ON WHAT MATTERS IN ASSESSING CONCEPTS

By

Denis Kazankov

Submitted to

Central European University

Department of Philosophy

In partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Supervisor: Asya Passinsky

Vienna, Austria

2025

Author's Declaration

I hereby declare that the dissertation contains no material accepted for any other degree in any other institution. Nor does it contain any material previously written and/or published by another person, except where appropriate acknowledgement is made in the form of bibliographic reference.

Denis Kazankov

Vienna, August 1, 2025

Abstract

This dissertation investigates a range of underexplored questions in conceptual ethics — the subfield of conceptual engineering concerned with assessing representational devices in language and thought. My focus is on linguistic conceptual ethics, whose target domain is linguistic expressions. The dissertation consists of seven chapters, divided into three parts.

Part I (Chapters 1-4) focuses on the idea that conceptual assessment should be sensitive to the functions of representational devices. I first argue that a satisfactory interpretation of conceptual function must align with how we generally understand tool functions. To this end, I propose interpreting conceptual function in terms of the actual, potential, or perceived effects that motivate a user group — often implicitly — to retain an expression in their conceptual repertoire. I then defend the theoretical viability of this interpretation and demonstrate its usefulness through two case studies: one on cross-linguistic conceptual ethics and another on thick ethical concepts.

Part II (Chapters 5-6) focuses on the ethics of conceptual abandonment, the subfield of conceptual ethics that investigates when expressions become so defective that it's preferable to stop using them. I examine one argumentative strategy for abandonment, which appeals to a mismatch between what an expression refers to and what its users, on reflection, want it to refer to. I argue that abandoning an expression is a proper response to such mismatches only under relatively demanding conditions. I then critically engage with Cappelen's (2023) recent argument for abandoning the expressions 'democracy' and 'democratic', and propose an alternative diagnosis that supports their retention.

Part III (Chapter 7) tackles the problem of how to non-arbitrarily address epistemic-moral conflicts in conceptual assessment. I argue that while the prospects for addressing them through the method of comparative reasoning are grim, there is underexplored potential in what I call the 'integrative approach', which aims to restructure our patterns of linguistic engagement so as to promote greater convergence between epistemic and moral standards in conceptual ethics.

For my mother, Dagmar.

Acknowledgements

I would first like to thank Prof. Asya Passinsky, who was the best supervisor I could have imagined for this project. Asya is a remarkably quick-witted and incisive philosopher, always able to identify precisely how my philosophical writing could be improved. Over the four years I worked with her, she consistently did so with warmth, generosity, patience, and humanity. It's no exaggeration to say that working with her was the best part of my PhD.

I also want to thank my mother, Dagmar, for her unwavering love and support throughout the writing of this dissertation. I dedicate it to her.

I'm grateful to Prof. Herman Cappelen for hosting me as a visiting researcher in the Department of Philosophy at the University of Hong Kong during the Fall Semester of 2024. It was a once-in-a-lifetime opportunity to be at the global centre of conceptual engineering and to engage in many inspiring exchanges with faculty and students. I will always associate this dissertation with long evenings spent philosophising and drinking tea by Victoria Harbour, taking in the sea panorama of Kowloon.

Lastly, I would like to thank all those who left their mark on this dissertation — whether by supporting me as friends or by inspiring me as colleagues through our intellectual exchanges. The following is an alphabetical list of some of them: Ivan Baltay, Bálint Békefi, Michal Čarný, Tim Crane, Max Deutsch, Anton Dolmatov, Yaren Duvarci, Matti Eklund, Shimpei Endo, Katalin Farkas, Maria Fedorova, Yufeng Fei, Hélène Frésard, Martin Garaj, Anca Gheaus, Henrique Gonçalves, Michael Griffin, Ying Heng, Ondrej Hinček, Daniel Holko, Frank Hong, Klára Hulíková, Taavet Kalda, Rael Kalda, Pelin Kasar, Daniel Kazankov, Matej Kerekréty, Hana Kerekrétyová, Rebeka Kerekréty Nagyová, Aleksandra Knežević, Miloš Kosterec, Pavlo Kostin, Rafal Klosek, Max Kölbel, Filip Kollárik, Pavel Krejčí, Maria Kronfeldner, Adam Líška, Yiping Lin, James Luong, Nikhil Mahant, Marta Matkovic, Takaaki Matsui, Matthew McKeever, Michael Memari, Ali Mousareza, Jennifer Nado, Matteo Pascusci, Stephania Pimentel, David Plunkett, Darren Rondganger, Liam Ryan, Jonathan Schaffer, Krzysztof Sękowski, David Simon, Matthew Simpson, Jason Stanley, Jessica Still, Rachel Sterken, Amie Thomasson, Justin Tiwald, Mano Toth, Emanuele Tullio, Sami Ulqani, Lorenzo Valenti, Daniela Vacek, Martin Vacek, Alex Vacar, Matúš

Vaňátka, Athina Vrosgou, Angie Wangnan, David Weberman, Michaela Wenzlová, Emily Williamson, Edison Yi, Xindi Ye, Kevin Zhang, Bojin Zhu, and others unnamed.

Table of Contents

Introduction

Conceptual engineering is the project of assessing and improving representational devices in our language and thought.[1] My dissertation focuses on the first component of this process, conceptual assessment, which, following Burgess and Plunkett (2013a, 2013b, 2020), is commonly referred to as 'conceptual ethics'. When pursuing conceptual ethics, we adopt a critical evaluative stance towards our representational devices: rather than simply presupposing that our existing representational devices are already in good enough shape, we ask how they *should* be designed to be valuable tools for their users. To illustrate the range of dimensions along which representational devices can be assessed, here are some questions that a conceptual ethicist might be interested in investigating: How should a representational device be used? What should be its semantic content? What function should it fulfil for its users? What features should it have to best fulfil its function? What bodies of information, valences, or emotions should users associate with it? What cognitive or practical effects should its use have on users and on the world at large?

Although conceptual engineering has been the subject of lively debate among philosophers in recent years, conceptual ethics hasn't been in the spotlight.[2] Instead, more attention has been devoted to questions such as what methodological significance conceptual engineering has in philosophical inquiry (e.g., Cappelen, 2018, chs.3-4; Chalmers, 2020; Deutsch, 2020; Scharp, 2020; Thomasson, 2021; Sękowski & Landes, 2024; Koch & Ohlhorst, 2024; Eklund, 2024; Isaac, 2025), what its proper target domain is (e.g., Cappelen, 2018, pp.137-162; Pinder, 2020; Nado, 2021b; Isaac, 2021; Nefdt, 2024), and whether — and how — it's practically implementable (e.g., Cappelen, 2018, pp.72-79; Jorem, 2021; Nimtz, 2024a, 2024b; Koch, 2021b; Matsui, 2024). The objective of this dissertation is to help fill this gap by investigating a range of underexplored questions I consider crucial for a serious engagement with conceptual ethics, as they guide us in how to appropriately assess representational devices. Hence, the title *On What Matters in Assessing Concepts*.

[1] See Cappelen (2018), Cappelen and Plunkett (2020), Chalmers (2020), Belleri (2021b), Eklund (2021), Isaac et al. (2022), and Koch et al. (2023) for an overview of the project.

[2] Notable exceptions include Burgess and Plunkett (2013a, 2013b), Simion (2018a), Podosky (2018, 2022), Isaac (2024), McPherson and Plunkett (2021), Queloz (2022, 2024a, 2024b, 2025), Köhler and Veluwenkamp (2024, 2025), Richardson (2024), Crisp (2025), and Smyth (n.d.).

In this dissertation, I use the term 'concept' as a generic label for representational devices, which may be either linguistic entities, such as lexical items, or mental entities, such as building blocks of thought or structured bodies of information about some category that explain our cognitive processes related to it.[3] My primary focus, however, is on *linguistic conceptual ethics*, which centres on the assessment of lexical items, particularly referential expressions. This linguistic focus is motivated by the fact that it aligns well with the central themes of the dissertation. Conceptual ethics treats representational devices as human tools with collective instrumental value for their users. Referential expressions fit especially well within this instrumentalist framework, as they are clearly tools human speakers collectively design by employing them in discourse to fulfil various (extra-)representational functions in communication and thought. Furthermore, I address several questions that are best tractable from a linguistic standpoint, such as how to conduct conceptual assessments across different languages or when a piece of terminology is so defective that we should stop using it.[4] That said, I believe my discussion is also relevant to theorists who view conceptual engineering as targeting mental entities (e.g., Isaac, 2020, 2021; Nefdt, 2024; Machery, 2021), since I argue that even the mental aspects of an expression — such as its associated bodies of information, its role in our thinking, and its surrounding conceptual networks — affect its value for users.

The dissertation consists of seven chapters and is divided into three parts. Part I, comprising the first four chapters, addresses what is perhaps the most widely discussed idea in the conceptual ethics literature: conceptual engineering, including conceptual assessment, should be sensitive to the functions representational devices serve for their users. Chapter 1 explores how best to interpret the notion of conceptual function. I argue that, since referential expressions are a subset of human tools, an adequate interpretation of their function should align with general observations about how we expect the functions of human tools to behave. I propose four desiderata for such an

[3] The building-block view of concepts is popular among philosophers of mind (e.g., Rey, 1983; Fodor, 1998; Laurence & Margolis, 1999), as well as some theorists in conceptual engineering and ethics (e.g., Burgess & Plunkett, 2013a, p.1095; Simion & Kelp, 2020, p.986; Koch, 2021a). In contrast, the view of concepts as structured bodies of information (e.g., exemplars, prototypes, and theories) is more common in psychological theories (see Machery, 2009; Löhr, 2020; and Isaac, 2020 for discussion).

[4] Additionally, linguistic expressions seem more natural primary targets for conceptual engineering than concepts understood as abstract objects akin to Fregean senses (Peacocke, 1992; Zalta, 2001). Since abstract concepts are immutable, they can only be engineered indirectly by changing which concept serves as an expression's semantic content. However, limiting assessment to semantic content is too narrow since, as noted above, semantic content is only one of many aspects for which representational devices can be assessed in conceptual ethics.

interpretation and argue that two candidate interpretations of conceptual function, which follow dominant philosophical approaches to function, fall short on some of them. As an alternative, I develop a novel motivation-based interpretation, which interprets conceptual function in terms of the effects that centrally motivate a user group, often implicitly, to retain an expression in their repertoire.

In Chapter 2, I lend further plausibility to the motivation-based interpretation of conceptual function by arguing that it can withstand Cappelen's scepticism about the possibility of specifying which of the plethora of effects associated with an expression qualify as its function in a theoretically fruitful yet not overly controversial way. I argue that there is a schema that meets this challenge — one based on the idea that a user group's central motivation for using an expression derives from the perceived significance of the category they expect it to represent, along with its extra-representational effects.

In Chapters 3 and 4, I present two case studies demonstrating how my proposed notion of conceptual function can serve as a useful theoretical resource in conceptual ethics. Chapter 3 focuses on its application to cross-linguistic conceptual ethics, i.e., the assessment of representational devices across different languages. I argue that functional similarities between expressions used by distinct linguistic communities can serve as a criterion for determining whether an expression in an unfamiliar community is sufficiently similar to one in the ethicist's own community, thereby allowing for reliable extrapolation of judgements about the latter to the former.

In Chapter 4, I apply the motivation-based notion of conceptual function to the discussion of thick ethical concepts. I first show that this notion sheds light on a neglected but useful distinction between two ways of characterising the properties that expressions track: their real definitions and their significance-explaining characterisations. I then argue that this distinction can clarify the debate over whether the evaluative and non-evaluative components of thick ethical concepts can be disentangled. This question is indirectly relevant to conceptual ethics, as understanding the relationship between these components enables us to better assess thick concepts.

Part II of the dissertation comprises two chapters on the ethics of conceptual abandonment — the subfield of conceptual ethics that investigates when expressions become so critically defective that it's preferable to abandon them. In Chapter 5, I examine a specific argumentative

strategy on which some conceptual abandonment proposals rely. The strategy involves identifying a mismatch between what, if anything, an expression actually refers to and what its users, on reflection, want it to refer to. While such a mismatch is sometimes argued to justify abandoning the expression, I argue that this is so only under two relatively demanding conditions. First, abandoning an expression must be clearly a more feasible strategy than semantically adapting it to match its users' referential expectations. Second, even if an expression resists semantic adaptation due to being associated with categorically unfulfillable referential expectations, it may still serve a useful role for its users — so the costs of its continued usage must outweigh its benefits.

In Chapter 6, I critically engage with Cappelen's (2023) recent argument for abandoning the words 'democracy' and 'democratic' (D-words), focusing primarily on his claim that D-words likely fail us semantically by being meaningless. Against Cappelen, I defend an alternative diagnosis: D-words are partially semantically settled expressions that foreground the category of fairness in people-centred decision-making. By foregrounding this category, they provide various epistemic and practical benefits to their users, which makes them worth retaining in our conceptual repertoire.

Finally, in Part III (Chapter 7), I address a core problem for conceptual ethics: conceptual assessments must be sensitive to epistemic and moral considerations that often pull in opposite directions, yet they are so different in kind that it's unclear how to adjudicate between them non-arbitrarily. I argue that we shouldn't expect epistemic-moral conflicts in conceptual ethics to be resolvable through comparative reasoning, as this method presupposes that an overarching standard for such comparisons — one whose existence and epistemic accessibility are questionable. I examine three strategies for resolving epistemic-moral conflicts through comparative reasoning but contend that they ultimately fail. Instead, I argue for the promise of what I call the 'integrative approach'. This approach is based on the idea that epistemic-moral conflicts often stem from contingent factors agents can modify to promote convergence in the overall verdicts of epistemic and moral standards regarding how an expression should be designed. I suggest that once we recognise what kind of tools linguistic expressions are, the potential for integrating epistemic and moral standards in conceptual ethics proves greater than it initially appears.

Part I – Conceptual Function and its Applications

Chapter 1. Unpacking Conceptual Function as Tool Function

1.1. Introduction

It has become increasingly common to argue in the conceptual engineering and ethics literature that when assessing representational devices and proposing their revisions, we must take into consideration the functions they serve for their users (e.g., Prinzing, 2018; Simion & Kelp, 2020; Thomasson, 2020; Haslanger, 2020a, 2020b; Nado, 2021a; Riggs, 2021; Jorem, 2022; Queloz, 2022; Köhler & Veluwenkamp, 2024; Zuber, 2025). This idea can be called 'function-sensitive conceptual engineering'. The objective of Part 1 of this dissertation is to defend this idea with respect to conceptual ethics in particular, arguing it's theoretically fruitful for conceptual assessment to be conducted in a function-sensitive way. This chapter sets the stage for my discussion by examining how the notion of conceptual function should be interpreted.

What motivates function-sensitive conceptual ethics? There's an ongoing debate as to what, if anything, makes the appeal to conceptual function theoretically useful; but its basic rationale strikes me as easy to appreciate.[5] As the above definition of conceptual engineering indicates, theorists in the field typically adopt the *instrumentalist approach* to what they assess: they consider themselves to be assessing *representational tools.*[6] Yet, generally speaking, it seems unlikely one can properly assess how well a tool should be designed in order to be useful for its users, or how useful it currently is without considering its function. Think, for example, of someone assessing how a hammer should be designed to be a useful tool while being completely unaware its function is to deliver impact force to a small area. Alternatively, think of someone assessing whether weapons are tools worth producing without asking what their function is. In both cases, the assessments are likely to go astray because they overlook the key information for determining

[5] Theoretical roles associated with conceptual functions include serving as a criterion for individuating concepts (Prinzing, 2018), explaining when two concepts concern the same subject (Sundell, 2020; Thomasson, 2020; Haslanger, 2020b), setting the limits of permissible conceptual revision (Haslanger, 2012, ch.7; Nado, 2021a; Jorem, 2022), explaining metalinguistic negotiations (Plunkett & Sundell, 2013), explaining the success conditions of conceptual innovation (Simion & Kelp, 2020), and explaining why we have reasons to adopt certain concepts over others (Queloz, 2022; Köhler & Veluwenkamp, 2024). For a sceptical perspective on whether conceptual functions can play some of these roles, see Riggs (2021).

[6] While conceptual engineering is predominantly seen as targeting *representational* tools, some exceptions exist: Löhr (2021) and Jorem and Löhr (2024) see it as targeting inferential tools, and Thomasson (2025) calls on conceptual engineers to recognise that language is a multipurpose tool whose many parts serve non-representational functions.

whether and when the tools in question are useful. Thus, the facts about the usefulness of tools are closely tied to the facts about their functions, and it seems very natural to extend this point also to representational tools.[7]

While the instrumentalist approach lends plausibility to function-sensitive conceptual engineering, it also prompts us to ask the following question: How exactly should conceptual function be interpreted to make sense of the idea that conceptual engineering is concerned with representational tools? Considering that representational tools are a subset of human tools, we should, plausibly enough, interpret their function in continuity with how we generally expect the functions of tools to behave. Admittedly, representational tools may have a functional profile that is, in some respects, distinctive compared to non-representational tools. Nevertheless, as long as we adopt the instrumentalist approach to conceptual engineering, we cannot treat conceptual functions as entirely divorced from the functions of other human tools. Thus, an adequate interpretation of conceptual function must respect at least those observations about tool behaviour that seem general enough to be reasonably extended to representational tools.

