitude and a concept that picks out a questioning attitude (see Chapter 4, §1.1). As such, speakers who use the word 'doubt', even in contexts where it is clear that what is under consideration is the concept that picks out the questioning attitude, will be primed to think about the propositional attitude of doubt. This will bring up a host of associations speakers have with the propositional attitude doubt, for example, that to doubt a proposition P involves thinking that P is probably false.[22] This can lead

[22] One might be tempted to think of this as an example of a "lexical effect" of the word 'doubt', on Cappelen's understanding (2018: 122-3). For Cappelen, lexical effects are cognitive or emotive effects that expressions can have on language-users that are not traceable to an element of the expression's semantics or pragmatics. For example, English speakers are made uneasy by the word 'niggardly', even though this has no semantic or pragmatic connection to the n-word slur (124). But this isn't quite right regarding 'doubt'. For 'doubt' is polysemous: it has multiple meanings, but these meanings are related to each other. So if language-users think of one meaning of 'doubt' in a context in which the word is used with its other meaning, this is not unrelated to the semantics of the word 'doubt'.

to confusion. In contrast, ordinary speakers are unlikely to have any associations with the expression '*epistemic anxiety*', due to its technical nature. As such, speakers are unlikely to bring in any such associations to their uses of the expression, or their interpretations of other speakers' sentences that involve the expression, in a way that would lead to confusion and miscommunication. As Peirce would say, the name '*epistemic anxiety*' is "ugly enough to be safe from kidnappers" (CP §5.414). So this is one reason to use the name '*epistemic anxiety*' for our explicated concept.

Here's a reason not to use '*epistemic anxiety*', but rather '*doubt*', for our explicated concept. As already noted in §4.2, the notion of epistemic anxiety seems to us to be a highfaluting notion with no connection to non-philosophers' concerns. It is an advantage of my account of epistemic anxiety, the epistemic emotion, that it actually has common currency under the name of 'doubt'. As the word 'doubt' is already in common use, it would encourage more widespread uptake of the explicated concept if '*doubt*' was retained for its name. It is unlikely that I could persuade legal theorists and practitioners, for example, to replace talk of the 'beyond reasonable doubt' standard with talk of a 'beyond fitting epistemic anxiety' standard.

This reason for using '*doubt*' for the explicated concept makes clear that I want my explicated concept to be put to use in existing debates, such as the debate over the value of sceptical doubts, and over what counts as a 'reasonable doubt' in the law. At such, I must address the Strawsonian challenge, as it applies to this explication. Recall from §4.1 of Chapter 2 the two objections to conceptual engineering that have been articulated under the banner of the 'Strawsonian challenge'. The first says that in changing a concept to solve a problem, advance a debate, and so on, we change the topic; thus we cannot be engaging with the very same problem, debate, and so on at the end of a process of conceptual engineering as we were when we began; so conceptual engineering cannot help us to solve problems, advance debates, and so on. The second says that concepts are individuated in terms of their intensions and extensions, so we cannot change a concept's intension or extension without abandoning the old concept in favour of a numerically distinct concept; as such, conceptual re-engineering is impossible. I called these the 'topic-preservation challenge' and the 'incoherence challenge', respectively.

As I noted there, the incoherence challenge does not apply to explication, nor to *de novo* conceptual engineering in general: in *de novo* conceptual engineering, of which explication is a sub-method, the theorist does not claim to preserve the

numerical identity of the concept to be engineered. Quite the opposite, in the case of explication: a theorist who engages in explication explicitly takes herself to be replacing the *explicandum* concept with a distinct concept, the *explicatum*. But the topic-preservation challenge does apply to explication. Insofar as the *explicatum* is supposed to be able to advance the same debates, solve the same problems, and so on, as did the *explicandum*, theorists who explicate concepts must have something to say about how topic is preserved through explication. I want my explicated concept to be put to use in existing debates, such as the debate over the value of sceptical doubts, and over what counts as a 'reasonable doubt' in the law. As I just pointed out, uptake of the new concept is much more likely to be successful if we retain the name '*doubt*' for this explicated concept. But retaining the name '*doubt*' isn't sufficient for topic-preservation – for "talking about the same thing" (Cappelen 2018: 97). What would be sufficient?

In §4.4 of Chapter 2, I argued that it is sufficient for topic preservation in a conceptual engineering project that speakers who use 'x' to refer to the pre-engineering concept can fluidly communicate with speakers who use 'x' to refer to the engineered concept. This doesn't require that 'x' has the same extension and intension for both speakers, just that the intensions and extensions are similar enough. Hence I called two speakers who can fluidly communicate using 'x' *similarsayers* with respect to 'x' (cf. Cappelen's "samesayers" (2018: 107)). I think that speakers who use '*doubt*' to pick out the *explicandum* concept *doubt* would be able to similarsay with those who use '*doubt*' to pick out the *explicatum* concept *epistemic anxiety* in relevant contexts. Both speakers will take their concepts to apply in roughly the same cases: most importantly, cases where the subject's situation with respect to a question Q is, or would seem to her to be, epistemically risky. Though there will be cases where *epistemic anxiety* clearly does or does not apply to a subject, yet unclear whether (the *explicandum* concept) *doubt* applies, so long as these cases are sufficiently infrequent, this won't preclude fluid communication for most purposes. In cases where users of the different concepts cannot fluidly communicate using '*doubt*', we should point out the theoretical advantages of the engineered concept (articulated in §6.3), and argue (with Simion and Kelp 2020; see Chapter 2, §4.5) that we should turn our back on the old concept and accept the consequence that we are unable to fluidly communicate with users of the old concept in these cases.

7. Conclusion

In this chapter, I developed an account of epistemic anxiety as an emotional response to epistemic risk. I argued that epistemic anxiety, like anxiety in general, is generated in response to two kinds of (epistemic) risk: normic and probabilistic. However, I showed that epistemic anxiety is nevertheless still valuable from the perspective of anti-risk epistemology, which must be understood as appealing to modal veritic risk. This is because epistemic anxiety motivates subjects to reduce normic veritic risk, and reducing normic veritic risk and reducing modal veritic risk are normatively coincident goals: one cannot aim for one without aiming for the other. I argued that my account of epistemic anxiety has advantages over extant accounts. It is more fleshed out than Nagel's, who says of the nature of epistemic anxiety only that it is a 'force' with a particular motivational power. My picture of epistemic anxiety as an emotion explains how epistemic anxiety has the motivational power that Nagel attributes to it. Thus my account has an explanatory advantage over Nagel's, though I hold that nothing in my account is incompatible with what Nagel says about epistemic anxiety. In contrast, my account conflicts with Vazard's picture of epistemic anxiety as an emotional response to potential practical threats in one's environment. But my account is preferable to Vazard's, as it avoids two problems facing her account: that it cannot distinguish epistemic anxiety from anxiety in general, and that it cannot capture all instances of epistemic anxiety. I argued that we should identify doubts with epistemic anxieties, and suggested two applications of doubt, so understood: precisifying the sense in which sceptical doubts are pathological doubts; and helping us to understand what constitutes a 'reasonable doubt' in the law. I then explicated the concept *doubt* that emerged from the conceptual reverse-engineering projects of Chapter 4, replacing it with my concept *epistemic anxiety*.

Conclusion

In this thesis, I have offered a novel account of the function-first approach to concepts, refined our understanding of the methods of conceptual engineering and conceptual reverse-engineering, and undertaken my own function-first approach to the concept *doubt* that makes use of these methods. I first reverse-engineered *doubt*, testing the plausible hypothesis that this concept functions to flag when inquiry should begin. I engaged in two conceptual reverse-engineering projects, one using a synchronic model and the other using a diachronic model. Both confirmed my hypothesis. The picture of *doubt* that emerged had it that this concept typically applies to some subject S with respect to some question Q when:

1. S has a questioning attitude to Q,
2. S doesn't believe any complete answer to Q,
3. S's situation with respect to Q is epistemically risky, or is represented to her as epistemically risky,
4. S is immediately motivated to inquire into Q.

I then undertook a conceptual engineering project on this concept *doubt*, a Carnapian explication, in which I replaced *doubt* with the more exact concept *epistemic anxiety*. A subject S is epistemically anxious regarding some possibility *w* iff S is having an emotional experience with the following profile:

(a) The emotional experience represents *w* to S as epistemically risky (i.e., as epistemically disvaluable and either modally close, normal, or probable),
(b) The emotional experience is affectively unpleasant,
(c) The emotional experience motivates S to take steps to reduce or avoid the risk of *w*'s obtaining.

We saw that *epistemic anxiety* can be used in place of *doubt* in relevant contexts, and has fewer borderline cases than does *doubt*. Further, *epistemic anxiety* is a fruitful concept for philosophical theorising, as it provides us with a framework for evaluating doubts as we evaluate anxieties more generally: in terms of their fittingness and their utility. The fittingness of a doubt is a matter of whether it is a response to genuine epistemic risk, and its utility a matter of whether it helps its experiencer to reduce or avoid some genuine epistemic risk, and the extent to which it does this. This allowed us to precisify the charge commonly raised against epistemological sceptics, that their

doubts are pathological (Chapter 5, §5.2), as well as to shed light on what are 'reasonable doubts' in the law (§5.3).

There are many further philosophical debates, problems, and projects in which I think my account of *doubt*, explicated in terms of *epistemic anxiety*, can be fruitfully deployed, but which I have not had the space in this thesis to explore. I'd like to end by gesturing towards three. Epistemologists are increasingly concerned with social and political issues, for example, conspiracy theories (Keeley 1999; Cassam 2016, 2019; Harris 2018; Dentith 2018, 2021; Coady 2019), vaccine scepticism (Baghramian and Croce 2021; Cassam 2021; Baghramian and Panizza 2022) and injustices perpetuated against persons in their capacities as epistemic agents (Fricker 2007; Mills 2007; Alcoff 2007; Medina 2012; Davis 2016; Luzzi 2016; Jenkins 2021; Tilton 2022). My account of doubts as epistemic anxieties can helpfully bear on these issues by providing us with a novel framework for theorising about what exactly is going wrong, epistemically speaking, with the subjects in these cases. This, in turn, opens up new avenues for engaging with these subjects, with the aim of fixing what has gone wrong.

Consider first conspiracy theorists. It is standard in the philosophical literature on conspiracy theory to understand conspiracy theorists as holding some belief or set of beliefs (see for example Keeley 1999: 112; Cassam 2016: 162; Harris 2018: 236; Dentith 2019: 2244). For example, a conspiracy theorist might believe that Lee Harvey Oswald did not act alone in the assassination of JFK (Keeley 1999: 109), that the World Trade Center attacks of September 11th 2001 were orchestrated by the U.S. government (Harris 2018: 237), or that the world is run by alien shape-shifting lizards (Dentith 2021: 9900). The conspiracy theorist's belief (or belief set) can then be evaluated as unjustified, unwarranted, or otherwise defective, insofar as it is badly supported by evidence, or simply false. But at least some of those who we might want to call conspiracy theorists need not have beliefs in the relevant propositions. For example, a conspiracy theorist might doubt whether Biden legitimately won the 2020 U.S. presidential election without outright believing that his win was a result of widespread voter fraud. Such a conspiracy theorist is surely still negatively evaluable *qua* epistemic subject. But this is not because he holds some belief that is ill-supported by evidence. He doesn't hold any such belief.

My picture of *doubt* provides us with the resources to say that the conspiracy theorist is doing badly, epistemically speaking, even though he doesn't hold any unjustified, unwarranted, or otherwise defective belief. He is doing badly, epistemically

speaking, because he doubts some question that does not involve any genuine epistemic risk, namely, whether Biden legitimately won the 2020 U.S. presidential election. Given the evidence that we have about how incredibly uncommon voter fraud is (see Minnite 2010; Levitt 2014; Bump 2016), it is extremely unlikely that Biden won due to widespread voter fraud, the closest world where this is the case is not close at all, and the most normal world in which it is the case is very abnormal. Herein lies a key benefit of applying my picture of *doubt* to the conspiracy theorist: we get a grip on what is going wrong with conspiracy theorists, even in cases where they do not hold defective beliefs.

Much the same can be said about vaccines sceptics. Regarding many high-profile vaccine sceptics, such as Andrew Wakefield (Wakefield et al. 1998), it is not clear that they do believe, for example, that the MMR vaccine causes autism, or that the COVID-19 vaccine is more dangerous than the disease itself (Putterman 2020). What matters is that they doubt – or at least try to promote doubt in others (see Oreskes and Conway 2010) – whether these vaccines are safe. Functionally, this doubt is just as effective as would be the relevant belief in doing harm, for example, by discouraging people from allowing their children to receive the MMR vaccine, which in turn could lead to more children having measles, mumps or rubella (*The Lancet* 2019). So having such a doubt about whether the MMR vaccine causes autism can lead to morally bad outcomes. My picture of *doubt* allows us to say that this is also epistemically bad. Suppose that a vaccine sceptic doubts whether the MMR vaccine is safe, because she is epistemically anxious about the possibility that it causes autism. Given the evidence that we have about the link between autism and the MMR vaccine, which overwhelmingly suggests that there is no causal connection between the two (WHO 2003), this possibility is very abnormal, very unlike the actual world, and very unlikely to be actual. So it is very low risk, on all three accounts of risk: normic, modal, and probabilistic. As such, the vaccine sceptic's doubt is unfitting: it is not a response to genuine epistemic risk. It is therefore an improper doubt.

Finally, consider a subject who, when a speaker testifies to him that she has been raped, fails to believe her, due to a background belief in what Katharine Jenkins calls the "dishonesty myth" (2021: 39): that women frequently lie about being raped. Jenkins and Emily Tilton (2022) have both independently argued that the hearer commits a testimonial injustice against the speaker in failing to believe her on the basis of this myth. On the standard picture of testimonial injustice, to commit a testimonial

injustice against a speaker is to assign her less credibility than she deserves on the basis of a prejudice about some social group of which she is a member (Fricker 2007: 28; though see Davis 2016 for an argument that hearers can commit testimonial injustices by assigning speakers *more* credibility than they deserve, due to prejudicial bias). As Jenkins notes, failing to believe someone who claims to have been raped can have a number of negative consequences: "she might find this re-traumatizing, an opportunity may be missed to prevent her rapist from committing further crimes in the future, and so on" (2021: 43). But importantly, Jenkins argues, failing to believe the speaker is a harm in itself: "the *very fact of not being believed* can be wrongful *in and of itself*, and independently of any bad consequences, if the withholding of belief stems from an identity-based prejudice – and dishonesty myths fit this criterion" (43, Jenkins's emphasis). But even though this is a wrong done to the speaker in her capacity as an epistemic agent, it is still a moral wrong: it is a harm, an injustice, and harms and injustices are moral bads.