Accordingly, this chapter explores how the notion of conceptual function is best interpreted in accordance with the instrumentalist approach. The chapter is organised as follows. In §1.2, I introduce four desiderata for an adequate interpretation of conceptual function that matches the general profile of function possessed by human tools. In §1.3, I use these desiderata to assess two interpretations of conceptual function that reflect dominant trends in how functions are understood elsewhere in philosophy: the proper-function interpretation, and the system-function interpretation. I demonstrate that both interpretations fail to satisfy some of these desiderata. In §1.4, I instead propose a motivation-based interpretation of conceptual function. This interpretation understands conceptual function in terms of those of an expression's actual, potential, or putative effects that are central to why a user group is, often implicitly, motivated to retain it in their conceptual repertoire. I argue that this interpretation, as a promising hybrid alternative to the proper-function and system-function interpretations, meets all four desiderata

[7] The fact that determining a representational tool's usefulness requires information about its function doesn't mean its function cannot be legitimately criticised and revised. After all, such information might reveal that the tool's present function is useless or even undesirable (cf. Nado, 2021a, p.1519; Jorem, 2022, pp.8-10, Koch, 2023, p.2139).

1.2. Four Desiderata for Interpreting Conceptual Function Adequately

In this section, I present and motivate four desiderata for an adequate notion of conceptual function, drawing upon some general observations about the functions of human tools. These desiderata will then guide us in our search for the best available interpretation of conceptual function, in the subsequent sections.

Firstly, whenever a human tool possesses a function, this function sets its *use and design standards*, i.e., the socially contingent standards against which we can evaluate whether it's used and designed in a way that enables it to produce the effects that its users generally expect it to produce when using it.[8] As an example, consider a mundane tool such as a teapot. The function of teapots — to serve tea — sets their use standards: teapots are used properly when employed for brewing tea leaves and pouring infusions, and used improperly when employed for other purposes, such as planting flowers or drinking beer. Also, this function sets their design standards: teapots are properly functioning teapots to the extent that they are designed to possess relevant features that enable them to be effectively used for brewing and pouring tea, such as a snug-fitting lid to prevent spills or a well-shaped spout for controlled pouring.

This brings us to the first desideratum: an extensionally adequate interpretation of what kind of effects qualify as a tool's function should yield plausible verdicts about its use and design standards. To illustrate, consider a bizarre interpretation of physical tools as having the function of producing whatever effects would make dogs happiest, if they were used to produce these effects. Such an interpretation would clearly be a non-starter, if only because, for many tools, it would yield verdicts about their use and design standards that are highly implausible *in virtue of being entirely disconnected from the expectations that we, as their users, have of them*. Obviously, flutes aren't properly used when employed as throwing toys for dogs; saucers aren't properly designed when designed as dog food bowls; and walking leashes aren't properly used when employed as pulling ropes in games with dogs.

[8] Note that such socially contingent standards may sometimes fail to track the normative facts about what effects a tool *ought to* produce in order to be a genuinely valuable tool for its users. Hence, they can only be used to assess a tool as properly functioning and properly used *relative to the imperfect expectations of its users*.

By the same token, an interpretation of conceptual function withstands scrutiny only if it yields plausible verdicts about what it takes to properly use and design expressions in specific cases. However, there is a noteworthy difference between expressions and non-linguistic tools. The design stage of non-linguistic tools generally precedes their use stage: a teapot is firstly designed to possess certain features that enable its proper functioning, and it's only subsequently used to produce the relevant effects, i.e., brewing and pouring tea, through these features. By contrast, expressions are designed *directly in the process of being used*. That is, the effects an expression can produce are determined by such factors as what descriptions its users associate with it, which things they apply it to, what valences and emotions it elicits by virtue of the associated descriptions and objects of application, or how it guides users in their actions. All these factors apply to an expression only if it's being circulated in a linguistic community, as a resource that its members can use in their thinking and communication.

The above point should be kept in mind when identifying the standards for an expression's proper usage and proper design, as it means that these standards overlap: to assess whether an expression is designed in a way that enables its proper functioning, we need to assess whether it's properly employed in its users' linguistic practices. While we can imagine a properly functioning teapot that none of its users can use properly due to their clumsiness, a community's systemic failure to use an expression in conformity with its function directly impedes its proper functioning.[9]

Let's now turn to the second desideratum. It seems theoretically useful to ascribe functions to tools, insofar as this allows us to cohesively identify a small number of effects they are primarily meant to produce, and to distinguish these from the endless variety of ancillary effects they can also bring about (cf. Wright, 1973, p.141; Houkes & Vermaas, 2010, p.5; Millikan, 1989, p.291). For example, characterising teapots as having the function of serving tea highlights their primary use, as opposed to more incidental potential uses — for instance, as paperweights, decorative objects, percussion instruments, or plant pots. Alternative theories that I discuss below then

[9] This doesn't blur the line between an expression's misuse and its malfunction. Even perfectly functioning expressions can be misused on occasion; they count as malfunctioning only when systematically misused by a significant portion of users (cf. Goetze, 2021, pp.25-26).

provide different accounts of what makes the primary use(s) of a tool — identified as its function — more important than its other uses.

It likewise seems important for conceptual function to be a compact category that serves to identify a small number of effects associated with expressions, which, in some relevant sense, encapsulate what they are primarily meant to be used for. If conceptual functions encompass multifarious effects that expressions might produce when used on particular occasions, then the notion of conceptual function seems like a disunified umbrella term, which no longer carries the identificatory virtue it holds for non-linguistic tools.[10] In that case, we might legitimately question what theoretical gains the appeal to conceptual function actually offers. Perhaps we should dispense with the talk of conceptual function and simply speak of individual effects of expressions.[11]

Next, the third desideratum. One of the appeals of characterising human tools in terms of their functions is that this allows us not only to identify their primary uses, but also to reconstruct the origins of various practices involving the instrumentalisation of objects. Agricultural practices offer a clear example. What marks the shift of human societies from foraging to subsistence farming, and thus the beginning of agricultural practices, is the invention of early agricultural tools such as digging sticks, whose function was rudimentary soil manipulation (López-Bultó et al., 2020). As societies transitioned to settled agriculture, these evolved into more complex implements, such as ploughs, which enabled deeper tillage and expanded both arable land and crop yields (Sherratt, 1981). Later, new irrigation and harvesting tools were developed in response to environmental factors and shifting social needs (Sojka et al., 2005; Anderson & Peña-Chocarro, 2014). This illustrates how the evolving trajectory of agricultural tools and their functions mirrors the increasing complexity of human agricultural practices.

[10] There are two conceivable ways of individuating conceptual function as a compact category. One way is to treat it as a category encompassing the plurality of effects that an expression is primarily associated with; the other is to treat each such effect as a distinct function. On the first approach, each expression has one function with plural content; on the second, multiple functions with singular content. I adopt the first because it foregrounds the interplay between various effects primarily associated with an expression, but little depends on this choice — readers are welcome to count one primary effect per function if it feels more natural.

[11] A similar point underlies Cappelen's (2018, ch.16) function-specification challenge, which I address in Chapter 2.

Similarly, it seems reasonable to expect conceptual function to serve as a theoretical resource for reconstructing the origins of human representational practices involving language. Conceptual function is up to this task only under an interpretation that is applicable to what might be called 'primordial expressions' — the expressions that emerged alongside the earliest forms of human language. To explain, consider the time when humans first began using language as a medium of representation, and ask how the first referential expressions might have emerged. We can imagine primordial speakers repeatedly uttering sound patterns like *STONE*, *TREE*, or *FOOD* while pointing to stones, trees, or pieces of food around them. Over time, these sound patterns became conventionally associated with certain types of objects; at which point we might say they evolved into expressions that represent stones, trees, or food.

Plausibly, the primordial speakers' practice of applying the sound patterns to objects around them must have been driven by some functional component, i.e., some initial purpose. Otherwise, the practice would appear to be merely an arbitrary convention, leaving us without an adequate explanation for how these sound patterns evolved into referential expressions. Since this evolution marks the beginning of human linguistic representational practices, identifying this component allows us to understand why humans began engaging in them. This understanding is crucial, as it provides a starting point for theorising about what drove further development of these practices. For instance, if we hypothesise that their initial purpose was to make human communication more effective, we are then led to reconstruct their development by tracing the increasing complexity of human communicative needs.[12] Consequently, an adequate theory of conceptual function must account for at least rudimentary functions in primordial expressions.[13]

Last but not least, the fourth desideratum for a conceptual function is that it should accommodate the phenomenon that Preston (1998, p.245; 2013, p.173) calls 'phantom functions'.[14] A phantom function of a human tool is an effect that the tool is used to produce, even

[12] This is just a tentative example. While effective communication is hypothesised by evolutionary linguists as one purpose driving language evolution (Pinker & Bloom, 1990), there are other potential purposes, such as maintaining social bonds and group cohesion (Dunbar, 1998), and facilitating internal thought processes (Chomsky, 2002).
[13] When first introduced, primordial expressions are a special case of newly created tool prototypes, which are often ascribed functions in the artefact literature (e.g., Millikan, 1999, pp.204-206; Kroes, 2012, pp.71-75; Vermaas & Houkes, 2003, p.266).
[14] For discussion of phantom functions, see also Holm (2017), Evnine (2016, pp.126-127), Parsons (2019), Juvshik (2023, pp.918-919), and Koslicki (2023, pp.226-233).

though it cannot in fact produce it. For example, feng shui mirrors are used to deflect bad energy, and amulets are used to bring luck. It seems attractive to characterise such tools' functions in terms of these putative effects, despite their unrealisability. After all, these effects play a theoretical role that tool functions typically play: as central effects for which the given tools are consistently used by their users, they help explain the practice of using them.

Moreover, some plausible candidates for tools possessing phantom functions can be also found among expressions. For example, when examining the history of 'race'-talk, Appiah (1996, pp.79-91) argues that a significant portion of the political and intellectual elite in the Anglosphere used the term 'race' for the purpose of tracking an essentialist kind, realised by a small number of large human groups sharing fundamental biological characteristics ('racial essences'), which were thought to explain their distinctive moral, intellectual, and cultural traits. Since the putative kind is biologically impossible, this purpose couldn't be successfully realised. Just as in the cases of the aforementioned non-linguistic tools, considering that this unrealisable purpose seems central to explaining why the given speakers engaged in the practice of using 'race', it seems appropriate to classify it as part of the term's function.[15]

1.3. Interpreting Conceptual Function Through Standard Approaches

In this section, I examine two candidate interpretations of conceptual function that follow dominant philosophical approaches to functions: the proper-function interpretation, and the system-function interpretation. I assess how these interpretations perform when tested against the four desiderata for conceptual functions outlined in the previous section, and show that neither satisfies more than two of the four.

[15] When discussing artefactual functions, Parsons (2019) and Koslicki (2023, pp.226-233) argue that an artefact's function can only include its realisable effects, and thus that phantom functions aren't genuine functions. When applied to conceptual function, this amounts to saying the real function of 'race', for the speakers in Appiah's example, wasn't to track the putative essentialist kind, but something else, such as oppressing those classified as belonging to inferior races. However, this interpretation is acceptable only if it can satisfactorily explain why speakers used 'race' without invoking its phantom function. After all, what makes it helpful to identify a tool's function is that doing so makes the practice of using it intelligible. Yet the interpretation overlooks something crucial: the oppressive use of 'race' was driven and post-hoc rationalised by the harmful belief that it tracked an essentialist kind marking some groups as inherently inferior (Mills, 1998). This supports an alternative interpretation of 'race' as having a dual function — comprising both the unrealised tracking effect and the oppressive effects.

1.3.1. Conceptual Function as Proper Function

The first interpretation of conceptual function under consideration views an expression's function as its *proper function*. An entity's proper function is standardly understood as consisting of the effects it's supposed to produce — such that failure to produce them counts as its malfunction— because they provide the best historical explanation for why it was selected for continued existence (Millikan, 1984, 1989). Accordingly, to speak of the proper function of expressions, we must clarify what it means for an expression to have been historically selected for continued existence in a language to produce certain effects. There are two dominant accounts of proper function that guide us towards different answers to this question. The first is the *intentionalist account*, which analyses the selection mechanism involved in determining proper functions as being governed by human intentions. The intentionalist account is defended in the literature regarding the function of artefactual kinds. Its proponents include Hilpinen (1993), Thomasson (2003, 2007, 2014), Dipert (1993), Baker (2007, ch.3), and Evnine (2016, chs.3-4). The intentions relevant to an artefactual kind's function are analysed by these authors as the intentions of those individuals who created it, i.e., its inventors. When the account is applied to a linguistic expression, the role of inventors is naturally assumed by the initial users who introduced it into the language. This leads us to the intentionalist account of conceptual proper function, I-Function:

> *I-Function*: An expression X has the proper function of producing effects E iff E are the effects that X's initial users intended it to produce.

A gesture towards I-Function can be found Prinzing (2018, p.869), who writes that 'something's function is what it was designed for', and goes on to suggest that concepts are individuated by the purpose for which they were designed.[16] More generally, I-Function can be situated within philosophy of language as a view that assigns an important explanatory role to initial users of expressions. It stands alongside the causal-historical theory in metasemantics (Kripke, 1980; Putnam, 1975), according to which these users are involved in fixing what expressions refer to; and originalist individuation theories, which state that lexical items or concepts are individuated by their historical origin (Stojnić, 2022; Sainsbury & Tye, 2012, respectively).

[16] Prinzing (2018, p.869) also writes that the design of some concepts can be explained in evolutionary terms, suggesting he defends a combined intentionalist-etiological account of conceptual function.

An alternative account of proper function is the *etiological account*. Its proponents, such as Wright (1973), Millikan (1984, 1989, 1999), Mitchell (1993), Neander (1991), Griffiths (1993), and Godfrey-Smith (1994), analyse an entity's function as determined by its reproductive history. What motivates the etiological account is an observation that biological and human kinds are often subject to various selective pressures from their environment, which make some of them more successful than others in reproducing themselves. Accordingly, the account suggests that such kinds' proper function consists of the effects produced by their past tokens, which historically gave them a selective advantage in competition with their alternatives, thereby causally contributing to their reproductive success. If we adopt the etiological account as an analysis of conceptual proper function, we arrive at E-Function.

> *E-Function*: An expression X has the proper function of producing effects E iff E were produced by the past uses of X and their past production contributes to the explanation of why X is successfully reproduced in its language.

Something like E-Function is embraced by Simion and Kelp (2020), who argue that conceptual innovation is successful to the extent that the introduced concepts are successfully reproduced to produce the effects that they were initially designed for, and their intentionalist function thereby evolves into their etiological function. Additionally, Thomasson (2020, pp.444-445) views the etiological account as a promising yet non-exhaustive approach to analysing conceptual functions.[17]

The intentionalist and etiological accounts of proper function can be treated as complementary rather than exclusive, as evidenced by the fact that some authors combine them in the discussion of artefactual proper function. For example, Millikan (1984, pp.17-19) distinguishes between *direct* and *derived* proper functions. An entity's direct proper function is established by the past reproductive successes of its ancestors. By contrast, derived proper function is inherited from the proper function of another entity. Millikan employs this distinction to argue that, while artefacts typically have their proper functions due to their reproductive history, artefactual prototypes that lack such a history derive their proper function from their makers' intentions (1999, pp.204-206; see also Griffiths, 1993, pp.418-419).

[17] See also Millikan (1984, ch.4), who, in a different context, argues that expressions have an etiological communicative function that provides a naturalistic explanation of their meaning.

The proper-function interpretation of conceptual function can echo this perspective by combining I-Function and E-Function. It can be argued that the expressions that have existed in a language long enough to become reproductively established possess a direct etiological proper function. In contrast, the expressions to which the etiological interpretation cannot be applied — either because they have entered the language only recently, or because they vanished shortly after entering — can be ascribed a proper function derived from their initial users' intentions.

It's time to test the proper-function interpretation against the four desiderata, starting with the first desideratum. Unfortunately, I-Function and E-Function seem inadequate with respect to this desideratum, as neither of them yields plausible verdicts about an expression's use and design standards whenever the effects for which it was introduced into a language, or for which it was successfully reproduced within it, no longer reflect its users' concerns.

To illustrate, imagine a community whose members originally introduced the expression 'afterlife' into their language in order to develop a conception of what follows death, which gives them emotional comfort by reducing death anxiety. Further, imagine that 'afterlife' was successfully reproduced in the community's language up to this point because it effectively served this purpose for them. Yet, suppose that the community's members have recently undergone transformative collective therapy that allowed them to entirely overcome death anxiety; thus, they no longer need 'afterlife' for its original purpose. Nonetheless, they decide to retain the expression for a different purpose — promoting a retributive moral framework that promises reward or punishment for earthly deeds after death.