My picture of *doubt* allows us to say that the hearer who fails to accept the speaker's testimony is also doing badly in a purely epistemic sense. Note that in this case the hearer doesn't need to believe that the speaker is lying, or that what she says is false. The harm done to the speaker lies in the hearer's simply *failing to believe* her. The hearer can fail to believe the speaker when she says that P without forming a belief that not-P. He can simply doubt whether P. In that case, we cannot say that the speaker has an unjustified, unwarranted, or otherwise defective belief. But we can say that he has an improper doubt about whether the speaker was raped, given that the dishonesty myth is not grounded in fact. As Jenkins notes (2021: 40), false allegations for rape are no higher than for other crimes, at around 3% according to Home Office statistics (Kelly et al. 2005). Given this evidence, it is unlikely that a given rape accusation is false. So there is a low probabilistic risk of the accusation being false. Further, as the hearer's evidence also includes that the speaker has testified to him that she was raped, explanation would be called for if this turned out to be false: why is the speaker lying? What does she have to gain? So the normic risk of the accusation being false is low, too. Supposing that the speaker is telling the truth, the modal risk of her accusation being false is also low. Then the hearer has a doubt that is an emotional response to a possibility that is not epistemically risky. As such, his doubt is unfitting, and therefore improper. It is a bad doubt.

In all three cases, I can say that the subject is epistemically defective in an important sense, even though s/he does not hold a defective belief. These subjects are doing badly, epistemically, because they have improper, because unfitting, doubts: doubts that are not responses to genuine epistemic risks. Thinking about these subjects as having bad doubts, rather than having bad beliefs, might change how we engage with them. For example, if we take the conspiracy theorist to believe that Biden's 2020 presidential election win was a result of widespread voter fraud, we will give him evidence that this belief is false: for example, that this election's ballots were carefully checked, and no evidence of substantial voter fraud was found (Cybersecurity and Infrastructure Security Agency 2020). The aim would be to get him to give up this belief, which may or may not involve coming to believe its negation. But if we take him to doubt whether Biden legitimately won the election because he is epistemically anxious about the possibility of widespread voter fraud, we would present him with evidence that this is not a genuine epistemic risk: we are not at risk of having false beliefs in the results of U.S. elections due to widespread voter fraud, because voter fraud in U.S. elections is generally very uncommon (Minnite 2010; Levitt 2014; Bump 2016). The aim would be to get him to give up this doubt in particular, which may or may not result in him forming a belief that Biden legitimately won the election. Insofar as he fails to give up his doubt, we will regard him as having a malfunctioning cognitive system producing his doubts: his capacity for experiencing epistemic anxiety in response to epistemic risk is overactive, or otherwise ill-attuned, generating anxiety where there is no genuine risk. A well-functioning doubting system receives epistemic risk as its input and outputs doubts (epistemic anxieties). This subject's system is malfunctioning, because it outputs doubts without an input of epistemic risk.

These subjects might be going wrong, epistemically, in other ways too. Mona Simion (2021) has recently argued that we have duties to believe that P if we have sufficient and undefeated evidence for P. This duty is grounded in proper epistemic functioning. When a subject's belief-forming processes are functioning as they ought to, then an input of sufficient and undefeated evidence for P will generate belief that P. In each of my three cases, the subject fails to believe some proposition for which they have sufficient and undefeated evidence: the conspiracy theorist that Biden legitimately won the 2020 U.S. presidential election; the vaccine sceptic that there is no causal link between the MMR vaccine and autism; and the hearer that the speaker

has been raped. On Simion's view, each of them thereby has a malfunctioning belief-forming process:

> Resistance to Evidence as Epistemic Malfunction (REEM): A subject S's belief formation process P is malfunctioning epistemically if there is sufficient evidence supporting *p* that is easily available to be taken up via P and P fails to output a belief that *p*. (2021: 3-4)

I don't see my explanation of what's going wrong epistemically in these cases as in conflict with Simion's. In fact, I think the two explanations can be applied in tandem to give us a fuller picture of the ways that these subjects are epistemically at fault. On Simion's picture, what each of these subjects have in common is that they violate an epistemic duty due to an instance of epistemic malfunctioning. But on Simion's picture, there needn't be anything common to each of them that explains why they are epistemically malfunctioning. Rather, the epistemic malfunction can be brought about by different causes, for example, "prejudice, optimism, lack of attention, partisanship, bias" (2). But if we supplement Simion's picture of the duty to believe with my picture of (im)proper doubt, we can note another unifying fault in each of these subjects, which explains why they are malfunctioning epistemically: each of these subjects takes seriously some possibility that is not genuinely epistemically risky, because it is modally distant, abnormal, and probabilistically unlikely to obtain. This can also be understood as a case of epistemic malfunctioning, but involving doubt, rather than belief: one's epistemic anxiety is triggered by possibilities that are not genuinely epistemically risky.

To wrap up: where there is a doubt to be evaluated, I've offered the means to do so. I hope this will be useful for many projects in epistemology. I also hope, more ambitiously, to have shown that doubt doesn't deserve its negative reputation. Though we can be racked, tormented, plagued by doubt, though doubts can be troubling, consuming, agonising, this is all for good reason: there's epistemic risk about. Doubt, on my picture, is epistemic anxiety: an epistemic emotion that helps us navigate epistemic risk in our environment, quickly and automatically. And the concept *doubt* serves a purpose for us that we couldn't do without: it flags when we ought to inquire. So our doubts are not "traitors", as Shakespeare says (1991, Act 1, Scene 1, line 77): they are our epistemic allies. Doubt is not the "death of the soul" (Flaubert 2001: 65): it is a prerequisite for living in the world as we do. This thesis has attempted to vindicate doubt. I hope that it has succeeded.

References

Alcoff, L. M. (2007). Epistemologies of ignorance: three types. In S. Sullivan and N. Tuana (eds.), *Race and epistemologies of ignorance*. Albany, NY: SUNY Press.

Anderson, C. (2014). Fallibilism and the flexibility of epistemic modals. *Philosophical Studies*, 167, 597-606.

Archer, A. (2018). Wondering about what you know. *Analysis*, 78(4), 596-604.

Armour-Garb, B. (2011). Contextualism without pragmatic encroachment. *Analysis*, 71(4), 667-676.

Armstrong, D. (1973). *Belief, Truth and Knowledge*. Cambridge: Cambridge University Press.

Arpaly, N. (2017). Epistemology and sandwiches. https://theviewfromtheowlsroost.com/2017/10/29/epistemology-and-sandwiches/

Austin, J. L. (1956). A plea for excuses. *Proceedings of the Aristotelian Society*, *57*(1), 1-30.

Babic, B. (2019). A theory of epistemic risk. *Philosophy of Science,* 86, 522-550.

Baghramian, M. and Croce, M. (2021). Experts, public policy and the question of trust. In M. Hannon and J. De Ridder (eds.), *The Routledge handbook of political epistemology*. London: Routledge.

Baghramian, M. and Panizza, S. C. (2022). Scepticism and the value of distrust. *Inquiry*. DOI: 10.1080/02691728.2022.2115325.

Barber, D. and Gordon, G. (1976). *Members of the jury*. London: Wildwood House.

Barlow, D. (2001). Anxiety and its disorders: The nature and treatment of anxiety and panic (2nd ed). New York, NY: The Guilford Press.

Bedau, M. (1991). Can biological teleology be naturalized? *The Journal of Philosophy*, 88, 647-655.

Blackburn, S. (1999). *Think: A compelling introduction to Philosophy*. Oxford: Oxford University Press.

Biddle, J. and Kukla, R. (2017). The geography of epistemic risk. In K. C. Elliot and T. Richards (eds.), *Exploring inductive risk: Case studies of values in science*. Oxford: Oxford University Press.

Boult, C. (2021). There is a distinctively epistemic kind of blame. *Philosophy and Phenomenological Research*, 103(3), 518-534.

Brandt, R. (1979). *A theory of the good and the right*. Oxford: Clarendon Press.

Bricker, A. (2018). Do judgements about risk track modal ordering? *Thought,* 7, 200-208.

Brigandt, I. and Rosario, E. 2020. Strategic conceptual engineering for epistemic and social aims. In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Brock, W. H. (1992). *The Norton history of chemistry*. New York: W. W. Norton.

Brown, J. (2008). Subject-sensitive invariantism and the knowledge norm for practical reasoning. *Noûs*, 42(2), 167-189.

Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis*, *81*, 1211-1241.

Brun, G., Doğuoğlu, U. and Kuenzle, D. (2008) Introduction. In G. Brun, U. Doğuoğlu and D. Kuenzle (eds.), *Epistemology and emotions*. Aldershot: Ashgate.

Buck, P. H. (1910). Medicine amongst the Māori's in ancient and modern times. MD Thesis, University of New Zealand. Available at: https://ourarchive.otago.ac.nz/handle/10523/10413 (accessed 24 May 2022.

Bump, P. (2016). There have been just four documented cases of voter fraud in the 2016 election. *The Washington Post*. https://www.washingtonpost.com/news/the-fix/wp/2016/12/01/0-000002-percent-of-all-the-ballots-cast-in-the-2016-election-were-fraudulent/ Accessed 27/11/2022.

Butler, G. and Mathews, A. (1987). Anticipatory anxiety and risk perception. *Cognitive Therapy and Research*, 11(5), 551-565.

Cappelen, H. (2012). *Philosophy without intuitions*. Oxford: Oxford University Press.

Cappelen, H. (2018). *Fixing language: An essay in conceptual engineering*. Oxford: Oxford University Press.

Cappelen, H. (2020.) Conceptual engineering: the master argument. In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual engineering and conceptual ethics*. Oxford: Oxford University Press.

Cappelen, H. and Plunkett, D. (2020). Introduction: a guided tour of conceptual engineering and conceptual ethics. In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual engineering and conceptual ethics*. Oxford: Oxford University Press.

Cappelen, H. and Dever, J. (2018). Puzzles of reference. Oxford: Oxford University Press.

Carnap, R. (1947). *Meaning and necessity*. Chicago, IL: University of Chicago Press.

Carnap, R. (1950). *Logical foundations of probability*. Chicago, IL: University of Chicago Press.

Carnap, R. (1963). Replies and systematic expositions. In P. A. Schilpp (ed.), *The philosophy of Rudolf Carnap*. La Salle, IL: Open Court.

Carruthers, P. (2018). Basic questions. *Mind & Language*, 33, 130-147.

Cassam, Q. (2016). Vice epistemology. *The Monist*, 99(2), 159-180.

Cassam, Q. (2019). *Conspiracy theories*. Cambridge: Polity Press.

Cassam, Q. (2021). Misunderstanding vaccine hesitancy: a case study in epistemic injustice. *Educational Philosophy and Theory*. DOI: 10.1080/00131857.2021.2006055.

Chalmers, D. (2011). Verbal disputes. *Philosophical Review*, *120*(4): 515-566.

Chalmers, D. (2020). What is conceptual engineering and what should it be? *Inquiry* DOI: 10.1080/0020174X.2020.1817141

Chisholm, R. (1977). *Theory of knowledge* (3rd edition). Englewood Cliffs, NJ: Prentice-Hall Inc.

Ciardelli, I., Groenendijk, J. and Roelofsen, F. (2019). *Inquisitive semantics*. Oxford: Oxford University Press.

Civil Partnership Act 2004. Available at: https://www.legislation.gov.uk/ukpga/2004/33/section/2/2019-12-02 (Accessed 28/06/22).

Civil Partnership, Marriages and Deaths (Registration etc) Act 2019. Available at: https://www.legislation.gov.uk/ukdsi/2019/9780111190784/contents (Accessed 28/06/22).

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19.

Clarke, M. (2010). Concepts, intuitions, and epistemic norms. *Logos & Episteme*, 1(2), 269-285.

Coady, D. (2019). Psychology and conspiracy theories. In D. Coady and J. Chase (eds.), *The Routledge handbook of applied epistemology*. London: Routledge.

Cohen, M. and Nagel, E. (1934). *An introduction to logic and scientific method*. London: Routledge & Kegan Paul Ltd.

Collins, A. (1996). Moore's paradox and epistemic risk. *The Philosophical Quarterly*, 46(184), 308-319.

Comesaña, J. (2020). A plea for falsehoods. *Philosophy and Phenomenological Research*, 100(2), 247-276.

Craig, E. (1990). Knowledge and the state of nature: An essay in conceptual synthesis. Oxford: Clarendon Press.

Craig, E. (2007). Genealogies and the state of nature. In A. Thomas (ed.), *Bernard Williams*. Cambridge: Cambridge University Press.

Creath, R. (1990). Dear Carnap, Dear Van: The Quine-Carnap correspondence and related work. Berkeley, CA: University of California Press.

Cybersecurity and Infrastructure Security Agency (2020). Joint statement from Elections Infrastructure Coordinating Council & The Election Infrastructure Coordinating Executive Committees. Retrieved from https://www.cisa.gov/news/2020/11/12/joint-statement-elections-infrastructure-government-coordinating-council-election Accessed 27/11/2022.

Davidson, D. (1990). The structure and content of truth. *The Journal of Philosophy*, 87(6), 279-328.

Davidson, D. (2005). *Truth, language and history*. Oxford: Clarendon Press.

Davies, M. and Gardner, D. (2010). *A frequency dictionary of contemporary American English*. Oxford: Routledge.

Davis, E. (2016). Typecasts, tokens, and spokespersons: a case for credibility excess as testimonial injustice. *Hypatia*, 31(3), 485-501.

Day, H. I. (1971). The measurement of specific curiosity. In H. I. Day, D. E. Berlyne, and D. E. Hunt (eds.), *Intrinsic motivation: A new direction in education*. New York: Holt, Rinehart & Winston.

Dentith, M. R. X. (2019). Conspiracy theories on the basis of evidence. *Synthese*, 196(6), 2243-2261.

Dentith, M. R. X. (2021). Debunking conspiracy theories. *Synthese*, 198(10), 9897-9911.

DeRose, K. (1991). Epistemic possibilities. *The Philosophical Review*, 100, 581-605.

DeRose, K. (2002). Assertion, knowledge, and context. *The Philosophical Review*, 111(2), 167-203.

Dever, J. (2016). What is philosophical methodology? In H. Cappelen, T. Gendler and J Hawthorne (eds.), *The Oxford handbook of philosophical methodology*. Oxford: Oxford University Press.

Dembroff, R. (2020). Beyond binary: genderqueer as critical gender kind. *Philosophers' Imprint*, *20*(9), 1-23.

Descartes, R. (2017). *Meditations on first philosophy with selections from the objections and replies*, 2nd ed., J. Cottingham (ed. and trans.). Cambridge: Cambridge University Press.

Diegan, M. (ms.). Questions should have answers. https://mikedeigan.com/pdfs/deigan-qsha-1.5.pdf

Dogramaci, S. (2012). Reverse-engineering epistemic evaluations. *Philosophy and Phenomenological Research*, 84(3), 513-530.

Dougherty T. and Rysiew, P. (2009). Fallibilism, epistemic possibility, and concessive knowledge attributions. *Philosophy and Phenomenological Research*, 78(1), 123-132.