The following question arises: Is 'afterlife' a properly functioning expression in the given community to the extent that it's used and designed to be effective for reducing death anxiety, or to the extent that it's used and designed to be effective for cherishing the hope that injustices in the present life will be rectified in the afterlife? Each option is likely to recommend a very different design and usage for the expression. For example, while the first might recommend associating it with a conception in which the misdeeds of the present life are largely forgiven after death, the latter might recommend associating it with a conception in which they are duly punished.

Contrary to what I-Function and E-Function predict, the second option seems more plausible than the first. That is, if we saw that the community adheres to a deeply emotionally comforting but unjust conception of the afterlife, we would conclude that they haven't designed

and used this expression in a way that enables its proper functioning. This intuitive response points us towards the following observation: It's natural for us to think of an expression's function in terms of those of its associated effects that somehow reflect its users' *present* concerns. After all, an expression's function is, plausibly enough, what makes the practice of using it at least seemingly pointful from the perspective of those who engage in it. But, if that's so, it would be strange to claim that an expression serves function X for its users even though whether X is realised is completely irrelevant to their present concerns.[18] As Queloz (2022, pp.1257-1260) notes, neither I-Function nor E-Function aligns well with this observation, as both adopt a *backward-looking* perspective — focusing on historical effects that merely explain why an expression entered a language or why it was successfully reproduced within it. As the 'afterlife' example demonstrates, there is no guarantee these effects in any way reflect what their users presently care about, which makes it implausible to classify them as its function.[19]

A supporter of I-Function might object that I underrated its potential to accommodate our intuitions about the 'afterlife' example. To explain, in the discussion on artefactual function, Evnine (2022, pp.5-6) argues that the intentionalist interpretation can, in fact, allow the intentions of an artefactual kind's later users to override those of its original makers in determining its function.[20] He thinks this is because later users can take on the role of inventors (cf. Houkes and Vermaas, 2004, p.66). Specifically, he argues that if later users start to intentionally 'counter-use' an artefactual kind by collectively using it in a non-standard way that diverges from its original makers' intentions, they can bring into existence a new artefactual kind whose function corresponds to its new intended use. Isn't the same move available to I-Function? Couldn't one say that, when the given community decides to repurpose 'afterlife', they coin a new expression — one for which they count as inventors, whose intentions determine its function?

[18] Relatedly, Smyth (n.d.) raises the concern that, insofar as the instrumentalist approach to conceptual engineering defines conceptual function independently of users' concerns, it promotes a wholly consequentialist method of conceptual assessment that neglects the agent-relative value concepts have for their users.

[19] Strictly speaking, Queloz's criticism comes from a slightly different angle. He proposes identifying an expression's function with the effects that reflect not just any of its users' present concerns, but those they would endorse upon reflection, thereby giving them a reason to adopt it. In §1.4, I argue that this stronger proposal is too restrictive.

[20] Evnine makes this argument in response to Koslicki's (2018, pp.227-228) counterexample to the intentionalist account of artefactual function, borrowed from Carrara and Vermaas (2009, p.135), in which the later users of the telephone use it for a different purpose from that of its inventor.

Unfortunately, this objection doesn't withstand scrutiny. For one thing, although, in the present example, the community repurposes 'afterlife' intentionally, we can easily imagine that they do so unintentionally, by changing the practical concerns that *tacitly* motivate their usage of the expression. In such a case, referring to the community's new *intended* use of 'afterlife' seems to overstate the extent to which its members are aware of the repurposing. Moreover, we could say that the community brought a new expression into existence through repurposing 'afterlife' only if words were individuated by their functions. How words are individuated is a complicated question that lies beyond this dissertation scope.[21] Nonetheless, function-based individuation seems inappropriate. To understand why, imagine that after repurposing 'afterlife', the community doesn't change its meaning. This could be because externalist metasemantics places the term's meaning beyond their control, or because what conception they associate with 'afterlife' doesn't affect its meaning. Then, we get the counterintuitive result that 'afterlife' becomes a different expression, despite its meaning and formal features remaining intact. That is, words would be too fluctuant entities if their identity changed every time their function changed.

Let's now turn to the second desideratum. I-Function and E-Function also perform poorly here. This is because, when applied to many expressions, both accounts tend to overgenerate the effects featuring in their function. As for E-Function, the full explanation of why many expressions persisted in a language arguably doesn't appeal to just a handful of effects for which their users uniformly reproduced them. Instead, it seems more likely that their users can be clustered into groups that reproduced them for different effects.

To illustrate, while the aforementioned 'afterlife' case serves as a neat toy example, the actual reproductive history of 'afterlife' is likely to have been more complicated. Different groups of English speakers presumably reproduced 'afterlife' for a variety of different effects. Some speakers might have indeed reproduced the expression for reducing death anxiety. Other speakers might have done so in order to promote the belief that there is a place where earthly deeds are rewarded or punished, thereby helping them uphold moral order. Alternatively, some speakers might have reproduced the expression simply because it carried cultural significance for them,

[21] For discussion of word-individuation, see notably Cappelen and Dever (2001), Kaplan (1990, 2011), Hawthorne and Lepore (2011), Irmak (2019), Miller (2021), and Stojnić (2022).

motivating traditions that reinforced community bonds.[22] These are just a few examples; many other effects contributing to the successful reproduction of 'afterlife' are conceivable. According to E-Function, the function of 'afterlife' encompasses all of them.

Furthermore, a similar proliferation of effects figuring into expressions' proper functions might also arise under I-Function. For instance, it wouldn't be surprising if historians found out that 'afterlife' was introduced into English by speakers with diverse intentions about its effects. While these speakers might have generally agreed that the expression served to refer to some form of continued existence after death, their differing conceptions of that continuation likely led to varied intentions regarding its further effects. Hence, both E-Function and I-Function render conceptual function too inclusive a category for it to be theoretically useful.

What about the third desideratum? If I-Function and E-Function are full-fledged accounts of conceptual function, they should accommodate the plausible hypothesis that early speakers introduced primordial expressions into language to serve a specific function. Unfortunately, I suspect that neither interpretation meets this demand. E-Function cannot do so, since primordial expressions lacked a history of successful reproduction immediately upon entering the language. But the prospects of I-Function seem scarcely more promising. Interpreting the initial function of primordial expressions through I-Function would amount to ascribing metalinguistic intentions to their initial users — intentions to use these expressions to verbally communicate about relevant objects in their environment. However, such an interpretation, I'm afraid, over-intellectualises the initial users' thoughts by overestimating the complexity of the intentions they could have formed at such an early stage of language development.

Suppose, for illustrative purposes, that the term 'stone' was introduced as a primordial expression. If so, I-Function entails that 'stone' had a communicative function for its initial users only if they introduced it with an intention whose content was something like [to use 'stone' for the purpose of verbally communicating about these objects], where 'these objects' refers to stones salient in a given context. Entertaining such content would require the initial users to have thoughts about the lexical item 'stone' and about the activity of verbal communication. Yet this seems

[22] The mitigation of death anxiety (Harding et al., 2005; Jonas & Fischer, 2006), the promotion of moral order through retributive justice (Johnson, 2015), and the maintenance of social cohesion (Sosis & Bressler, 2003) are among the effects studied in anthropology, social psychology, and religious studies as potential motivations for religious beliefs — including afterlife beliefs.

implausible, as it attributes to them a level of abstract thinking that is likely to have developed only later in history, when their linguistic capacities became more advanced. Hence, I-Function doesn't adequately capture the situation faced by primordial speakers when introducing the very first referential expressions of their language.

The only desideratum that the proper-function interpretation shows a promise of fulfilling is the fourth, which concerns the possibility of expressions possessing phantom functions. Even this, however, is true only for I-Function. I-Function allows that if the speakers who introduced 'race' into English did so with the intention of tracking an essentialist human kind, then this intended effect can form part of the expression's function, despite its unrealisability. E-Function doesn't offer the same advantage, as it requires that 'race' was successfully propagated over time at least in part due to actually tracking the given kind — which is impossible.

Overall, when tested against the four desiderata, the proposal to interpret conceptual functions as proper functions underperforms: it accommodates only one desideratum while portraying conceptual function as an overgenerating and backward-looking category.

1.3.2. Conceptual Function as System Function

Let's now turn to the proposal to interpret conceptual function in terms of Cummins' system functions (1975). On Cummins' view, entities can be ascribed a function relative to a system in which they operate, and this function can be identified with capacities through which they contribute to the exercise of this system's capacities (see also Nagel, 1977; Bigelow & Pargetter, 1987; Kitcher, 1993). For example, the heart's function in the circulatory system is to pump blood, thereby maintaining the system's capacity to circulate nutrients and oxygen. Alternatively, a fuel injector's function in an engine system is to deliver fuel to the combustion chamber, thus ensuring the engine's capacity to produce power.

The system-function interpretation of conceptual function is explicitly defended by Haslanger (2020a, 2020b) and Thomasson (2020, 2025). Haslanger (2020a, pp.250-256; 2020b, pp.250-255) seeks to demonstrate the utility of this interpretation by employing it to examine the expressions 'family' and 'marriage', arguing that we can critically evaluate and ameliorate these expressions only if we attend to their functioning within relevant social systems. By comparison, Thomasson (2025, ch.7) focuses on the functions that expressions serve within linguistic systems,

arguing that to identify these functions accurately, we must first understand the functions of language as a whole and then ask how the given expressions help realise them.[23]

Let's test the system-function interpretation against the four desiderata for conceptual function. Consider the first desideratum. As explained earlier, the proper-function interpretation yields implausible conclusions about how an expression should be used and designed to function properly in cases where its historical effects are no longer relevant to speakers in light of their present concerns. Can the system-function interpretation avoid a similar flaw?

As Queloz (2022, p.1260) points out, scepticism is warranted here if the systems relative to which conceptual functions are ascribed to expressions are impersonal systems, detached from what humans care about in their lives. However, the systems most relevant to expressions' functioning seem to be conceptual systems of their users. These are personal systems that don't develop in a vacuum but rather in response to their users' concerns. Accordingly, the core capacity of a conceptual system, which defines the conditions of its proper functioning, is its ability to serve its users' concerns. To think otherwise is to overlook the obvious truth that language, along with its constituent expressions, is a human tool. Humans are generally driven to use tools by desires, interests, needs, and practical problems they encounter. There is no reason to believe that linguistic tools such as expressions are an exception.

This observation paves the way for a version of the system-function interpretation that does justice to the close connection between an expression's function and its users' concerns. Queloz himself guides us to such an interpretation when he introduces what he calls the 'concern-relative function' (c-function) of concepts:

> *C-Function*: A concept X has the c-function of type C of producing effect E iff (1) users of X have among their present concerns — their needs, interests, desires, projects, aims, and aspirations — a concern of type C; (2) under propitious circumstances, applications of X produce E; (3) E stands in an instrumental relation to the concern of type C, which is to say

[23] Haslanger's and Thomasson's proposals are complementary, as Thomasson draws on systemic functional linguistics to argue that language is a social tool serving various regulative, interpersonal, and expressive functions, which must be uncovered and incorporated into critical concept evaluation (2020, pp.454-455; 2025, pp.210-213).

> that under propitious circumstances, producing E contributes to the satisfaction of a concern of type C (Queloz, 2022, p.1261).

Although Queloz presents C-Function as an alternative to the system-function interpretation of conceptual function, we can draw on it to formulate a concern-relative version of that interpretation, namely, CS-Function:

> *CS-Function*: An expression X has the system function of producing effects E within the conceptual system of a user group G iff each effect in E is such that: (1) under propitious circumstances, applications of X produce it; and (2) under propitious circumstances, producing it contributes to the satisfaction of some of G's present concerns.

By integrating the concern-relative and system-function interpretations, CS-Function shows that the two interpretations can complement each other. In fact, Thomasson is the first to recognise their complementarity, as she expresses sympathy for Queloz's concern-relative interpretation, viewing it as an enhancement of her own system-function interpretation (2025, pp.202-209). Specifically, she argues that examining various general concerns that a linguistic system serves for its users is a good starting point for understanding how its constituent expressions contribute to its overall functioning (2025, pp.203-204). Additionally, she argues that without considering more specific concerns that explain why individual expressions are useful in its users' linguistic systems, the system-function analysis wouldn't be fine-grained enough to track differences between their functions (2025, pp.204-205). CS-Function accommodates both points, as a group's present concerns that it appeals to may include both general and specific concerns.

Most notably, CS-Function ensures that relevant systems whereby expressions are ascribed functions tie in with the present concerns of those using them. This feature makes CS-Function immune to counterexamples similar to those that challenge I-Function and E-Function due to their backward-looking orientation.

Next, CS-Function also fares better than I-Function and E-Function in addressing the third desideratum, as ascribing a system function to primordial expressions doesn't seem problematic. For instance, we can say that 'stone' functioned for early speakers to enable verbal communication about stones because they started using the term in response to their need to develop a conceptual system for talking about salient objects in their environment, including stones. This need extended

beyond their consciously held attitudes and required only a basic awareness of the communicative activity they were engaged in. Such a minimal cognitive demand contrasts with the more sophisticated metalinguistic intentions posited by I-Function.

As shown, CS-Function outperforms I-Function and E-Function on the first and third desiderata. However, it struggles with the remaining two. CS-Function clearly falters on the fourth desideratum concerning phantom functions, as it requires that an effect can figure into an expression's function only if the expression is capable of producing it under propitious circumstances. This requirement risks obscuring the important explanatory role of unrealisable effects that nonetheless influence why speakers adopt certain expressions.

The most pressing challenge for CS-Function, however, is the second desideratum, which requires it to interpret conceptual function as a compact category. CS-Function's prospects of doing so are questionable. Admittedly, CS-Function *does* slightly mitigate the overgeneration problem that plagues the proper-function interpretation. As a pluralistic interpretation of conceptual function, it allows that what function is assigned to an expression varies relative to various subsets of its users individuated by some shared concerns. The number of the effects that qualify as parts of the expression's function relative to such subsets might be lower than, for example, the number of all the effects that contributed to its reproduction.

However, this number might often still be excessively high. After all, even subsets of users may share many different concerns. Consider once again the earlier 'afterlife' example. There are countless effects that the usage of 'afterlife' can produce for its users; I listed only a few of them above. Many of these effects address concerns that people widely share, even if they ascribe to them various degrees of importance, depending on their priorities. Hence, if we were to divide all the English speakers into groups based on what concerns 'afterlife' serves for them, each group would likely share many such concerns. Therefore, the system function of 'afterlife' would still overgenerate its constituent effects. This shows that we cannot solve the overgeneration problem unless we differentiate various concerns that an expression addresses, in terms of their importance for the expression's users.

1.4. Motivation-Based Interpretation of Conceptual Function

As we can see, neither proper-function nor system-function interpretations adequately capture the expected behaviour of conceptual functions *qua* tool functions. This raises the question of whether an alternative interpretation can fulfil all the four desiderata. To this end, I propose what I call the 'motivation-based interpretation of conceptual function' (M-Function).

> *M-Function*: An expression X has the function of producing effects E relative to a user group G iff (1) E are actual, potential, or putative effects of X, and (2) the realisation of E contributes to the satisfaction of those of G's concerns that are central to G's (possibly implicit) motivation to keep X in their conceptual repertoire.

What I want to show is that M-Function can be seen as a *hybrid* interpretation of conceptual function, drawing on both the system-function and proper-function interpretations by combining their strengths while overcoming their respective shortcomings.

First, M-Function resembles the proper-function interpretation in one respect. Like I-Function, it treats conceptual function as a resource for explaining why an expression exists in a conceptual repertoire, without requiring that the relevant effects explaining its existence be ones that the expression actually produces or has the potential to produce. Instead, it allows that users may reproduce an expression for putative effects it merely purports to produce. This makes M-Function well-suited to meet the fourth desideratum. Importantly, M-Function doesn't suggest that when investigating what function an expression has for a user group, we should disregard the effects it can in fact produce. In many cases, such effects can indeed help identify what motivates users to rely on the expression in their lives. M-Function only recommends that we also attend to what users *assume* the expression can do for their lives — whether or not those assumptions are true.

Where M-Function differs from the proper-function interpretation is in not relying on backward-looking explanations for an expression's presence in a conceptual repertoire. It avoids the unwarranted assumption that the factors behind an expression's introduction and past reproduction also explain its present use. In this respect, M-Function aligns more with CS-Function: both tether conceptual function to users' present concerns. This allows M-Function to satisfy the

first desideratum, as it precludes identifying an expression's function for a user group with effects unrelated to why they presently need it.