Dretske, F. (1970). Epistemic operators. *The Journal of Philosophy*, 67(24), 1007-1023.

Ebert, P., Smith, M. and Durbach, I. (2020). Varieties of risk. *Philosophy and Phenomenological Research*, 101(2), 432-455.

Eder, A. M. A. (2021). Explicating the concept of epistemic rationality. *Synthese*, *199*, 4975-5000.

Elliot-Graves, A. (2020). What is a target system? *Biology and Philosophy*, *35*(2).

Enoch, D., Spectre, L. and Fisher, T. (2012) Statistical evidence, sensitivity and the legal value of knowledge. *Philosophy and Public Affairs,* 40(3), 197-224.

Fantl, J. and McGrath, M. (2002). Evidence, pragmatics, and justification. *Philosophical Review*, 111(1), 67-94.

Fantl, J. and McGrath, M. (2009). *Knowledge in an uncertain world*. Oxford: Oxford University Press.

Fassio, D. and McKenna, R. (2015). Revisionary epistemology. *Inquiry*, *58*(7-8), 755-779.

Feldman, R. (1981). Fallibilism and knowing that one knows. *The Philosophical Review*, 90(2), 266-282.

Feldman, R. (2000). The ethics of belief. *Philosophy and Phenomenological Research*, 60(3), 667-695.

Feldman, R. (2002). Epistemological duties. In P. Moser (ed.), *The Oxford handbook of epistemology*. Oxford: Oxford University Press.

Feltz, A. and Zarpentine, C. (2020). Do you know more when it matters less? *Philosophical Psychology*, 23(5), 683-706

Field, C. (2021). Giving up the enkratic principle. *Logos & Episteme*, 12(1), 7-28.

Fine, G. (2014). The possibility of inquiry: Meno's Paradox from Socrates to Sextus. Oxford: Oxford University Press.

Flaubert, G. (2001). *Memoirs of a madman* (trans. and ed. T. Unwin). Liverpool: Liverpool Online Series Critical Editions of French Texts.

Flores, C. and Woodard, E. (forthcoming). Epistemic norms on evidence-gathering. *Philosophical Studies*.

Foley, R. (1987). *The theory of epistemic rationality*. Cambridge, MA: Harvard University Press.

van Fraassen, B. (1977). The pragmatics of explanation. *American Philosophical Quarterly*, 14(2), 143-150.

Fricker, M. (1998). Rational authority and social power: towards a truly social epistemology. *Proceedings of the Aristotelian Society*, 98(2), 159-177.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.

Fricker, M. (2016). What's the point of blame? A paradigm-based explanation. *Noûs*, 50(1), 165-183.

Fricker, M. (2019). Forgiveness: an ordered pluralism. *Australasian Philosophical Review*, 3(3), 241-260.

Friedman, J. (2013). Question-directed attitudes. *Philosophical Perspectives*, 27(1), 145-174).

Friedman, J. (2017). Why suspend judging? *Noûs*, 51(2), 302-326.

Friedman, J. (2019a). Checking again. *Philosophical Issues*, 29(1), 84-96.

Friedman, J. (2019b). Inquiry and belief. *Noûs*, 53(2), 296-315.

Friedman, J. (2020). The epistemic and the zetetic. *Philosophical Review*, 129(4), 501-536.

Gardiner, G. (2015). Teleologies and the methodology of epistemology. In J. Greco and D. Henderson (eds.), *Epistemic evaluation: Purposeful epistemology*. Oxford: Oxford University Press.

Gardiner, G. (2019). The reasonable and the relevant. *Philosophy and Public Affairs*, 47(3), 288-318.

Geach, P. (1956). Good and evil. *Analysis*, 17(2), 33-42.

Gerken, M. (2017). On folk epistemology: How we think and talk about knowledge. Oxford: Oxford University Press.

Ginzburg, J. (1995). Resolving questions, I. *Linguistics and Philosophy*, 18(5), 459-527.

Goddard, C. (2010). Universals and variation in the lexicon of mental state concepts. In B. Malt and P. Wolff (eds.), *Words and the mind*. Oxford: Oxford University Press.

Godfrey-Smith, P. (2009). Abstractions, idealizations, and evolutionary biology. In A. Barberousse, M. Morange and T. Pradeu (eds.), *Mapping the future of biology: Evolving concepts and theories*. Dordrecht: Springer.

Goldie, P. (2002). *The emotions: A philosophical exploration*. Oxford: Oxford University Press.

Goldie, P. (2004). Emotion, reason, and virtue. In D. Evans and P. Cruse (eds.), *Emotion, evolution, and rationality*. Oxford: Oxford University Press.

Goldman, A. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy*, 73(20), 771-791.

Graham, P. (2014). Warrants, functions, history. In A. Fairweather and O. Flanagan (eds.), *Naturalizing epistemic virtue*. Cambridge: Cambridge University Press.

Greco, J. (2008). What's wrong with contextualism? *The Philosophical Quarterly*, 58(232), 416-436.

Greco, J. and Henderson, D. (2015). The point and purpose of epistemic evaluation. In J. Greco and D. Henderson (eds.), *Epistemic evaluation: Purposeful epistemology*. Oxford: Oxford University Press.

Grimm, S. (2015). Knowledge, practical interests, and rising tides. In J. Greco and D. K. Henderson (Eds.), *Epistemic evaluation: Purposeful epistemology*. Oxford: Oxford University Press.

Grimshaw, J. (1979). Complement section and the lexicon. *Linguistic Inquiry*, 10(2), 279-326.

Grimshaw, J. (1990). *Argument structure*. Cambridge, MA: MIT Press.

Groenendijk, J. and Stokhof, M. (1984). *Studies on the semantics of questions and the pragmatics of answers*, Dissertation, University of Amsterdam. Amsterdam.

Hacking, I. (1967). Possibility. *The Philosophical Review*, 76(2), 143-168.

Hamblin, C. L. (1958). Questions. *Australasian Journal of Philosophy*, 36(3), 159-168.

Hannon, M. (2015). Stabilizing knowledge. *Pacific Philosophical Quarterly*, 96, 116-139.

Hannon, M. (2019). What's the point of knowledge? A function-first epistemology. Oxford: Oxford University Press.

Hansson, S. O. (2018). Risk. *Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/risk/ Retrieved August 29, 2021.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press.

Harris, K. (2018). What's epistemically wrong with conspiracy theorising? *Royal Institute of Philosophy Supplement*, 84, 235-257.

Haslanger, S. (1999). What knowledge is and what it ought to be: feminist values and normative epistemology. *Philosophical Perspectives*, 13, 459-480.

Haslanger, S. (2000). Gender and race: (what) are they? (What) Do we want them to be? *Noûs* 34(1), 31-55.

Haslanger, S. (2006). What good are our intuitions? *Proceedings of the Aristotelian Society*, *Supplementary Volumes*, *80*, 89-143.

Haslanger, S. (2012a). Resisting reality: Social construction and social critique. Oxford: Oxford University Press.

Haslanger, S. (2012b). Social construction: the 'debunking' project. In *Resisting reality: Social construction and social critique*. Oxford: Oxford University Press.

Hawthorne, J. (2004). *Knowledge and lotteries*. Oxford: Oxford University Press.

Hawthorne, J. and Stanley, J. (2008). Knowledge and action. *The Journal of Philosophy*, 105(10), 571-590.

Hedden, B. (2015). Reasons without persons: rationality, identity, and time. Oxford: Oxford University Press.

Helm, B. (2001). Emotional reason: Deliberation, motivation, and the nature of value. Cambridge: Cambridge University Press.