Also, the central motivation in M-Function is to be understood not as a consciously held attitude like intention, but rather as a dispositional mental state that speakers may possess *implicitly* due to their other mental features. That is, what centrally motivates speakers to use an expression simply depends on what interests, evaluative outlooks, desires, and needs most significantly influence their disposition to use it. Consequently, M-Function doesn't require speakers to consciously entertain the content of their central motivation for using an expression.[24] Thus, unlike I-Function, M-Function can treat primordial expressions as having functions while avoiding the over-intellectualisation of early users' thoughts.[25]

Perhaps the most notable virtue of M-Function is that it escapes the overgeneration problem that afflicts both the system-function and proper-function interpretations, and thereby fulfils the second desideratum. This is because M-Function is a pluralistic interpretation of conceptual function that is more restrictive than CS-Function, regarding which concerns of a user group are relevant to what function an expression has for them. Recall that, according to CS-Function, for an effect to be part of an expression's function, it only needs to satisfy *any* concern that the group might have. M-Function demands more. Specifically, M-Function allows an effect to feature in an expression's function relative to a user group only if it contributes to satisfying the concerns central to their collective motivation to keep it in their repertoire.

The following analogy with a non-linguistic tool helps clarify the kinds of concerns I have in mind: a ruler can have many uses that address different needs. It can be used as a measuring tool, a bookmark, a guide for cutting fabric, an aid for opening envelopes and packages, or even to scrape off dust, among other things. However, only some of the needs served by these uses are

[24] This isn't to say that, as a dispositional state, speakers' motivation for using an expression cannot be interpreted as a propositional attitude; rather, that such an interpretation shouldn't be taken as directly capturing the propositional structure of their motivation. Instead, it should be seen as an idealised model that helps us understand which concerns are central to their disposition to use the expression. This approach to interpreting unconscious states with a dispositional structure is defended in Crane (2017) and Crane and Thompson (2023).

[25] Additionally, the dispositional structure of the central motivation enables M-Function to account for what Queloz (2021, pp.54-59; 2019, pp.1136-1137) identifies as *self-effacing functionality* of some concepts. For example, if 'afterlife' serves the function of reducing death anxiety for a community, this function may be best fulfilled not when members consciously aim to achieve it, but rather when they conceive of themselves as using the term for less instrumental reasons.

central to most people's motivation for keeping a ruler in their toolbox. That is, while people might occasionally need a ruler for all these uses, only a few of the given needs are such that a person's disposition to keep or discard their ruler depends heavily on whether they have them. Arguably, few people would be disposed to throw away their ruler if dust no longer bothered them, but many might do so if they no longer needed to measure anything.

The same applies to linguistic tools like expressions. For instance, many users of 'water' are avid tea drinkers who use the term to identify and talk about a suitable liquid for brewing tea. Yet, if they stopped caring about tea, this is unlikely to significantly affect their disposition to retain 'water' in their conceptual repertoire, since they would still rely heavily on the term to satisfy more basic human needs. In contrast, if God fundamentally altered biological facts so that users of 'water' no longer needed to identify and communicate about water in order to survive, they would likely use the term significantly less often or even begin to question the point of using it altogether. According to M-Function, only the latter concern is central to explaining why speakers are motivated to use 'water' and thus relates to its function.

How does this help M-Function avoid the overgeneration problem? The key point to appreciate is that, according to M-Function, an expression can have a function only relative to a group of users *individuated by a shared central motivation for retaining it in their conceptual repertoire*. Hence, unlike CS-Function, M-Function doesn't aim to identify all kinds of ways in which an expression might be useful for its users based on how various concerns relevant to it are distributed among them. M-Function focuses narrowly on what *centrally* motivates a user group to use the expression, abstracting away from their peripheral concerns. Since only a few of all the relevant concerns are typically central to speakers' motivation for using an expression, M-Function imposes a selection constraint on which concerns count towards its function. This then rules out scenarios in which an expression's function encompasses too many effects for a given group. Consequently, the resulting notion of conceptual function that M-Function advances is sufficiently compact for it to be theoretically useful.

Two important clarifications about M-Function are in order. Firstly, as a pluralistic interpretation of conceptual function, M-Function allows a single expression to serve multiple functions relative to different user groups. But this pluralism isn't problematic. When there is no unified explanation for why an expression is used by an entire linguistic community, the proper

response isn't to cling to an invariant notion of conceptual function that creates the illusion of such an explanation. Rather, the proper response is to temper our expectations about the explanatory power of conceptual function, which is precisely what M-Function does. Moreover, M-Function can still deliver a unified explanation whenever one exists. Some expressions can indeed be used because all their users share a common central motivation for using them. In such cases, M-Function permits that an expression serves a single function across the board. Therefore, M-Function doesn't diminish the explanatory role of conceptual function; rather, it better delineates the scope of what it's supposed to explain.

Here is the second clarification. There clearly exist human tools that have objectionable functions, such as chemical weapons or torture devices. Moreover, this point plausibly extends to representational tools, as, unfortunately, many expressions have deeply objectionable functions and, for that reason, call for uncovering, critical assessment, and either abandonment or radical revision.[26] Terms like 'hysteria', 'poverty culture', 'welfare queen', 'race', and 'civilisation' are just a few examples. M-Function possesses two features that make it well-suited for analysing such expressions.

The first feature marks a point of difference between M-Function and Queloz's C-Function. As Queloz presents it, an expression's c-function can only include concerns users would endorse upon reflection (2022, pp.1265, 1268). This restriction serves to exclude concerns lacking normative grip for users — those formed through easily corrected misapprehensions or deception against users' interests, which would vanish once users recognised their origin (2022, pp.1268-1269). However, the restriction seems overly idealising, as it unduly rules out the possibility that some expressions have deeply objectionable functions due to containing a phantom component. Consider the earlier example from Appiah: it's plausible that, for many historical speakers, the phantom function of 'race' was to track putative racial essences. After all, their implicit interest in doing so — to rationalise their oppressive racist practices — largely explains why many speakers engaged in 'race'-talk. Unless we identify the tracking task with the phantom function of 'race', we risk neglecting this important consideration in our critical assessment of such speakers' 'race'-talk. Yet C-Function doesn't allow us to do so, since the speakers would likely have abandoned the

[26] For discussion of such expressions, see Thomasson (2025, pp.210-213) and Haslanger (2012, ch.7), who discuss oppressive functions tied to gender- and race-based classifications.

given interest upon reflection, having realised that it was unsatisfiable. By contrast, M-Function avoids this problem, allowing conceptual function to include concerns formed through false beliefs or deliberative error.

The second feature is the one that M-Function shares with Queloz's C-Function. The concerns that meet Queloz's restriction yield normative reasons. That is, if an expression's effect addresses a concern that its users would endorse upon reflection, they have a normative reason to use the expression to produce it. This implies that, even under M-Function, speakers might have a reason to adopt an expression whose function is objectionable, insofar as that function stems from concerns that they didn't form based on false beliefs or deliberative error — and thus concerns they would retain upon reflection. Yet, as Queloz (2022, p.1269) points out, these concerns might be merely *pro tanto* reasons — ones that are weak enough to be outweighed by countervailing reasons against adopting the expression. Hence, M-Function allows that speakers shouldn't adopt objectionable expressions when they lack an all-things-considered reason to do so.[27]

1.5. Conclusion

In this chapter, I investigated which interpretation of conceptual function best aligns with our general understanding of tool function. I first critically examined the proper-function and system-function interpretations of conceptual function, upon which I developed an alternative motivation-based interpretation that combines their strengths while overcoming their respective limitations. Still, M-Function isn't yet a fully developed notion of conceptual function. While it offers general guidance on how to identify an expression's function — namely, by focusing on the effects that centrally motivate a user group to retain the expression — more detail is needed on how to specify these effects. This issue is the focus of the next chapter.

[27] It's helpful to contrast M-Function here with Köhler and Veluwenkamp's 'normative function' interpretation of conceptual function (2024). On this interpretation, a concept's function consists of the effects it has when deployed in relevant circumstances, and which its users have objectively very good reasons to deploy it for (2024, pp.407, 420-423). These effects are different from those that serve its users' concerns whenever their concerns fail to track what is objectively good for those affected by the concept's usage. While this interpretation allows even objectively very good reasons for adopting an expression might be overridden by countervailing reasons (2024, pp.420-422), it implausibly implies concepts cannot have functions that are objectively very bad. This makes it unsuitable for analysing objectionable expressions (cf. Zuber, 2025, pp.15-24).

Chapter 2. The Function-Specification Challenge in Conceptual Ethics

2.1. Introduction

In Part I of this dissertation, I aim to develop a theoretically adequate notion of conceptual function and show how this notion can be useful for conceptual ethics. In Chapter 1, I took the first step in this direction by defending the motivation-based interpretation of conceptual function, M-Function, on which an expression's function for a user group is defined in terms of their central motivation for retaining it in their repertoire. M-Function is, however, currently an incomplete notion of conceptual function: it addresses the general question of what defines an expression's function, but not how to analyse it in specific terms. For example, it's one thing to interpret the function of 'family' for a group in terms of their central motivation for retaining the term, and another to specify exactly what that function is.

What makes the latter question especially pressing is that Cappelen (2018, pp.180-188) casts doubt on whether there is any theoretically significant answer to it. His sceptical challenge goes as follows. Although a single expression can generate a wide range of diverse, non-trivial effects depending on its context of use, none of these effects — or any subset thereof — is produced stably enough to serve as an uncontroversial candidate for its function. As a result, Cappelen concludes that the only uncontroversial way to specify the function of an expression 'A' is via the disquotational schema: 'A' functions to enable speakers to talk about A(s). For example, the function of 'family' is to enable speakers to talk about families, and that of 'responsibility' is to talk about responsibility. Yet such specifications are entirely uninformative, thereby undermining the theoretical significance of conceptual function for conceptual ethics. Thus, Cappelen claims, we face a dilemma: either we provide a disquotational and uncontroversial functional specification, which lacks theoretical depth; or we offer a theoretically interesting but non-disquotational one, which invites controversy. We cannot have it both ways — or so Cappelen contends.

Cappelen's function-specification challenge can be extended to M-Function. To be sure, M-Function doesn't leave us at a complete loss when trying to pin down an expression's function for a user group amid the vast array of its associated effects. It instructs us to specify it in terms of those effects that centrally motivate the group's usage of it. Nevertheless, it might be thought, *pace*

Cappelen, that the only plausible candidates for these centrally motivating effects are trivial, disquotationally specified ones. For example, what centrally motivates speakers to use 'family' may simply be that it enables them to talk about families. Beyond this clear but uninformative purpose, speakers may use 'family' for a wide range of different purposes across diverse situations, without assigning greater importance to any of them. If this diagnosis can be generalised, the very idea that certain non-trivial effects of expressions motivate their retention more strongly than others may be nothing more than a philosopher's fiction.[28]

I don't find the diagnosis under consideration to be well-motivated, as it implies a hard-to-explain asymmetry between how humans engage with expressions and how they engage with other human tools. In our daily interactions with ordinary tools, people rarely adopt an egalitarian evaluative outlook towards their various uses. Instead, we typically ascribe differing degrees of importance to these uses, depending on which best serve the situations that matter most to us. For example, a book might be used as a reading medium, a doorstop, a tool for pressing flowers, or a mouse mat. Still, people typically regard a book's use as a reading medium as most consequential for their motivation to keep it. It's difficult to see why our evaluative outlook should differ when it comes to representational tools like expressions.

Nonetheless, it would be premature to conclude that M-Function defeats Cappelen's scepticism. Cappelen might argue that even if an expression's central effects are stable for individual speakers, they exhibit too much *interpersonal variation*. An expression's users can widely disagree on which effects render it worth using. This often makes it difficult to identify recurring motivations for using an expression, on the basis of which its users can be categorised into subgroups that primarily use it for the same *non-disquotationally specifiable* effects. Even if such subgroups exist, they are often small and homogeneous, limiting their theoretical relevance for conceptual ethics. Conceptual ethics is arguably most interesting when it asks what an expression should be like for a user group that is heterogeneous enough to disagree on this question.

[28] Even this diagnosis is compatible with the deflated notion of conceptual function independently advocated by Nado (2021a), Riggs (2021), and Jorem (2022). As they see it, an expression's function can only be specified in relation to individual situations and the purposes for which it may be useful in them. By contrast, M-Function is less relativised: it ties function to user groups, allowing for a stable function across situations within those groups.

Thus, it would be disappointing if, under M-Function, expressions served a non-disquotationally specifiable function only within a few very homogeneous user groups.

In what follows, I argue that such interpersonal variation doesn't sufficiently motivate scepticism about conceptual function as defined by M-Function. My strategy is to show that, although disagreements about what makes expressions worth using are common, it would be a misinterpretation to conclude that they prevent us from specifying the central effects for which a heterogeneous group of users collectively employ an expression.

Here is the chapter's roadmap. In §2.2, I introduce a non-disquotational schema for specifying conceptual functions in line with M-Function, drawing upon the idea that a user group's central motivation for using an expression derives from the perceived significance of the category that they expect it to represent and from its extra-representational effects. In §2.3, I show that, while the schema allows users to disagree about an expression's function, it resists Cappelen's scepticism because these disagreements are more plausibly interpreted as concerning not what its function is, but rather its more specific aspects. In §2.4, I argue that, unlike the disquotational schema, the proposed schema boasts some theoretical virtues, making the appeal to functions in conceptual ethics attractive. In §2.5, I address two objections to the proposed schema.

2.2. Specifying Conceptual Function

In this section, I introduce what strikes me as the best possible schema for specifying the function of expressions in line with M-Function. This schema will provide me with theoretical resources to address Cappelen's scepticism about conceptual function in the subsequent section.

2.2.1. Two Clarifications

To frame my discussion, I want to start with two important clarifications. First, it's important to remember that, in accordance with the prevailing trend in the literature on conceptual engineering, my focus here is on those expressions that serve as representational tools. There are two ways in which an expression can qualify as a representational tool.

Suppose that an expression is applied to an object. If the expression is a singular term, it represents this object as its referent. But if the expression is a predicate, it represents the object as a member of its domain of application in the world and time in which it's situated, i.e., its extension

in the given world and time. In the latter case, the information about what objects belong to an expression's extension across different worlds and times calls for an explanation of why the expression applies specifically to these objects and not to others. The most common explanation is that there is a property that all these objects instantiate in the worlds and times in which they are situated.[29] The expression then also represents the given property as its referent, i.e., it serves as a linguistic proxy standing for the property that enables us to verbally communicate about it. By doing so, it introduces a new distinction through which we can carve up the world.[30]

Hence, expressions function representationally at two levels. First, they represent the objects they are applied to — either as their referents if they are singular terms or as the members of their extensions if they are predicates. Second, in the latter case, they also represent properties in virtue of instantiating which the objects in their extensions belong to them. Due to space constraints, I follow common practice in conceptual engineering and focus solely on expressions functioning as predicates, leaving open the possibility of extending the argument to singular terms.[31]

Secondly, when I characterise expressions as 'representational tools', I don't mean that they are *merely* representational tools. What I mean is that expressions serve to produce effects that involve representation, whether as their final end or only as a means to some further extra-representational end. For example, an expression may represent a property to direct attention, aid communication, or promote social norms. Thus, speakers can value expressions as representational tools while still expecting them to do more than represent.

[29] Some properties may also be referred to as 'kinds'. While the distinction between properties and kinds is somewhat murky, a common view is that properties are entities instantiated by objects, whereas kinds are clusters of properties shared by objects based on systematic similarities between them. For a comprehensive discussion, see Boyd (1991). Here, I will refer to what an expression represents as either a 'property' or a 'kind' depending on which description is more common in the literature.

[30] The notion of representation I employ here is broader than what Price (2011, p.20) calls the 'external notion of representation', which understands representation as co-variance with some external factor or environmental condition in the world. By contrast, my account allows that an expression may represent not an actual feature of the external world but rather a putative feature — such as a possible object or property — that helps us conceive of how the world could be carved up under certain counterfactual conditions.

[31] Additionally, I'm neutral in the dispute between Frege (1892/1980) and Russell (1903) over whether singular terms can refer to properties. Should this be possible for some singular terms, I see no principled reason why my argument couldn't be extended to them.

2.2.2. Representational and Extra-representational Effects

Let's consider what might prompt speakers to use an expression to represent a specific property. Merely 'stumbling upon' a property — that is, finding it instantiated or even just conceivable — doesn't seem sufficient to spark interest in representing it. Although there are countless actual and conceivable properties, we aren't interested in representing all of them.

To illustrate, consider the following example, similar to one given by Habgood-Coote (2019b, p.695). The property whose instances are consumable and at least minimally nutritious is most likely conceptualised by an expression in all human languages. In English, we refer to it by the expression 'food'. However, there are alternative properties in its vicinity which aren't commonly conceptualised in human languages. Consider the property of being consumable but not even minimally nutritious. Or consider even a more 'gruesome' property such as that of being consumable, at least minimally nutritious and observed before time t, or being consumable, not even minimally nutritious and unobserved before time t. This difference calls for explanation: why do humans use 'food' to represent the first property but not the latter two?

A natural answer is that it's only the former property that we think renders the expression 'food' worth using. That is, we consider it to be valuable to have an expression in our language that represents the property shared by the objects that can be consumed by animals and whose consumption provides the energy necessary for their subsistence. By contrast, the latter two properties aren't significant for us in a similar way, as we don't consider it valuable to have an expression that represents the property shared by the objects that are consumable but whose consumption is ultimately useless for our subsistence, or the property that groups together consumable objects not only depending on their minimum nutritiousness but also on when we observe them. Conceptualising these properties in our language just seems wasteful, given our limited conceptual resources.[32]

This example highlights that a group's motivation to use an expression can be largely explained by the fact that its members are disposed to consider the category they use it to represent

[32] Metaphysicians who, *pace* Lewis (1984) and Sider (2011), prefer to evaluate properties based on their naturalness might regard the latter two properties as insignificant not only for practical reasons but also because, being structurally disjunctive, less explanatorily powerful, and less generalisable, they fail to carve the world at its joints. However, the idea that a property's significance primarily depends on whether it's joint-carving can be misleading, since, as Barnes (2014, pp.335-341) argues, joint-carvingness seems ill-suited as a criterion for assessing social kind concepts.

to carry a *certain significance*, which makes it worth conceptualising.[33] Accordingly, the group expects the category to satisfy *certain desiderata* that explain why it's significant. Given M-Function, the expression can be described as having the function of representing the category that best satisfies these desiderata. As our example shows, such desiderata aren't disquotational: what motivates us to use 'food' isn't an interest in representing food *per se*, but rather an interest in representing *the property whose instances are consumable and provide the energy necessary for our subsistence*. Even if both interests are directed towards the same property, the latter presents it in a more informative guise that explains why we find it worth conceptualising.

In line with the second clarification, note, however, that a group might find an expression worth including in their repertoire not only because it's supposed to represent a property meeting certain desiderata but also because it's supposed to produce further extra-representational effects. Accordingly, M-function invites us to interpret an expression's function as partly extra-representational. But, to be clear, by 'extra-representational effects', I mean effects that depend in part on an expression's representational capacity, but also extend beyond it. Therefore, interpreting an expression's function as encompassing extra-representational effects doesn't commit us to the neo-pragmatist project of explaining the function of expressions with representational appearances but without treating them as genuinely representational.[34] Instead, we can treat such expressions at face value as representational, while focusing on how their representational capacity enables them to do more than merely represent. This approach allows us to analyse conceptual function in terms of the interplay between representational and extra-representational effects, rather than focusing exclusively on one or the other.

Let's consider some examples of how an expression's function can incorporate extra-representational effects. Sometimes, an expression's extra-representational effects may simply involve drawing its users' attention to a property and its instances, thereby allowing them to communicate about these properties or to study them. Consider, for example, the properties that

[33] This point underscores the idea that metalinguistic disagreements about expressions are seldom exclusively or mostly about expressions (Plunkett & Sundell, 2021a, p.9). Rather, they are closely tied to object-level issues regarding properties and kinds that expressions represent.

[34] Prominent defenders of the neo-pragmatist project include Brandom (1994), Blackburn (2013b, 2017), Price (2011), Price et al. (2013), Macarthur and Price (2007), and Williams (2011).

specialised scientific terms track. Plausibly, what primarily motivates scientists to keep these terms in their jargon is simply that they are useful for drawing attention to these properties (and their instances), discussing them, and studying them.

Yet, there are also some expressions for which it seems compelling to think that what primarily motivates us to use them is something other than attentional and communicative needs. One example is expressions whose function includes world-making effects. These are the expressions we are motivated to use because they directly influence the world around us even as they represent it. For example, we presumably use the expression 'family' in our language not only to represent a kind instantiated by groups of people who meet some desiderata, and to talk about them, but also in order to subject these groups to some norms regarding how they should organise their domestic lives (cf. Haslanger, 2020a, pp.250-251). After all, it's difficult to imagine our community's motivation to retain 'family' in our conceptual repertoire wouldn't significantly decrease if we didn't have a collective interest in normatively regulating people's domestic lives. Similarly, we probably wouldn't find the expression 'legally dependent' worth having in our repertoire if it didn't allow us to qualify individuals for various benefits, rights, or special considerations based on their relationship to another person. [35] [36]

Some expressions seem to have a world-making function even in a stronger sense, by being directly involved in constructing the social world. As many have argued, several social kinds are such that they couldn't exist if society were unable to entertain thoughts about them or their instances (Searle, 1995, ch.3; Thomasson, 2003; Khalidi, 2015). Yet, to think about them, society often must rely on certain linguistic items that represent them. For example, it's difficult to imagine a society where universities, money, MPs, or tax laws could exist without having expressions that stand for these kinds in their repertoire. Moreover, what likely motivates our use of such terms is, to a large extent, their role in constructing these very kinds: if we lived in a society where practical

[35] Another widely discussed example of expressions that regulate people through norms is gender terms, whose usage is often analysed as being motivated by society's interest in subjecting those to whom they are applied to certain norms of treatment (e.g., Haslanger, 2006; Gheaus, 2023).

[36] Relatedly, there is extensive discussion of how classifying people into social kinds and subjecting them to associated norms can change their self-understanding and behaviour (e.g., Hacking, 1999; Cooper, 2004; Allen, 2021). These changes may, in turn, prompt revisions to the classification itself. For example, labelling individuals as 'obese' may affect their self-perception, social interactions, or identity, leading to shifts in how 'obesity' is understood over time.

constraints made establishing universities impossible, we would probably use the word 'university' significantly less often than we do now.

Another example of expressions serving an extra-representational function involves those whose usage triggers various associations, emotions, memories, or narratives. Such expressions produce what Cappelen (2018, ch.11) calls 'lexical effects' — cognitive and emotional responses that build on but also extend beyond an expression's representational content.[37] For example, one likely reason the term 'mansplaining' gained traction is that, due to the meanings of its morphemes 'man' and 'splain', it immediately evokes a negative image of an arrogant man condescendingly explaining something to a woman (cf. Riggs, 2021, pp.11565-11568). Some expressions persist in users' repertoires precisely because of their negative lexical effects. For instance, drawing on Alexopoulos's 2017 book on Soviet gulags, Beaver and Stanley (2024, pp.437-439) discuss how mass violence in the gulags was obscured through bureaucratic language tied to technocratic ideals. Within the Gulag bureaucracy, the inhumanity of forced labour for sick prisoners was masked by the term 'labour therapy', which carried positive ideological associations. These associations were probably a primary reason why those involved in organising the gulags used the expression.

The last example of extra-representational effects I wish to highlight involves *action-guiding expressions*. These are the expressions whose applicability provides users with motivational reasons for or against taking certain actions. Bernard Williams (1985/2006, pp.140-141) identifies these effects in thick ethical concepts, which not only describe what they are applied to in a specific way but also morally evaluate it.[38] Depending on whether this evaluation is positive or negative, our disposition to apply a thick concept to an object affects the evaluative perspective from which we interpret it, often motivating us either to take action towards it or to refrain from acting in a certain way. For instance, if someone is disposed to use the term 'crime' for the act of stealing, this discourages them from performing it; whereas the disposition to describe an act of giving up one's seat to an older person as 'kind' gives them a reason to do so.

Action-guidingness shouldn't be seen as exclusive to thick concepts. It extends to most expressions that shape our background understanding of the environment by influencing the norms and possibilities for action — also called 'affordances' (Gibson, 1979) — that we perceive in it.

[37] See Landes (forthcoming) for discussion of lexical effects.
[38] See also Queloz (2025, chs.1-2) for discussion of the power and authority of action-guiding concepts.

For example, we see objects referred to as 'chairs' as affording the possibility of sitting and as being accompanied by some norms about how to sit on them. This, in turn, gives us a pro tanto reason to sit on them in a certain way. Of course, the reasons arising from affordances are sometimes too weak to motivate action on their own, but they still subtly shape the normative lenses through which we view our surroundings and guide our behaviour.

The above examples underscore a key point made earlier: we rarely value expressions solely for their representational features, independent of the extra-representational effects those features enable. These effects are central to why expressions are used. If conceptual ethicists overlook them, their normative engagement with expressions risks becoming overly simplistic. This is especially important because, as we will see later, many questions in conceptual ethics cannot be adequately addressed without considering their extra-representational effects.

2.2.3. The Variability and Multiplicity of Representational Desiderata

For simplicity, I assumed that the only relevant desideratum for the referent of 'food' concerned its practical value for our sustenance. But representational desiderata aren't limited to basic survival needs; they vary widely in the aspects of the world they track. Some are *epistemic*, sensitive to how well properties meet our epistemic needs. For example, 'knowledge' can be seen as a useful expression because it tracks a mental state that signals the end of inquiry (Kelp, 2014) or indicates good informants (Craig, 1990). Other desiderata might be characterised as *moral*, reflecting our attitudes about the well-being and dignity of others. Moral terms like 'kind', 'courageous', 'cruel', or 'selfish' are valued because they track properties warranting moral evaluation due to their impact on others' well-being and dignity. Still other desiderata can be described as *social*, as they concern the social world. Even the 'food' example above might be oversimplified, as linguistic communities arguably seek to track by it not just what is consumable and minimally nutritious, but also what fulfils a social function, such as serving as a centrepiece of gatherings. These examples illustrate that representational desiderata can be classified in many ways, depending on which aspects of the world they direct our attention to.

Importantly, the perceived value of many expressions depends on multiple desiderata for the category they track. This affects how complex the judgements are that one must make to properly assess what makes such expressions valuable representational tools. To illustrate, consider the following three desiderata for the kind we plausibly expect 'money' to represent. First,

we expect 'money' to represent a kind that functions as a medium of exchange. Second, we expect it to serve as a unit of account — to provide a standard measure of value for recording debts, keeping financial accounts, or comparing prices. Third, we expect it to function as a store of value — something that can be saved and retrieved in the future while maintaining its purchasing power. Properly assessing what makes 'money' a valuable expression, therefore, requires weighing these three desiderata in their respective importance.

Yet this isn't a simple task, as reasonable disagreements about which of these desiderata best capture what makes 'money' worth conceptualising seem conceivable. For example, David Graeber (2011) famously critiques the standard economic narrative that money arose to solve the inefficiencies of barter, arguing instead that its earliest purpose was to track obligations — such as tax payments, tributes, or soldiers' rations — long before it took the form of coinage or facilitated market exchange. His critique could be used to challenge the view that 'money' is useful primarily for its role in representing and constructing a medium of exchange, and only secondarily its role in representing and socially constructing a unit of account. Also, Passinsky (2021, p.285) discusses how economists disagree on whether bitcoin qualifies as 'money', depending on whether they prioritise the medium of exchange function or weigh all three functions equally. Such disagreements suggest that the conditions for assessing why 'money' is a valuable representational tool are more complex than if it were associated with a single representational desideratum.

2.2.4. The Value-based Schema

We are now ready to formulate the schema for specifying conceptual function as interpreted by M-Function. The effects central to a group's motivation for keeping an expression in their repertoire can be analysed in terms of the effect of tracking a property that meets certain desiderata, as well as various additional extra-representational effects produced along the way. Together, these effects capture the primary expectations of the group towards the expression that explain *why they treat it as worth using*. On this analysis, the function of predicate expressions can be specified through what I call the 'Value-based Schema'.

> *Value-based Schema*: The function of an expression 'A', relative to a group of its users G, consists of a representational effect R of tracking a property that best satisfies desiderata D_i-D_n, and of extra-representational effect(s) E_i-E_n, such that both R and E_i-E_n are central

to explaining why G's members regard 'A' as a valuable enough expression to include in their conceptual repertoire.

Recall from Chapter 1 that an expression's function sets its use and design standards. These are the standards for what Jorem (2022, p.106) calls its 'functional goodness' — they specify how an expression should be used and designed to properly perform its function. We can assess expressions based on how well they perform their function by examining how closely their actual usage aligns with these standards. This depends on how closely the effects of their actual usage approximate the effects the Value-based Schema identifies as those which users expect it to produce.

Importantly, how well an expression performs the representational and extra-representational components of its function often depends both on what *mental associations* surround it in its users' minds (such as its associated descriptions, bodies of information, paradigms, relations to other concepts, speaker meanings) and on what category its use is causally linked to, given what things speakers (or some relevant subset of them) tend to apply it to in their thought and talk.

To illustrate, suppose the function of 'knowledge' for a group of speakers is to represent the property whose instantiation is indicative of good informants (the representational component), and to facilitate the pooling of true information by enabling them to identify these informants (the extra-representational component).[39] I don't know whether what 'knowledge' (and similar expressions) *refers to* is the category that causally regulates its usage, or the category that best satisfies the descriptions that its users associate with it, whenever the two diverge. This is a contested metasemantic question that I cannot afford to address here.[40] Nonetheless, it seems plausible that, regardless of whether the reference of 'knowledge' is governed by descriptivist or causal metasemantic mechanisms, the expression can function optimally only if the group's

[39] This example is inspired by Craig (1990), who argues that the concept of knowledge serves to flag good informants, thereby helping us pool true information about the world.
[40] For discussion of this question, see, e.g., Jackson (1998) in the descriptivist camp; Kripke (1980) and Putnam (1975) in the causal theory camp; and Evans (1973) and Devitt (1981) somewhere in between.

members both associate it with descriptions compatible with the envisioned property's nature and causally link its applications to that property.[41]

To see this, suppose the group's members associate 'knowledge' with the descriptions satisfied by the property whose instantiation is indicative of good informants, but that their use of 'knowledge' isn't causally linked to the actual instances of this property because they predominantly apply it to false or luckily true beliefs. Then, they don't effectively use the expression for their desired representational and extra-representational ends, i.e., tracking the property indicative of good informants and pooling true information. Conversely, suppose the members are intuitively able to apply 'knowledge' to actual instances of the given property, but hold a misconception about what qualifies individuals as good informants. As a result, they associate 'knowledge' with descriptions such as 'knowledge can be acquired by luck' or 'knowledge doesn't have to be justified'. In that case, while they are able to use 'knowledge' to identify good informants, they fail to grasp what distinguishes good informants from poor ones. This is likely to hinder the expression's role in their thinking and communicating about the property, as well as in pooling true information. Hence, it seems that, for the optimal fulfilment of its function, 'knowledge' must be used so that its causal links and associated descriptions converge, selecting the same property — the one the group is interested in tracking with the expression.

Next, one might wonder how we can identify an expression's function for a user group through the Value-based Schema. To answer this question, I propose a systematic method for making reasonable hypotheses about what centrally motivates a user group to rely on an expression in their lives. My proposed method consists of two steps. In the first step, we need to collect empirical information about the user group and their interaction with the expression, and employ this information to identify those of their concerns that bear on their motivation to use the expression. The relevant information about the user group can be gathered by investigating what non-linguistic concerns — such as desires, interests, aspirations, evaluative outlooks, and needs — its members have, what capacities limit what they are capable of, and what circumstances they

[41] This point aligns with theorists who argue that even if an expression's meaning is externally determined, conceptual engineering — and thus conceptual assessment — can still legitimately target internal aspects of the expression, such as its speaker meaning (Pinder, 2021; Riggs, 2019), its cognitive content (Koch, 2021a), bodies of information associated with it (Isaac, 2021), their internal role in our conceptual network (Pollock, 2021) or classificatory procedure associated with it (Nado, 2021b, 2023).

are situated in.[42] By comparison, the relevant information about the group's interaction with the expression can be gathered by investigating such things as how the group uses it; what effects relevant to their lives its usage produces or is at least believed to produce; which of these effects they respond to positively and negatively; what claims they make about how the expression should be defined; which of these claims seem least negotiable for them; what inferences they tend to make when using it; what objects they treat as its paradigms and anti-paradigms; and what other expressions it's associated with, in their conceptual network. Taken together, this information provides us with a trail of clues as to which of the users' concerns the expression is used to address.

In the second step, we can use the gathered information to determine which of the identified concerns are the central reasons why the group's members are disposed to value the expression's presence in their conceptual repertoire. I propose we do this by asking one of the following two counterfactual questions, depending on which of them has a more easily conceivable antecedent. The first question is: which of their concerns are individually such that, if the users no longer had them, they would become stably disposed to use the expression significantly less often than they do now? The second question is: which of these concerns are individually such that, if the users discovered the expression was hopelessly incapable of addressing them, they would become stably disposed to use it significantly less often than they do now? Both questions effectively seek to identify those concerns whose presence and absence are consequential for whether the users are disposed to see the expression as worth using, thereby guiding us in reverse-engineering its function for them.[43] Of course, this two-step method isn't bulletproof, and its applicability may be limited by the amount of relevant empirical information accessible to us. Nonetheless, the method's availability demonstrates that the task of identifying an expression's function doesn't

[42] See Queloz (2025, ch.7) for discussion of how information about users' non-linguistic concerns, circumstances, and capacities can be used to identify the concerns (what he calls 'conceptual needs') to which their concept responds. Queloz (2025, pp.230-233) helpfully points out that, even if we lack information about one parameter of the concerns–circumstances–capacities triad, we can often reliably infer it from the remaining two. For example, if we know enough about the circumstances an expression's users face and the capacities that limit them, this narrows the search space enough to reliably identify the concerns they are likely to have.