Higginbotham, J. (1992). Truth and understanding. *Philosophical Studies*, *65*, 1-18.

Ho, H. L. (2008). *A philosophy of evidence law*. Oxford: Oxford University Press.

Hobbes, T. (2008). *Leviathan*, J. C. A. Gaskin (ed.). Oxford: Oxford University Press.

Hookway, C. (1998). Doubt: affective states and the regulation of inquiry. *Canadian Journal of Philosophy*, 24, 203-225.

Hookway, C. (2008). Epistemic immediacy, doubt and anxiety: on a role for affective states in epistemic evaluation. In G. Brun, U. Doğuoğlu and D. Kuenzle (eds.), *Epistemology and emotions*,. Aldershot: Ashgate.

Ichikawa, J. J. (2017) *Contextualising knowledge: Epistemology and semantics*. Oxford: Oxford University Press.

Jackson, S. (1991). The lottery. In *The Lottery and Other Stories*. New York, NY: Farrar Press.

James, W. (1884). What is an emotion? *Mind*, 9(34), 188-205.

Jenkins, K. (2016). Amelioration and inclusion: gender identity and the concept of woman. *Ethics*, *126*(2), 394-421.

Jenkins, K. (2021). Rape myths: what are they and what can we do about them? *Royal Institute of Philosophy Supplement*, 89, 37-49.

Jensen, K. K. (2012). A philosophical assessment of decision theory. In S. Roeser, R. Hillerbrand, P. Sandin and M. Peterson (Eds.), *Handbook of risk theory: Epistemology, decision theory, ethics, and social implications*. Dordrecht: Springer.

Joyce, R. (2001). *The myth of morality*. Cambridge: Cambridge University Press.

Kang, M. J., Hsu, M. Krajbich, I. M., Loewenstein, G., McClure, S.M., Wang, J. T. Y., Camerer, C. F. (2009). The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, 20(8), 963–973.

Kant, I. (2012). *Groundwork of the Metaphysics of Morals*, 2nd ed., C. Korsgaard (ed.), M. Gregor and J. Timmermann (trans.). Cambridge: Cambridge University Press.

Kappel, K. (2010). On saying that someone knows: themes from Craig. In A. Haddock, A. Millar, and D. Pritchard (Eds.), *Social epistemology*. Oxford: Oxford University Press.

Kashdan, T. B., Rose, P. and Finchham, F. D. (2004). Curiosity and exploration: facilitating positive subjective experiences and personal growth opportunities. *Journal of Personality Assessment*, 82, 291-305.

Kauppinen, A. (2018). Epistemic norms and epistemic accountability. *Philosophers' Imprint*, 18(8), 1-16.

Keeley, B. L. (1999). Of conspiracy theories. *The Journal of Philosophy*, 96(3), 109-126.

Kelly, K., Lovett, J. and Regan, L. (2005). A gap or a chasm? Attrition in reported rape cases. London: Home Office Research, Development and Statistics Directorate.

Kelp, C. (2011). What's the point of "knowledge" anyway? *Episteme* 8(1): 53-66.

Kelp, C. (2014). Two for the knowledge goal of inquiry. *American Philosophical Quarterly*, 51, 227-232.

Kelp, C. (2021a). *Inquiry, knowledge, and understanding*. Oxford: Oxford University Press.

Kelp, C. (2021b). Theory of inquiry. *Philosophy and Phenomenological Research*, 103(2), 359-384.

Kemeny, M. E. and Shestyuk, A. (2008). Emotions, the neuroendocrine and immune systems, and health. In M. Lewis, J. M. Haviland-Jones, L. Feldman Barrett (Eds.), *Handbook of emotions* (3rd ed.). New York, NY: The Guilford Press.

Kiefer, F. (1988). On the pragmatics of answers. In M. Meyer (ed.), *Questions and questioning*. Berlin: De Gruyter.

Koch, S. (2021). Why conceptual engineers should not worry about topics. *Erkenntnis*. DOI: 10.1007/s10670-021-00446-1.

Kornblith, H. (2002). *Knowledge and its place in nature*. Oxford: Oxford University Press.

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

Kurth, C. (2015). Moral anxiety and moral agency. In M. Timmons (ed.), *Oxford studies in normative ethics* (Vol. 5). Oxford: Oxford University Press.

Kurth, C. (2018). The anxious mind: An investigation into the varieties and virtues of anxiety. Cambridge, MA: MIT Press.

Kusch, M. (2009). Testimony and the value of knowledge. In A. Haddock, A. Millar and D. Pritchard (eds.), *Epistemic value*. Oxford: Oxford University Press.

Kusch, M. and McKenna, R. (2020). The genealogical method in epistemology. *Synthese*, 197, 1057-1076.

Kvanvig, J. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge: Cambridge University Press.

Kvanvig, J. (2009). Assertion, knowledge, and lotteries. In P. Greenough and D. Pritchard (eds.), *Williamson on knowledge*. Oxford: Oxford University Press.

Lader, M. and Marks, E. (1973). *Clinical Anxiety*. London: Heinemann.

The Lancet (2019). Vaccine hesitancy: a generation at risk. Editorial. *The Lancet Child & Adolescent Health*, 3(5), 281.

Lasonen-Aarnio, M. (2008). Single-premise deduction and risk. *Philosophical Studies*, 141, 157-173.

Lasonen-Aarnio, M. (2010). Unreasonable knowledge. *Philosophical Perspectives*, 24(1), 1-21.

Lasonen-Aarnio, M. (2014). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, 88(2), 314-345.

Lazerowitz, M. (1970). A note on 'metaphilosophy'. *Metaphilosophy*, 1(1), 91.

Leslie, S. (2017). The original sin of cognition: fear, prejudice, and generalization. *Journal of Philosophy*, *114*(8), 393-421.

Levitt, J. (2014). A comprehensive investigation of voter impersonation finds 31 credible incidents out of one billion ballots cast. *The Washington Post*. https://www.washingtonpost.com/news/wonk/wp/2014/08/06/a-comprehensive-investigation-of-voter-impersonation-finds-31-credible-incidents-out-of-one-billion-ballots-cast/ Accessed 27/11/2022.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, 70(17), 556-567.

Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74(4), 549-567.

Loewenstein, G., Weber, E. U., Hsee, C. K., and Welch, N. (2001). Risk as feeling. *Psychological Bulletin*, 127(2), 267-286.

Luzzi, F. (2016). Testimonial injustice without credibility deficit (or excess). *Thought*, 5(3), 203-211.

Lynch, M. (2005). *True to life: why truth matters*. Cambridge, MA: MIT Press.

Marriage (Same Sex Couples) Act 2013. Available at: https://www.legislation.gov.uk/ukpga/2013/30/contents/enacted/data.htm (Accessed 28/06/22).

Marsh, H. (2014). *Do no harm*. New York, NY: St. Martin's Press.

McGrath, M. (2021). Being neutral: Agnosticism, inquiry and the suspension of judgment. *Noûs*, 55(2), 463-484.

McKenna, R. (2017). Pluralism about knowledge. In A. Coliva and N. Pedersen (eds.), *Epistemic pluralism*. Cham: Palgrave Macmillan.

McLaughlin, P. (2001). *What functions explain*. Cambridge: Cambridge University Press.

Medina, J. (2012). The epistemology of resistance: gender and racial oppression, epistemic injustice, and the social imagination. Oxford: Oxford University Press.

Mellor, D. H. (2005). *Probability: A philosophical introduction*. London: Routledge.