[43] Another potentially useful way to reverse-engineer an expression's function through counterfactual reasoning is *pragmatic genealogy* (Queloz, 2021; Pettit, 2018). This method involves constructing an idealised state-of-nature story about which practical needs might lead individuals similar to the target speakers to introduce the expression. For the method to be reliable, however, it must draw on sufficient evidence about the target users and their current engagement with the expression. Otherwise, it risks tracing the expression's practical origin to needs that don't reflect the users' present concerns (cf. Smyth, 2023).

have to be speculative guesswork, but can be approached in a systematic, empirically informed way.

Let me illustrate how this method works in practice with a simple example of the term 'water'. We have an abundance of empirical information about the biological needs and capacities that users of 'water' have as humans, the concerns they possess due to how they organise their lives, the environment they inhabit, as well as how they engage with the term in their linguistic practices. Based on this information, it isn't difficult to identify the concerns that motivate them to use 'water': they use the term because they seek to track and verbally communicate about the substance, whose consumption is indispensable for their subsistence, whose presence on Earth is essential for the preservation of ecosystems, which has distinctive chemical properties of scientific interest, supports hygiene and sanitation, can be used in the preparation of beverages, plays a role in various cultural rituals, and serves as a medium for recreational activities such as swimming, among other things.[44]

What I described above are various communicative effects of using 'water' that are relevant to why people use the expression. However, the function of 'water' relative to a linguistic community, as per M-Function, includes only those of them that are central to the community's motivation to keep the term in their repertoire. This brings us to the second step of my proposed method. The method recommends that we ask which of the identified concerns are individually such that, if humans no longer had them, they would become stably disposed to use 'water' significantly less often than they do now.

It seems perfectly possible to reliably hypothesise about this question. Arguably, we have sufficient knowledge about which aspects of water matter most to humans. We know humans are biological creatures who must drink water to survive and who have developed hygienic practices to prevent the spread of disease. Given this, we clearly use 'water' in our everyday thoughts and utterances most often to track and verbally communicate about the transparent liquid substance that is drinkable and can be used for cleaning. We also know we are existentially dependent on water's role in terrestrial ecosystems: its presence in the physical environment is crucial for agriculture, industry, and transportation, as well as for the prevention of natural disasters.

[44] See Schroeter and Schroeter (2015, p.426) for a list of possible roles the referent of 'water' can play in human life.

Furthermore, we standardly treat ‘water’ as a chemical term whose paradigmatic instances have the H_2O structure. This indicates that we recognise that if we want to understand what enables water to perform all its other roles in human life, we must attend to its underlying chemical structure revealed in scientific laboratories.

This knowledge makes it very plausible to think that the desiderata most consequential for human disposition to use ‘water’ are those reflecting the significance of the substance it tracks for their survival (the subsistence desideratum), hygiene and sanitation (the hygienic desideratum), the terrestrial environment (the ecology desideratum), and chemistry (the scientific desideratum). That is, each of these four desiderata seems such that our disposition to use ‘water’ would be significantly affected if we were no longer interested in tracking and verbally communicating about the kind that satisfies it. By contrast, while humans certainly appreciate that ‘water’ enables them to refer to a substance used in preparing beverages, swimming, or performing rituals, it seems far-fetched to think that, if we no longer needed water to engage in any of these activities because we had lost interest in it, we would be significantly less disposed to use the term than we are now. Even avid swimmers, beverage enthusiasts, or ritual practitioners obviously rely on ‘water’ primarily as a tool for satisfying more fundamental human needs and interests than those tied to their idiosyncratic activities.

Now that we have completed our reflection on what centrally motivates human society to use ‘water’, we are ready to specify the function of ‘water’ for human society through the Value-based Schema. By filling ‘water’ into the placeholder ‘A’, human society into the placeholder G, the subsistence desideratum, the hygienic desideratum, the ecology desideratum, and the scientific desideratum into the placeholders D_i–D_n, tracking a kind satisfying these desiderata into the placeholder R, and the facilitation of verbal communication about the kind satisfying the desiderata into the placeholder E_i, we arrive at the following specification of the function of ‘water’ for our community, WS.

> *WS*: The function of ‘water’, relative to human society, consists of the representational effect of tracking a kind that best satisfies the subsistence desideratum, the hygienic desideratum, the ecology desideratum, and the scientific desideratum, and of the extra-representational effect of verbally communicating this kind, such that both these effects are

central to explaining why human speakers regard ‘water’ as a valuable enough expression to include in their conceptual repertoire.

The examples of ‘knowledge’ and ‘water’ presented above are neat and tidy. Yet it should also be recognised that the story behind why speakers value an expression’s presence in their repertoire is often messier than in these examples. There are three ways in which this messiness may arise. First, we shouldn’t expect that representational desiderata associated with an expression are always specific enough to single out a unique property that satisfies them. Often, speakers may have only a partial idea of what they expect an expression to represent. In Chapter 6, I argue that the term ‘democratic’ functions in this way: speakers use it to represent some property of decision-making systems that depends at least partly on whether those systems enable people to influence outcomes through fair processes — while leaving unsettled what exactly that property is.

Second, we shouldn’t assume that the representational desiderata associated with an expression are always mutually compatible and satisfiable. People often hold inconsistent or unrealistic expectations, and representational desiderata are likely no exception. For instance, many users of the term ‘sustainable’ may expect it to represent the property of promoting economic growth while minimising environmental harm. Yet one could argue that these two desiderata are in tension, as economic growth typically involves significant environmental costs. Consider also Appiah’s analysis of ‘race’ (1996) from Chapter 1. If correct, it suggests that for much of modern history, many speakers used ‘race’ to track an essentialist human kind in the world. However, our current understanding of human biology shows no such kind exists.

Third, the relationship between an expression’s representational desiderata and its extra-representational effects is often less straightforward than in the case of ‘water’. There, the term enables communication only if it successfully refers to a substance meeting the four desiderata. But other expressions may achieve their extra-representational effects *merely by attempting to represent something*, even if that attempt fails. For instance, many speakers may use ‘sustainability’ to guide their economic activities towards environmental friendliness. The term may serve this purpose even if the putative ideal that it’s expected to represent is incoherent for the aforementioned reasons. Similarly, racist speakers likely use racial terms to enforce oppressive norms based on imagined traits. Tragically, racial terms have historically proven highly effective for this purpose, despite failing to fulfil their representational promises. Recognising these untidy

connections between the representational and extra-representational effects is important for accurately understanding the conditions under which an expression fulfils its function.

2.3. Two Contentious Questions about the Value-based Schema

In this section, I employ the Value-based Schema to distinguish between two questions one can raise about conceptual function. This will help us approach the scepticism about conceptual function from a more nuanced perspective and ultimately mitigate its initial force. The scepticism stems from the concern that the representational and extra-representational effects centrally motivating an expression's use are too interpersonally variable to allow for an uncontroversial specification of its function — except, perhaps, within a few homogeneous user groups.

I'm happy to grant the sceptic that speakers use expressions to produce various different effects, which results in a lot of controversy about what their function is. Nevertheless, it would be too quick to conclude that this concession prevents us from specifying conceptual function relative to heterogeneous user groups. This is because speakers' attitudes regarding the generic question of what the desiderata for an expression's referent are may often converge even if their attitudes regarding more specific points concerning these desiderata diverge. I will clarify my point using WS.

Plausibly enough, the four representational desiderata in WS reflect those attitudes about the value of 'water' that our linguistic community widely shares. At the same time, WS leaves ample room for disagreement over more specific points regarding these desiderata. Namely, our responses to the following two questions may differ.

> *Interpretative Question*: How exactly should we interpret the content of the four desiderata?
> *Comparative Question*: How do the four desiderata compare to each other in their relative importance with respect to the value that 'water' has for us?

The answers to these two questions might be related but shouldn't be conflated. Let me elaborate on each question in turn.

The Interpretative Question is the crucial question for identifying the kind speakers expect 'water' to represent: different speakers might identify a distinct kind as the one tracked by 'water' even if they interpret only one of the four desiderata differently. For example, speakers might agree the scientific desideratum concerns the *chemical structure of the watery stuff around them*, yet still interpret what counts as 'chemical structure' differently. Some speakers might interpret it simply as the chemical composition of watery stuff, yet other speakers might interpret it as including not only its chemical composition but also some other microstructural properties related to its atomic mass and molecular structure. Such different interpretations of the scientific desideratum can result in disagreements over the membership conditions of the kind 'water' serves to track. For example, two individuals might disagree over whether the kind includes samples of deuterium oxide, an isotopic variant of H_2O that has some distinctive nuclear properties that also ground its distinctive manifest properties, such as lethality to some organisms when consumed at higher concentrations (cf. Hendry, 2006, pp.866-869; LaPorte, 2004, p.107).

Disagreements over how to interpret a representational desideratum are common. For instance, we might agree that 'knowledge' functions to represent the property indicative of good informants, yet still disagree on who qualifies as 'good informant'. Or we might agree that 'murder' functions to represent the killing of a human being with malice aforethought and no legal justification, but differ on what counts as 'malice aforethought' or 'legal justification'. And even for mundane expressions, like 'fruit', such disagreement is possible: we may agree it functions to represent the sweet edible part of a plant containing seeds, but disagree over whether a tomato counts as a fruit, depending on how we interpret 'sweet'.

Let's now turn to the Comparative Question. This question concerns the relative importance of each of the four desiderata in contributing to the overall value of 'water'. While WS states that 'water' functions optimally when it represents the kind that satisfies all four desiderata, this doesn't mean each carries equal weight. This is a general point: just because an entity requires multiple features to function optimally, it doesn't mean that all features are equally important in its functioning. For example, optimally functioning shoes need both soles and laces, but without soles, the shoes are unusable regardless of the laces. So, soles are more important to the functioning of shoes than laces.

Why does the Comparative Question matter? To see this, we must first realise that we may not always use 'water' optimally in accordance with all four desiderata. For example, imagine a substantial portion of the H_2O on Earth were replaced by a substance with the same observable properties but a different microstructure. As a result, speakers would find it difficult to consistently apply the term 'water' to H_2O, as they would frequently conflate it with the other substance. In such a case, 'water' would fail to optimally fulfil the expression's function as specified by WS because it would underperform with respect to the scientific desideratum.

For each of the other three representational desiderata in WS, we can imagine similar cases in which 'water' violates them. However, the extent to which the usage of 'water' deviates from optimal functioning in such cases may vary depending on the relative weight of the violated desideratum. The more important the desideratum, the further the usage of 'water' that violates it takes us from the expression's optimal functioning. Hence, the answer to the Comparative Question provides information about how close or far our suboptimal usage of 'water' is from its optimal usage.

Generally, such information can be useful for expressions we routinely struggle to use in optimal alignment with their function, despite being able to identify the desiderata relevant to that function. This is particularly relevant to the above-discussed expressions associated with representational desiderata that have contested interpretations. In such cases, knowing the relative weights of these desiderata can help us in settling on the correct interpretation of which desideratum is most urgent to ensure the expression's proper functioning. This, in turn, can assist in allocating our interpretive and ameliorative efforts more effectively.

That said, answering the Comparative Question can sometimes be just as challenging as answering the Interpretative Question. One reason is that the importance of a desideratum may only be determined after deciding how to interpret it. Moreover, even if there is agreement on its interpretation, there may still be disagreement over its relative weight. For example, suppose our linguistic community agrees that 'the chemical structure' in the scientific desideratum for 'water' refers simply to its H_2O chemical composition. Still, there might be disagreement over the relative weight of the scientific desideratum compared to the hygienic desideratum and the subsistence desideratum.

Some speakers may think the scientific desideratum should take precedence over the other three. They might argue that 'water' is a natural kind term whose value primarily derives from how joint-carving its referent is, and that chemical structure is the most joint-carving classificatory criterion. Others may disagree, arguing that the value of 'water' for our linguistic community primarily derives from the practical use of its referent in hygiene and subsistence. On this view, a hypothetical kind lacking the H_2O microstructure but usable for hygiene and subsistence might better fit the function of 'water' as its referent than an H_2O substance that is too polluted to be consumed and used for hygienic purposes.

The Interpretative Question and the Comparative Question certainly raise contentious issues about specific aspects of the function of 'water' as articulated by WS. However, it's important to recognise that such disagreements can occur only against the backdrop of consensus that 'water' is a valuable expression insofar as it represents the kind satisfying the four desiderata and facilitates communication about it.[45] These questions, therefore, presuppose agreement on the correctness of WS. In the case of 'water', such agreement is plausibly even global. As shown above, generic human needs motivate speakers to use the term to represent the kind that enables subsistence, is suitable for hygiene and sanitation, is essential to terrestrial ecosystems, and is defined by a specific chemical structure.

Surely, other expressions may enjoy far less agreement about which representational desiderata their referents should satisfy. Some divisive expressions may even be subject to such disagreement across the entire linguistic community. Yet even then, there remains a non-negligible possibility that there are some recurrent valuing patterns within the community, based on which we can cluster its members into subgroups that exhibit considerable heterogeneity but still collectively value an expression for shared generic reasons. That such subgroups are likely to disagree over the expression's function doesn't warrant dismissing this possibility. These disagreements may concern the Interpretative Question and the Comparative Question rather than the generic question of how to specify an expression's function using the Value-based Schema.

[45] Cf. Plunkett (2015, p.857; 2016, p.242) and Plunkett and Sundell (2013, p.20; 2021b, p.148), who also interpret many metalinguistic disagreements as occurring against the backdrop of an agreement about the function an expression serves.

Also, denying that heterogenous groups of speakers can form clusters that converge on generic reasons for which an expression is worth using would give rise to at least two explanatory gaps. As for the first gap, consider an everyday tool such as a blackboard. One very general way of characterising the function of blackboards is as serving as a reusable writing surface for displaying information. Yet it also seems possible to cluster the users of blackboards into subgroups that primarily use them for a common, more specific cause stemming from overlaps between the environments in which they organise their lives. For example, the people who work in corporate offices primarily use blackboards for brainstorming ideas and visualising workflows; those who work in educational institutions, for explaining ideas during instruction; those who work in design studios, for sketching visual layouts; and those who work in restaurants, for displaying menus and announcements. Given this, it seems very reasonable to also characterise blackboards as serving these more specific functions for their respective user groups. It would create yet another hard-to-explain asymmetry if we couldn't find similar commonalities among members of broader linguistic communities — shared ways of organising life, developed through interaction in the same environment, that give rise to unifying patterns in what primarily motivates their use of certain expressions.

Secondly, if speakers valued expressions for completely different reasons, one would expect successful communication by means of those expressions to be very difficult. But this often isn't the case. For example, our communication involving even contested expressions such as 'knowledge' seems relatively coordinated because, despite differences, the class of mental states to which we apply the expression is relatively narrow. If the representational desiderata associated with 'knowledge' were completely divergent, it would be puzzling how such coordination in our representational practices is even possible. It would be akin to explaining how two people managed to jointly carve a statue with chisels they each keep in their toolbox for entirely different reasons.

Given these explanatory gaps, it seems appropriate to view disagreements about the Value-based Schema as often concerning not which desiderata are relevant to an expression's perceived value for a group of speakers, but how these desiderata should be interpreted and weighed in their relative importance. If so, the Value-based Schema provides a way of specifying conceptual function that resists Cappelen's scepticism even across many heterogeneous groups. Even if these groups don't form an entire linguistic community, assessing how an expression should be used

within them remains a worthwhile *local* conceptual ethics project. Moreover, when speakers value an expression due to more generic human needs — as possibly illustrated by 'water' — the schema may also serve the *global* conceptual ethics project, which aims to assess how an expression should be used not just within a subgroup, but across the board.

2.4. Theoretical Virtues of the Value-based Schema

Since the Value-based Schema identifies an expression's function by providing information about the evaluative outlook of a user group, unlike the disquotational schema, it avoids reducing conceptual function to a trivial phenomenon. I will now demonstrate this point by flagging its three theoretical virtues.

The first virtue is related to functional continuity in conceptual engineering. I don't think that conceptual revisions are permissible only if they preserve the function of a target expression. As noted in Chapter 1, some expressions plausibly serve objectionable functions that should be revised if we wish to continue using them. More generally, insofar as an expression's function is given by what centrally motivates a user group to retain it, it's difficult to see why speakers should always trust their present motivations as the best possible ones, rather than remain open to reconsidering them whenever there are good reasons to do so.[46] Nonetheless, the Value-based Schema points to a different reason why conceptual ethicists should pay attention to whether their proposed revisions of an expression are continuous with its present function: doing so can help them better estimate how practically difficult it may be to persuade its users to adopt these revisions.

To explain, according to the Value-based Schema, conceptual function reflects a user group's evaluative stance on what makes an expression worth using. Accordingly, when a proposed revision aims to change an expression's function, it seeks to alter the group's stance on which representational and extra-representational effects are central to its value for them. Conversely, when the revision aims to preserve the expression's function, it seeks only to change the group's beliefs about when those effects are best realised. Unsurprisingly, the former type of revision typically requires more sustained argumentative effort than the latter: convincing people to

[46] See also Koch (2023, pp.2138-2139), who likewise criticises the view that conceptual engineering is permissible only if original conceptual functions are preserved. He discusses functional continuity as one possible reading of topic continuity but aligns with others who argue that changing topics during conceptual revision can sometimes be desirable (e.g., Prinzing, 2018, p.871; Knoll, 2020, pp.17-20; Belleri, 2021a, p.18; Nado, 2021a, p.1514).

reconsider their prior values is more demanding than persuading them to reconsider how best to realise those values. This difference is worth keeping in mind for conceptual engineers who aim to propose revisions that are, in practice, implementable.

Next, as noted earlier, an expression's function sets use standards. Hence, we can assess the expression in terms of how closely its usage conforms to these standards, given the effects it's conducive to. However, the Value-based Schema also unveils two other interesting dimensions along which an expression can be assessed in conceptual ethics, besides how well its actual usage aligns with its function. This leads us to its second and third theoretical virtues.

The second virtue is that the schema reveals that whereas an expression's usage is a matter of how speakers apply it, its function is also a matter of what values and ends its application serves to promote. Consequently, an expression can be used in exactly the same way in two groups, yet each group can place very different representational desiderata on it. Such disparity often also translates into differing *extra-representational* effects that the expression serves to produce for each group.

Consider two groups that apply the term 'prisoner' to the same individuals — those deprived of freedom through incarceration. Suppose, however, that each group holds a different view of the purpose of incarceration: the first sees it as a means of rehabilitation and crime prevention, while the second adopts a more retributive stance, viewing it as a practice designed to justly punish criminals. Accordingly, whereas the first group uses 'prisoner' to represent incarcerated individuals needing rehabilitation, the second uses it to represent incarcerated individuals deserving punishment.

This difference in representational desiderata will likely lead to a difference in the extra-representational effects that the expression is instrumental in producing for each group. Each group is likely to treat those they classify as 'prisoners' according to very different norms. A conceptual ethicist assessing which group has the better concept of 'prisoner' shouldn't focus solely on how each group uses the expression, because their usages are indistinguishable. They must also consider the function the term serves in each group, including both its representational and extra-representational effects. A virtue of the Value-based Schema is that it highlights the need to incorporate these considerations into conceptual assessment.

The third virtue of the Value-based Schema is that it frames an expression's function not in terms of the effects that objectively make it a valuable representational tool, but in terms of the effects a user group subjectively values it for. Since these judgements are fallible, the schema's output serves not only as a benchmark for assessing and revising actual usage, but also as a legitimate target for assessment and revision. As a result, it encourages conceptual ethicists to employ *the method of reflective equilibrium*, involving mutual adjustments between an expression's usage and its function. This method allows conceptual ethicists not only to align an expression's usage with its function but also to re-evaluate the function in light of insights gained from its usage.[47]

To illustrate, imagine a community whose central motivation for using the expression 'art' is to represent creative works made to produce aesthetic pleasure. Suppose, however, that the community begins applying the expression also to certain modern creative practices — like performance art or installation pieces. While they initially assume that these practices should be appreciated based on whether they elicit aesthetic pleasure, over time they come to realise that this criterion is unfitting for many of them, as they often aim more at provoking political reflection or emotional discomfort than at aesthetic pleasure.

One response to this realisation would be to narrow their usage of 'art' so that it excludes such practices. Yet such a response might impoverish the community's conceptual repertoire, as it could deprive its members of an expression that carves out a broader category — one that allows them to recognise the diversity of human creative expression and explore links between traditional aesthetic objects and modern creative practices. Therefore, the community might consider an alternative response: to repurpose 'art' so that its function is to track all kinds of creative works and practices. With both responses available, the community's deliberation over how to reconcile the actual usage of 'art' with its function is more likely to yield an optimal outcome than if they took the expression's function as non-negotiable and merely strived to bring its actual usage closer to it. The fact that the Value-based Schema supports such *bi-directional assessments* is its virtue.

[47] Brun (2022) also argues that reflective equilibrium can play a vital role in conceptual ethics. However, his version is aimed at theory development and thus targets a different domain: it involves mutual adjustments between initial pre-theoretic commitments associated with a concept and the theory to which it's imported.

2.5. Objections and Replies

Finally, let's consider two objections to the Value-based Schema. First, I have proposed a systematic method for reverse-engineering an expression's function and illustrated its application using 'water'. Yet, the reader might doubt how realistic it is for philosophers to apply this method to more complex expressions that didn't develop in response to obvious human needs. After all, the method requires philosophers to hypothesise about a user group's conceptual concerns and dispositions based on empirical research about them and their interaction with the expression. Yet philosophers aren't standardly trained to conduct such research. Consequently, the objection goes, the project of conceptual ethics that incorporates functional considerations into the assessment and design of expressions isn't viable.

In reply, while I agree that philosophers often aren't in the best position to conduct the kind of empirical research under consideration, the proper response to this limitation isn't to deem the method unviable. Rather, it should be seen as viable — *but only if* research in conceptual ethics becomes more interdisciplinary. Philosophers who want to do conceptual ethics should recognise the limits of armchair inquiry, step beyond its confines, and collaborate with (or at least consult the work of) experimental philosophers or scholars from other disciplines (e.g., anthropology, history, sociology, and linguistics), who are better trained to conduct the relevant empirical research.

To illustrate, imagine that a conceptual ethicist sets out to assess the translation equivalents of the expressions 'prisoner' and 'incarceration' as they are used in Scandinavian societies. To do so properly, she needs to consider the functions these expressions serve in the given societies. What I would advise such a philosopher to do is consult the work on the phenomenon known in criminology as 'Scandinavian penal exceptionalism', which concerns unusually humane and mild prison conditions and low incarceration rates in Scandinavian countries. There is no shortage of interesting research on this topic that tries to explain the phenomenon as having roots in the egalitarian cultural values and social systems of Scandinavian countries (Pratt, 2008a; Pratt and Eriksson, 2014), that examines its future developments (Pratt, 2008b), that tries to confirm the phenomenon by conducting surveys among prisoners in Norway on their experience of incarceration (Crewe et al., 2023), or that questions the phenomenon (or its extent) by criticising the methodology behind its investigation (Mathiesen, 2012; Ugelvik, 2013) or by providing

countervailing evidence about the treatment of foreign national prisoners and pre-trial prisoners in Scandinavian prisons (Smith, 2012; Ugelvik, 2012).

Even if the philosopher doesn't conduct this research herself, by carefully attending to it, she can still gain valuable insights into how Scandinavian people conceptualise incarceration, what effects this conceptualisation has on the treatment of prisoners, and how their concepts of 'prisoner' and 'incarceration' relate to concepts such as punishment, equality, rehabilitation, or forgiveness. Once she acquires these insights, she still has substantial theoretical work to do in order to reverse-engineer the functions of the translation equivalents of 'prisoner' and 'incarceration' for Scandinavian speakers. Namely, she must assess the relevance of the learnt insights to her inquiry and use them to form hypotheses about the concerns that these terms serve for Scandinavian speakers — or for particular subsets of them — and which of these concerns are most consequential for their disposition to use the terms.

Accordingly, my proposed method neither asks conceptual ethicists to do something beyond their training nor suggests that they fully outsource their work to other disciplines. Rather, it recommends that conceptual ethics become an interdisciplinary and empirically informed project in which philosophers and other researchers divide the empirical and theoretical labour between themselves.[48]

I will now address the second objection. It may be suspected that the Value-based Schema and M-Function seem plausible only if we idealise speakers' primary motivations for using expressions. Yet, in de-idealised reality, speakers are often primarily motivated to use an expression because of prudential effects that seem irrelevant to its function. Consider this example from Riggs (2021, p.11567).[49] The expression 'philosophical zombie' is standard jargon among philosophers of mind. However, the main reason why the expression is used by them might be sociological. For example, it might be because it was introduced and popularised by the famous philosopher David Chalmers (1996), and its usage allows philosophers to express allegiance to him. Or else, it might be because the expression sounds cool and catchy, making it easier to produce

[48] I'm not the first to argue that conceptual engineering can benefit from empirical input. Nado (2021c), for example, contends that empirical research helps clarify the functions concepts serve and what enables them to do so. Others argue that empirical input increases the chances of successful implementation of proposed revisions (Pinder, 2017; Koslow, 2022; Landes, 2025). See also Andow (2020) and Torregrossa (2022) on the role experimental philosophy can play in assessing expressions.

[49] Riggs (2021) uses this example to argue that an expression's function is often messier than it initially seems.

a publishable philosophical paper when using it. Under the Value-based Schema, these prudential effects would count as the function of 'philosophical zombie' since they explain why users value its presence in their jargon. However, this seems unpalatable; we would expect the expression's function to be something more substantive and theoretically relevant, such as providing a modal argument against physicalism or illustrating the conceivability of consciousness-lacking human duplicates.

Here is my reply to this objection. Even if the prudential effects enter the central motivation for which philosophers retain 'philosophical zombie' in their jargon, it seems highly implausible that they alone do so exhaustively. Arguably, philosophers wouldn't use 'philosophical zombie' just because it sounds catchy, or because David Chalmers popularised it, if they didn't expect the expression to also have some substantive theoretical utility. That is, 'philosophical zombie' presumably would never have entered philosophical jargon and wouldn't persist in it if it referred to something irrelevant to philosophers' inquiry, such as chewing gums or traffic cones. Hence, the theoretical utility of what the expression represents is what enables its prudential effects. Therefore, according to the Value-based Schema, the function of 'philosophical zombie' cannot be solely constituted by the given prudential effects but only partially so, alongside its other theoretical effects.

Some readers may find it problematic to say that the function of 'philosophical zombie' includes both prudential and theoretical effects. But at this point, I caution these readers against idealising philosophical terms as immune to fashion and disciplinary trends. If prudential effects *do* indeed play an important role in explaining why philosophers use the expression, this insight should inform our assessment of its value and potential for improvement. Rather than a flaw, it's a strength of M-Function and the Value-based Schema that they foreground such effects as part of the expression's function.

A problematic outcome would arise only if the Value-based Schema led us to an uncharitable interpretation of 'philosophical zombie' as a term that merely pretends to serve theoretical purposes while actually being a mere product of intellectual fashion. However, the schema doesn't do that. After all, it allows not only that philosophers would stop using 'philosophical zombie' if it were no longer useful for their inquiries, but also that the reverse doesn't hold: they would keep using the expression — albeit less frequently — even if Chalmers

and zombies fell out of fashion. Accordingly, nothing in the schema rules out the possibility that, while the function of 'philosophical zombie' encompasses both theoretical and prudential effects, theoretical effects still take precedence.

2.6. Conclusion

In this chapter, I argued that M-Function can address the function-specification problem by presenting the Value-based Schema for specifying the representational and extra-representational effects that centrally motivate a user group to keep an expression in their repertoire. The schema is theoretically useful without its applicability being limited to extremely homogeneous groups, as it yields specifications of conceptual function that are generic enough to accommodate intra-group disagreements over its specific details. With this schema in place, M-Function is now a well-developed notion of conceptual function.

Chapter 3. A Guide to Reliable Extrapolation in Cross-Linguistic Conceptual Ethics

3.1. Introduction

Having developed the motivation-based notion of conceptual function, I will devote the next two chapters to demonstrating how this notion enriches inquiry in conceptual ethics. This chapter demonstrates that conceptual function can guide theorists in pursuing conceptual ethics *cross-linguistically*, i.e., in assessing representational devices across different languages.

Let's first clarify what kind of conceptual assessments I will focus on here. Conceptual assessments can, broadly speaking, be divided into two types of judgements. On the one hand, we can make judgements that recommend that an expression should have some features on the grounds that possessing them will enable it to become a better representational tool for its users than it is currently. Let's call these judgements 'advisory judgements'. Advisory judgements may concern such questions as what the expression should refer to, what should be in its extension, what descriptions should be associated with it, what emotions and valences it should elicit, which objects should be perceived as its paradigmatic instances, what effects it should bring about, what role it should play in its users' lives, or how it should be related to other concepts in its vicinity. These judgements can then further guide what may be called 'diagnostic judgements'. Diagnostic judgements diagnose how good a representational tool an expression is, based on whether it exhibits the features that advisory judgements recommend it to have. Taken together, advisory judgements and diagnostic judgements help philosophers determine what revisions should be made to expressions to improve them. My focus will be on advisory judgements in cross-linguistic conceptual ethics. I will investigate how to determine whether these judgements have cross-linguistic applicability, i.e., whether advisory judgements about an expression in one language can be appropriately applied to similar expressions in other languages.[50]

More specifically, the template situation I explore is this: imagine a conceptual ethicist who argues that an expression X should have certain features to be a valuable representational tool for

[50] In addition to the uni-directional approach to conceptual ethics, which is my current focus, there is also the possibility of the bi-directional approach, involving the assessment of expressions in both communities to create a new concept that fosters cross-cultural dialogue (Vaidya, 2020, p.312).

her own community.[51] Suppose her advisory judgements about X are correct. She then learns of a different community with an expression Y, which is similar enough to X to be its best available translation. This leads her to wonder whether her conclusions about X apply to Y. However, she cannot simply assume that they do, since, despite being the best translations of each other, X and Y might differ in why they matter to their respective communities. Two examples will illustrate this.

Suppose first that X is the English expression 'inference' as used by contemporary analytic philosophers, and Y is the Sanskrit expression 'anumāna' as used by classical Indian philosophers. If one searches for a Sanskrit expression whose meaning best approximates the meaning of 'inference' in English, it's 'anumāna'. This is because the two expressions serve similar roles in the intellectual traditions in which they respectively developed: both groups of philosophers use them to talk about reasoning from conclusions to premises. Yet, they aren't perfect translations of each other. This is because, in classical Indian philosophy, 'anumāna' is defined as one of the five instruments of knowing (*pramanas*), and, as such, it doesn't allow for the conclusions of the reasonings that are in its extension to be false (Siderits, 2020; Vaidya, 2020). By contrast, contemporary Western philosophers use 'inference' as a broader notion defined as any reasoning from premises to conclusions, including one that is unsound due to its conclusion being false.

Alternatively, suppose that Y is the English expression 'humility' as used by Christian ethicists, and X is its best available but imperfect translation into Classical Chinese 'qiānxū' ('謙虛') as used by Confucian ethicists. In Christian ethics, 'humility' is a moral virtue requiring that an agent attributes her salvation, talents, abilities, accomplishments, and virtues to God, feels gratitude towards God, and maintains a lower opinion of herself in relation to God (Austin, 2018, pp.60-65). By contrast, in Confucian ethics, 'qiānxū' is a moral virtue that requires no relationship to God, but instead that an agent moderates her ambitions by realistically assessing her abilities and recognising externally imposed human limitations (Rushing, 2013, pp.176-180). External factors are understood here in terms of fate and historicity, so they don't imply the existence of

[51] I use the term 'community' broadly, making it applicable to both global and local conceptual ethics. It refers not only to entire linguistic communities but also to subgroups sharing similar motivations for using an expression, rooted in common intellectual traditions, histories, cultures, values, religions, lifestyles and more.

any divine entity to which one should permanently feel grateful and inferior. Hence, while the two expressions refer to similar moral virtues in their respective ethical frameworks, they aren't semantically equivalent.

In both examples, the differences in how X and Y are respectively defined in the two communities of their users signal some important differences in what roles these expressions play in the communities' cultures, systems of thought and overall life patterns of their members. This exposes a problem for conceptual ethics. On the one hand, the conceptual ethicist clearly shouldn't assume that all the advisory judgements she makes about X also hold for Y. Such an assumption would be patronising and overgeneralising, as it disregards the possibility that different features might contribute to the value of each expression for its respective users.

At the same time, it would also be unsatisfactory if, upon recognising this issue, the conceptual ethicist chose preventive silence — refraining from assessing whether a judgement about a familiar expression in her own language could extend to similar expressions in other languages. For one thing, philosophical inquiry is often understood, *pace* Sellars (1963/2007, p.369), as seeking to answer questions about 'how things hang together in the broadest possible sense of the term'. Such questions aren't language-relative, and if we give up on the possibility of cross-linguistic conceptual assessment entirely, one might reasonably doubt how insightful conceptual ethics and conceptual engineering are as philosophical methods. Their scope would then be confined to individual languages, severely limiting their insight. Moreover, considering that conceptual engineering — like most analytic philosophy — is predominantly published in English, preventive silence about the cross-linguistic applicability of one's conceptual assessments risks fostering active ignorance toward non-English conceptual systems, i.e. the ignorance of the socially dominant that is actively cultivated rather than a mere passive occurrence.[52] Even if motivated by caution against overgeneralisation, this approach is problematic: it unjustifiably biases conceptual ethics towards English, excluding valuable normative insights about representational devices of other linguistic traditions.