Metzger, C. (2021). How to do the curly girl method for beginners. *Cosmopolitan* https://www.cosmopolitan.com/style-beauty/beauty/a34292024/curly-girl-method-how-to/ Accessed 29 December 2021.

Millar, A. (2011). Why knowledge matters. Proceedings of the Arisotelian Society, Supplementary Volume, 85, 63-81.

Millikan, R. G. (1984). Language, thought, and other biological categories: New foundations for realism. Cambridge, MA: MIT Press.

Mills, C. (2007). White ignorance. In S. Sullivan and N. Tuana (eds.), *Race and epistemologies of ignorance*. Albany, NY: SUNY Press.

Millson, J. (2021). Seeking confirmation: a puzzle for norms of inquiry. *Analysis*, 80(4), 683-693.

Minnite, L. C. (2010). *The myth of voter fraud*. Ithaca: Cornell University Press.

Möller, N. (2012). The concepts of risk and safety. In S. Roeser, R. Hillerbrand, P. Sandin and M. Peterson (eds.), *Handbook of risk theory: Epistemology, decision theory, ethics, and social implications*. Dordrecht: Springer.

Morton, A. (2010). Epistemic emotions. In P. Goldie (ed.), *The Oxford handbook of philosophy of emotion*. Oxford: Oxford University Press.

Moser, P. L. (2015). Metaphilosophy. In R. Audi (ed.), *The Cambridge Dictionary of Philosophy* (3rd edition). Cambridge: Cambridge University Press.

National Health Service (2019). Women's health. https://www.nhs.uk/common-health-questions/womens-health/ Accessed 18 February 2022.

Nagel, J. (2010a). Epistemic anxiety and adaptive invariantism. *Philosophical Perspectives*, 24, 407-435.

Nagel, J. (2010b). Knowledge ascriptions and the psychological consequences of thinking about error. *The Philosophical Quarterly*, 60(239), 286-306.

Navarro, J. (2019). Luck and risk: how to tell them apart. *Metaphilosophy*, 50(1-2), 63-75.

Navarro, J. (2021). Epistemic luck *and* epistemic risk. *Erkenntnis*, 88, 929-950.

Navarro, J. (Forthcoming). Hacia una epistemología del secreto. In M. Moscoso Pérez (ed.), *El secreto: restricción y circulación de la información en la sociedad del conocimiento.* Madrid: Plaza y Valdés/CSIC.

Nelson, J. S. (2006). *Fishes of the world* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Newton, L. (2022). Epistemic anxiety and epistemic risk. *Synthese*, 200(324). DOI: 10.1007/s11229-022-03788-7.

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, 17(2), 141-161.

Nietzsche, F. (2006). *On the genealogy of morality*, K. Ansell-Pearson (ed.) and C. Diethe (trans). Cambridge: Cambridge University Press.

Nozick, R. (1981). *Philosophical explanations*. Cambridge, MA: Harvard University Press.

Öhman, A. (1993). Fear and anxiety as emotional phenomena: clinical phenomenology, evolutionary perspectives, and information processing mechanisms. In M. Lewis & J. M. Haviland (eds.) *Handbook of emotions.* New York, NY: Guilford Press.

Oreskes, N. and Conway, E. M. (2010). Merchants of doubt: how a handful of scientists obscured the truth on issues from tobacco smoke to global warming. New York: Bloomsbury Press.

O'Sullivan, A. and Newton, L. (ms.). A function-first approach to *risk*.

Overgaard, S., Gilbert, P. and Burwood, S. (2013). *An introduction to metaphilosophy*. Cambridge: Cambridge University Press.

Parfit, D. (1984). *Reasons and persons*. Oxford: Clarendon Press.

Peirce, C. S. (1958). *The collected papers of Charles Sanders Peirce*. Vols. I-VI C. Hartshorne and P. Weiss (eds.), vols. VII-VIII A. W. Burks (ed.). Cambridge, MA: Harvard University Press.

Peirce, C. S. (1877). The fixation of belief. *Popular Science Monthly*, 12, 1-15.

Peschard, I. (2010). Target systems, phenomena and the problem of relevance. *The Modern Schoolman*, *87*(3-4), 267-284.

Pesetsky, D. (1990). *Zero syntax: experiencers and cascades*. Cambridge, MA: MIT Press.

Pettigrew, R. (2019). *Epistemic risk and the demands of epistemic rationality*. Oxford: Oxford University Press.

Plantinga, A. (1993). *Warrant and proper function*. Oxford: Oxford University Press.

Pritchard, D. (2005). *Epistemic luck*. Oxford: Oxford University Press.

Pritchard, D. (2015). Risk. *Metaphilosophy*, 46(3), 436-461.

Pritchard, D. (2016). Epistemic risk. *The Journal of Philosophy*, 113(11), 550-571.

Pritchard, D. (2018). Legal Risk, Legal Evidence and the Arithmetic of Criminal Justice. *Jurisprudence*, 9(1), 108-119.

Pritchard, D. (2022). Varieties of epistemic risk. *Acta Analytica*, 37, 9-23.

Prinz, J. (2004). *Gut reactions: A perceptual theory of emotions*. Oxford: Oxford University Press.

Prinzing, M. (2018). The revisionist's rubric: conceptual engineering and the discontinuity objection. *Inquiry*, 61(8), 854-880.

Putterman, S. (2020). No, the new coronavirus vaccines are not more dangerous than COVID-19. *Politifact*. https://www.politifact.com/factchecks/2020/dec/18/facebook-posts/no-new-coronavirus-vaccines-are-not-more-dangerous/ Accessed 27/11/2022.

Queloz, M. (2020). From paradigm-based explanation to pragmatic genealogy. *Mind*, *129*(515), 683-714.

Queloz, M. (2021). *The practical origin of ideas*. Oxford: Oxford University Press.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Railton, P. (1986). Moral realism. *The Philosophical Review*, 95(2), 163-207.

Railton, P. (1989). Naturalism and prescriptivity. *Social Philosophy and Policy*, 7(1), 151-174.

Ralph, P. and Coop, G. (2013). The geography of recent genetic ancestry across Europe. *PLoS Biology*, *11*(5). DOI: 10.1371/journal.pbio.1001555.

Richard, M. (2019). *Meaning as species*. Oxford: Oxford University Press.

Rorty, R. (1995). Is truth a goal of enquiry? Davidson vs. Wright. *The Philosophical Quarterly*, 45, 281-300.

Rorty, R. (2000). Universality and truth. In R. Brandom (ed.), *Rorty and his critics*. Oxford: Blackwell.

Russell, B. (1948), H*uman knowledge: Its scope and limits*. London: Allen and Unwin.

Rysiew, P. (2012). Epistemic scorekeeping. In J. Brown and M. Gerken (Eds.), *Knowledge ascriptions.* Oxford: Oxford University Press.

Sainsbury, R. M. (1997). Easy possibilities. *Philosophy and Phenomenological Research*, 57(4), 907-919.

Saul, J. (2012). Politically significant terms and philosophy of language: methodological issues. In S. Crasnow and A. Superson (eds.), *Out from the shadows: Analytical feminist contributions to traditional philosophy*. Oxford: Oxford University Press.

Sawyer, S. (2020). Talk and thought. In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual engineering and conceptual ethics*. Oxford: Oxford University Press.

Scerri, E. R. (2001). Prediction and the Periodic Table. *Studies in History and Philosophy of Science, 32*(3), 407-452.

Scerri, E. R. (2006). *The Periodic Table: its story and significance*. Oxford: Oxford University Press.

Schaffer, J. (2008). Knowledge in the image of assertion. *Philosophical Issues*, 18, 1-19.

Scharp, K. (2013). *Replacing truth*. Oxford: Oxford University Press.

Scharp, K. and Shapiro, S. (2017). Revising inconsistent concepts. In B. Armour-Garb (ed.), *Reflections on the liar*. Oxford: Oxford University Press.

Schopenhauer, A. (2010). *The world as will and representation*, vol. 1, J. Norman and A. Welchman (ed. and trans.), C. Janaway (ed.). Cambridge: Cambridge University Press.

Searle, J. (1995). *The construction of social reality*. New York: Simon & Schuster.

Sennet, A. (2016). Polysemy. In E. N. Zalta, *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/spr2016/entries/ambiguity/ Accessed 24/10/2022.

Shakespeare, W. (1991). *Measure for measure. The Oxford Shakespeare*, N. W. Bawcutt (ed.). Oxford: Oxford University Press.

Shatz, M., Wellman, H. M., and Silber, S. (1983). The acquisition of mental verbs: a systematic investigation of the first reference to mental state. *Cognition*, *14*, 301-314.

Shulman, E. P. and Cauffman, E. (2014). Deciding in the dark: Age differences in intuitive risk judgment. *Developmental Psychology*, 50(1), 167-177.

Simion, M. (2023). Resistance to evidence and the duty to believe. *Philosophy and Phenomenological Research*. DOI: 10.1111/phpr.12964.

Simion, M. and Kelp, C. (2020). Conceptual innovation, function first. *Noûs*, *54*(4), 985-1002.

Slaby, J. and Wüschner, P. (2014). Emotions and agency. In S. Roeser and C. Todd (eds.), *Emotion and value*. Oxford: Oxford University Press.

Smith, M. (2012). Two notions of epistemic risk. *Erkenntnis*, 78, 1069-1079.

Smith, M. (2016) Between probability and certainty: What justifies belief. Oxford: Oxford University Press.

Smith, M. (2018). When does evidence suffice for conviction? *Mind*, 127(508), 1193-1218.

Smith, M. (2022). Two accounts of assertion. *Synthese*, 200(3), DOI: 10.1007/s11229-022-03745-4.

Solomon, R. (1993). *The passions: Emotions and the meaning of life* (2nd ed.). Indianapolis: Hackett.

Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives*, 13, 141-153.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25(5/6), 701-721.

Stalnaker, R. (2014). *Context*. Oxford: Oxford University Press.

Stanley, J. (2005a). *Knowledge and practical interests*. Oxford: Oxford University Press.

Stanley, J. (2005b). On the linguistic basis for contextualism. *Philosophical Studies*, 199(1-2), 119-146.

Sterken, R. (2020). Linguistic intervention and transformative communicative disruptions. In H. Cappelen, D. Plunkett, & A. Burgess (eds.), *Conceptual engineering and conceptual ethics*. Oxford: Oxford University Press.

Stich, S. (1979). Do animals have beliefs? *Australasian Journal of Philosophy*, 57(1), 15-28.

Stine, G. (1976). Skepticism, relevant alternatives, and deductive closure. *Philosophical Studies*, 29(4), 249-261.

Strawson, P. (1963). Carnap's views on conceptual systems versus natural languages in analytic philosophy. In P. A. Schilpp (ed.), *The philosophy of Rudolf Carnap*. La Salle, IL: Open Court.

Strawson, P. (1992). *Analysis and metaphysics: an introduction to philosophy*. Oxford: Oxford University Press.

Tanesini, A. (2008). Virtues, emotions and understanding. In G. Brun, U. Doğuoğlu and D. Kuenzle (eds.), *Epistemology and emotions*,. Aldershot: Ashgate.

Tappolet, C. (2016). *Emotions, value and agency*. Oxford: Oxford University Press.

Thagard, P. (2008). How cognition meets emotion: beliefs, desires and feelings as neural activity. In G. Brun, U. Doğuoğlu and D. Kuenzle (eds.), *Epistemology and emotions*,. Aldershot: Ashgate.

Thomasson, A. (2020). A pragmatic method for normative conceptual work. In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual engineering and conceptual ethics*. Oxford: Oxford University Press.

Thorstad, D. (2022). There are no epistemic norms of inquiry. *Synthese*, 200(410). DOI: 10.1007/s11229-022-03896-4.

Tilton, E. C. R. (2022). Rape myths, catastrophe, and credibility. *Episteme*. DOI: 10.1017/epi.2022.5.

Turri, J. (2010). Epistemic invariantism and speech act contextualism. *The Philosophical Review*, 119(1), 77-95.

Vargas, M. (2013). Building better things: A theory of moral responsibility. Oxford: Oxford University Press.

Vazard, J. (2018). Epistemic anxiety, adaptive cognition, and obsessive-compulsive disorder. *Discipline Filosofiche*, 28(2), 137-158.

Vazard, J. (2021). (Un)reasonable doubt as affective experience: obsessive–compulsive disorder, epistemic anxiety and the feeling of uncertainty. *Synthese*, 198, 6917-6934.

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

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature*, *118*, 558-560.

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., Berelowitz, M., Dhillon, A. P., Thomson, M.
A., Harvey, P., Valentine, A., Davies, S. E., & Walker-Smith, J. A. (1998). Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, 351(9103), 637-641.

Weinberg, J. M. (2007), How to challenge intuitions empirically without risking skepticism. *Midwest Studies in Philosophy*, 31(1), 318-343.

Weisberg, M. (2007). Who is a modeler? *British Journal of the Philosophy of Science*, *58*, 207-233.

Weisberg, M. (2013). *Simulation and similarity*. Oxford: Oxford University Press.

Willard-Kyle, C. (forthcoming). The knowledge norm for inquiry. *The Journal of Philosophy*.

Williams B. (1982). Practical necessity. In D. M. Mackinnon, B. Hebblethwaite and S. R. Sutherland (eds.), *The philosophical frontiers of Christian theology: essays presented to D. M. Mackinnon*. Cambridge: Cambridge University Press.

Williams, B. (2002). *Truth and truthfulness: An essay in genealogy*. Princeton, NJ: Princeton University Press.

Williams, B. (2014). Why philosophy needs history. In his *Essays and reviews 1959-2002*, M. Woods (ed.). Princeton, NJ: Princeton University Press.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Williamson, T. (2005). Knowledge and scepticism. In F. Jackson and M. Smith (eds.), *The Oxford handbook of contemporary philosophy*. Oxford: Oxford University Press.

Williamson, T. (2007). *The philosophy of philosophy*. Oxford: Wiley-Blackwell.

Williamson, T. (2011). Very improbable knowing. *Erkenntnis*, 79 5), 971-999.

World Health Organization (2003). MMR and autism: extract from report of GACVS meeting of 16-17 December 2002. https://www.who.int/groups/global-advisory-committee-on-vaccine-safety/topics/mmr-vaccines-and-autism#cms Accessed 27/11/2022.

World Health Organization (2018). ICD-11: International classification of diseases for mortality and morbidity statistics (11th ed). Retrieved from https://icd.who.int/

Wright, C. (1992). *Truth and objectivity*. Cambridge, MA: Harvard University Press.

Wright, C. (2004). Warrant for nothing (and foundations for free?). *Aristotelian Society Supplementary Volume*, 78(1), 167-212.

Wright, L. (1973). Functions. *Philosophical Review*, *82*(2), 139-168.

Zeidner, M. and Matthews, G. (2005). Evaluation anxiety: current theory and research. In A. Elliot and C. Zweck (eds.), *Handbook of competence and motivation*. New York: Guilford Publications.