Hence, while cross-linguistic conceptual assessments carry the risk of overgeneralisation, conceptual ethics cannot avoid them without incurring the opposite risk of English-centred

[52] For discussion of active ignorance in social epistemology, see e.g. Mills (2007), Pohlhaus (2012), Medina (2013, ch.1), and Tilton (2024).

parochialism. To my knowledge, this problem has been largely underexplored in the current conceptual ethics literature. This isn't to say that the problem — or its echoes — have flown entirely under the radar in conceptual ethics and adjacent fields. The tendency to overgeneralise conceptual assessments across cultures can lead to the imposition of foreign conceptual resources on other communities, often at the expense of their own. Such conceptual imposition has been famously criticised by Kwasi Wiredu (1995, 1998) as part of his argument for the conceptual decolonisation of African philosophy, which he advocates through avoiding the uncritical acceptance of concepts imported from outside into African conceptual systems. Relatedly, the risk that a philosophical study of concepts may be biased towards Western languages and cultures has been discussed in connection with descriptive conceptual analysis driven by the method of cases by Machery et al. (2004), Mallon et al. (2009), and Sytsma et al. (2015). These authors present data suggesting cross-cultural variation in intuitions about the reference of proper names among speakers of Western and East Asian languages. [53]

Furthermore, early gestures towards the given problem can also be found in Haslanger (2000), whose paper on revising the definitions of gender and race terms is often regarded as one of the foundational texts in the conceptual engineering literature.[54] After proposing her revisionary definition of gender terms — where these terms refer to positions in a social hierarchy occupied based on observed or imagined sex-related features — Haslanger notes:

> 'It is a virtue, I believe, of these accounts, that depending on context, one's sex may have a very different meaning and it may position one in very different kinds of hierarchies. The variation will clearly occur from culture to culture (and sub-culture to sub-culture); e.g., to be a Chinese woman of the 1790's a Brazilian woman of the 1890s, or an American woman of the 1990's may involve very different social relations, and very different kinds of oppression. Yet on the analysis suggested, these groups count as women insofar as their subordinate positions are marked and justified by reference to (female) sex.' (Haslanger, 2000, p.39)

[53] Note that the results of Machery et al.'s original study (2004) have been extensively questioned by theorists who identified possible confounds within it (see, e.g., Ludwig, 2007; Devitt, 2011; Lam, 2010; Izumi et al., 2018).
[54] Haslanger (2005, 2006) later reframed what she originally presented as revisionary definitions of gender (and race) terms, as the best externalist interpretation of what these terms in fact mean.

This passage suggests Haslanger recognises that her claims about how gender terms should be understood to combat sex-based oppression are based primarily on observations of such oppression in Western societies. That is, she acknowledges that these claims cannot be automatically extended to gender-terms in non-Western societies, where oppression may operate differently. She can even be read here as recognising the need for separate ameliorative analyses of gender terms in these societies and inviting others to pursue them. Yet this invitation went largely unheeded. Despite Haslanger's paper having sparked an extensive debate on gender terms, one would search in vain for any attempt within this debate to conduct a revisionary assessment of gender terms in non-Western societies.[55] This gap may stem from the debate's participants being predominantly U.S. and European scholars who didn't feel competent to conduct such assessment — but if so, it underscores my concern that *preventive silence* perpetuates ignorance of non-Western conceptual resources.

What, then, should a conceptual ethicist do if they want to avoid ignorance about the cross-linguistic relevance of their conceptual assessments, while also avoiding the mistake of overgeneralising them? A plausible approach is to remain open to extending their conceptual assessments to other communities, but to do so only through a reliable belief-forming method. By following such a method, the conceptual ethicist should be able to develop a good track record in properly applying advisory judgements made about an expression X in their own community, to an expression Y used in a less familiar community.[56]

Here is a first-pass proposal for what such a method might look like: the conceptual ethicist should assess Y separately, just as she previously assessed X, and check whether the advisory judgements made about Y match those made about X. This may seem like a promising and reliable method — if properly applied. Yet, this method is highly epistemically demanding. After all, it requires the conceptual ethicist to double her work by subjecting Y to an assessment parallel to that of X. This is a substantive evaluative task, as it requires that she identifies all relevant considerations bearing on the value of Y for the unfamiliar community, weighs them in relative

[55] See, e.g., Mikkola (2009), Saul (2006), Jones (2014), Jenkins (2016), Simion (2018b), McKenna (2018), and Marques (forthcoming).

[56] To be clear, some advisory judgements may hold for most expressions simply because they track general normative facts — such as that an expression shouldn't be oppressive or should have a minimally coherent definition. These, however, aren't my focus. Instead, I examine the cross-linguistic application of judgements that are responsive to those features of an expression that make it a distinctive tool for a user group.

importance, and forms advisory judgements based on that weighing. To do this well, she must have a detailed understanding of Y's role in the community — something she may often lack and be unable to acquire due to practical constraints. Moreover, outsourcing the task is difficult, as it would require someone who both understands Y's role and is comparably trained in conceptual assessment. Such an individual may often be unavailable. Therefore, despite its initial appeal, it's highly doubtful that conceptual ethicists can realistically be expected to apply this method reliably in practice.

Accordingly, it's worth asking whether the conceptual ethicist could reliably extend some of her advisory judgements about X to Y using a less epistemically demanding method — one that doesn't require separately assessing Y. In this chapter, I explore whether extrapolation can sometimes serve this purpose. Broadly, extrapolation involves inferring that conclusions drawn about a well-known domain also apply to a less familiar one, based on a relevant similarity relation between them. In conceptual ethics, this means inferring that an advisory judgement about a familiar expression also holds for a less familiar one, which hasn't been separately assessed, due to a relevant similarity relation between the two.

The key question, then, is whether there are conditions under which the extrapolation of advisory judgements can serve as a reliable belief-forming method in conceptual ethics, i.e., one that typically yields a high ratio of true to false conclusions. What makes a belief-forming process reliable under normal conditions is that it produces conclusions based on strong supporting evidence rather than arbitrarily.[57] Accordingly, for extrapolation to be a reliable belief-forming method in cross-linguistic conceptual ethics, there must be strong evidence that the original and target expressions share relevant similarities decisive for the extrapolated judgement's truth. As we will see, identifying what these similarities are isn't straightforward. What's clear, however, is

[57] Two clarifications are in order. First, by 'strong evidence', I mean evidence whose acquisition makes it highly likely that the supported conclusion is true. What counts as strong evidence depends on the agent's total evidence; initially strong evidence may be weakened by defeating evidence that reduces its confirmatory power. Second, by 'reliable under normal conditions', I mean non-deceptive conditions in which agents with *ordinary* human cognitive faculties operate. Epistemologists often discuss cases — like Gettier-style scenarios — where strong evidence leads to unreliable belief formation due to deception (e.g., Cohen, 1984; Pollock, 1984; Goldman, 1976), or cases involving clairvoyance where beliefs are reliably formed without strong evidence (e.g., BonJour, 1980; Plantinga, 1993). While useful for debates on the relationship between reliability, justification, and knowledge, such cases aren't directly relevant here, since conceptual ethicists presumably lack clairvoyant abilities and aren't typically in Gettier-style situations.

that these similarities must be identifiable *without* separately assessing the target expression — otherwise, the judgement wouldn't qualify as extrapolated.

In what follows, I examine various candidates for these similarities, aiming to identify a *criterion for reliable extrapolation in cross-linguistic conceptual ethics*. My investigation will lead me to a *function-based criterion* that conceptual ethicists can use to determine (1) which advisory judgements, if any, can be reliably extrapolated to other communities, and (2) which expressions these judgements apply to within them.

The chapter is structured as follows: In §3.2, I consider proposals to frame the reliable extrapolation criterion in terms of semantic similarities between expressions but argue that they fail, since what we ultimately value in expressions are their effects, not their semantic features. In §3.3, I build on this point by exploring an effect-based criterion focused on overlaps in expressions' functions. I then argue that this criterion needs an additional condition requiring that the advisory judgement be true for the original expression independently of idiosyncratic factors, or due to idiosyncratic factors similar to those affecting the target expression. In §3.4, I address two objections to my proposal. Finally, in §3.5, I briefly outline how conceptual ethicists can make use of the criterion.

3.2. The Semantic Approach to Reliable Extrapolation

Expressions are assessed in conceptual ethics *qua* representational tools, and many representational features of expressions are conferred on them through their semantic features. This makes it natural to think that reliable extrapolation in conceptual ethics requires strong evidence that the originally assessed expression and the expression targeted by extrapolation have something in common with respect to their semantics. If so, the question is what semantic similarity relation could be relevant here.

The relevant relation arguably cannot be a semantic equivalence relation, as that would be too stringent a requirement. Consider, for example, the translation pair 'humility'-'qiānxū' discussed above. We have seen that there are semantic differences between the two expressions when they are respectively used by Christian ethicists and Confucian ethicists. This might indicate that some advisory judgements that are correct about one of them are incorrect about the other. Yet, considering they share certain semantic similarities, it's premature to say no advisory judgements

can be reliably extrapolated between them. For instance, suppose we judge that 'humility' in Christian ethics refers to a moral virtue primarily reflected in inner attitudes, and we learn that this judgement rests on characteristics shared with 'qiānxū'. If this evidence was acquired without separately assessing 'qiānxū', we could use it to reliably extrapolate this judgement to 'qiānxū'. Setting aside whether this example is true, it would be surprising if such extrapolations were never possible simply because two expressions aren't semantically equivalent.

Nevertheless, it might still be argued that what facilitates reliable extrapolation is some semantic similarity relation that is weaker than semantic equivalence but aligns closely with what some authors in the conceptual engineering literature discuss under the notion of *topic continuity*.[58] Topic continuity is often discussed as a criterion for distinguishing between conceptual revisions and conceptual replacements, and for delineating the limits of permissible conceptual engineering as opposed to just 'changing the subject' that an expression concerns before being engineered. However, one might also try to appeal to topic continuity in order to formulate a reliable extrapolation criterion such as RE1.

> *RE1*: An assessor A can, under normal conditions, reliably extrapolate a correct advisory judgement J made about what features an expression X should have when used by one community to an expression Y used by a different community iff A has strong evidence E that X's and Y's semantic contents are similar enough for them to concern the same topic.

As explained earlier, the reliable extrapolation criterion presupposes that E is a type of evidence the assessor can obtain without conducting a separate assessment of Y or deferring to someone who has. Otherwise, the agent couldn't be said to extrapolate based on E, which would undermine the very purpose of the criterion: to guide conceptual ethicists in determining when they can extend advisory judgements about one expression to different expressions in other communities' repertoires without separately assessing the expression. Accordingly, all candidates for the reliable

[58] The discussion of topic continuity in conceptual revision can be traced back to Strawson's argument (1963, pp.505-506) against Carnap's account of conceptual explication (1950, pp.1-18). More recent discussions appear in Cappelen (2018), Nado (2021a), Sawyer (2018), Koch (2023), Belleri (2025), McPherson and Plunkett (2024), and Kocurek (2022), among others. Note, though, that topic continuity is sometimes framed as functional rather than semantic similarity (e.g., Sundell, 2020; Nado, 2021a, pp.1515-1522; Thomasson, 2020, p.443). I avoid this framing, as it can mislead by implying an expression's function depends solely on its topic. However, as argued in Chapter 2, an expression's function often involves extra-representational effects that don't fit the mould of topic continuity.

extrapolation criterion considered here are formulated against the backdrop of this presupposition.[59]

RE1 commits us to saying more about what exactly it means for two expressions to concern the same topic. One proposal for explicating reliable extrapolation in terms of topic continuity builds on a heuristic method that Cappelen (2018, pp.109-111) proposes for testing whether a single expression preserves topic across variations in its intension and extension. The method runs as follows: when an expression is used in two utterances of the same sentence to convey distinct contents, those contents concern the same topic if it's permissible to report the speakers as having said the same thing. For example, suppose speakers A and B both say 'Leibniz is well-dressed', but A means 'well-dressed for a philosopher' and B means 'well-dressed for a 17th-century male adult'. The method holds that 'well-dressed' concerns the same topic in both utterances if we can truthfully report, 'Both A and B said the same thing'. We can call this method the 'samesaying test'.

Although RE1 appeals to topic continuity between different expressions, the samesaying test can be easily adapted to it: When two expressions are used in utterances of two sentences, the only difference between which can be attributed to the expressions' semantic contents, those contents concern the same topic if it's permissible to report the speakers as having said the same thing. For example, suppose an English-speaking Christian ethicist says 'Peter is humble' and a Classical Chinese-speaking Confucian ethicist utters its best translation, 'Bǐdé hěn qiānxū', both referring to the same person. If we can report that both said the same thing, this indicates that 'humble' and 'qiānxū' concern the same topic despite minor meaning differences. This yields the reliable extrapolation criterion RE2.

> *RE2*: An assessor A can, under normal conditions, reliably extrapolate a correct advisory judgement J made about what features an expression X should have when used by one community to an expression Y used by a different community iff A has strong evidence E that X and Y pass the samesaying test.

Now, as Cappelen (2020, p.526) himself recognises, the dispositions to report speakers as samesaying aren't invariant among their interpreters. Instead, they vary relative to interpreters'

[59] More needs to be said about how an assessor can obtain E. However, since I believe RE1 requires significant revision, I set this question aside until we have a more satisfactory version of the criterion.

interests. If interpreters are interested only in what 'qiānxū' and 'humble' roughly mean in the given utterances, they will be disposed to judge that the Christian ethicist and the Confucian ethicist in the above example said the same thing. But the interpreters might also be interested in more fine-grained aspects of what the expressions mean, for example, because they want to draw inferences about whether Peter having the property that 'humble' picks out in Christian ethics or that 'qiānxū' picks out in Confucian ethics is compatible with him not considering himself to be inferior to a divine entity. In that case, they aren't disposed to report that the two speakers said the same thing. As a result, topic continuity is a shifty feature that varies depending on the interests of the expressions' interpreters.

I leave open whether the shiftiness of topic continuity is independently problematic.[60] However, it clearly throws RE2 into doubt, since the reliability of extrapolation doesn't seem to shift in the way RE2 suggests. Reliability concerns whether a belief-forming process tends to produce true beliefs. So RE2 could only be true if the reliability of extrapolation — and thus the truth value of the advisory judgements it yields — were relative to the interests of those interpreting the expressions. This leads to an unacceptable consequence for two reasons.

Firstly, even if someone's interests are involved in determining what features a target expression should have, these are arguably the interests of those directly affected by it — its active users. But according to RE2, whether two expressions pass the samesaying test, and thus whether advisory judgements about the original expression can be reliably extrapolated to the target expression, depends on the interests of *anyone interpreting the target expression* — even those who never use it or don't speak the language to which it belongs. This seems overly permissive.

Secondly, even if we fixed this problem by tweaking RE2 to condition the reliability of extrapolation on whether expressions pass the samesaying test relative to a target group of its active users, RE2 would still be unacceptably permissive. These users may have diverse interests across different situations, not all of which are relevant here. For instance, suppose the Christian ethicist says 'Peter is virtuous' instead of 'Peter is humble'. There might be some interpreter, another Confucian ethicist using 'qiānxū', who is only interested in whether Peter is in the extension of

[60] See, e.g., Belleri (2025), who argues that the shiftiness of topic continuity is problematic because it brings debates over the permissibility of concept engineering to a standstill.

‘qiānxū’, because if he is, this entails that he has a moral virtue. This interest might lead her to judge that ‘Peter is virtuous’ says the same thing in the mouth of the Christian ethicist as ‘Bǐdé hěn qiānxū’ does in that of the Confucian ethicist. Yet this gives us no reason to think that correct advisory judgements about ‘virtuous’ in Christian ethics can be reliably extrapolated to ‘qiānxū’ in Confucian ethics. They clearly cannot, as the former denotes a much broader category than the latter. This is another example showing that it’s much easier for two expressions to pass the samesaying test than to be suitable for reliable extrapolation. Hence, basing the success condition of the latter on that of the former is misguided.

An alternative approach to understanding topic continuity is to treat it as something that is grounded in how much overlap there is between expressions’ extensions across different worlds and times. This paves the way for the reliable extrapolation criterion, RE3.

> *RE3*: An assessor A can, under normal conditions, reliably extrapolate a correct advisory judgement J made about what features an expression X should have when used by one community to an expression Y used by a different community iff A has strong evidence E that there is a stable overlap between X’s and Y’s extensions.

For two expressions to have a stable extensional overlap is for them to systematically coincide in what they are correctly applied to across different worlds and times. RE3 is, however, too demanding a criterion. This is because it seems possible to reliably extrapolate some advisory judgements across expressions even if there cannot be any extensional overlap between them. To illustrate two ways in which this may be so, consider the following two conceivable cases:

> *Proposition*: Imagine multiple groups of philosophers of language working on propositions. One group defines ‘proposition’ as denoting sets of possible worlds. The second group defines ‘proposition’ as denoting real-world structured entities composed of objects, properties, and relations. The third group defines ‘proposition’ as denoting abstract structured entities composed of modes of presentation. The fourth group defines ‘proposition’ as denoting acts of predicating a property of an object. For each group, the term ‘proposition’ refers to a metaphysically different type of entity, resulting in no overlap between their extensions. Nevertheless, in all groups, the central point of using the ‘proposition’ is the same: it enables philosophers to discuss an entity that serves as the

content of sentences and the bearer of truth values.[61]

Conscript: Imagine multiple countries with different languages, each using a local term — let's call it a 'conscript-term' — that translates to 'conscript' in English, as it's correctly applied to individuals who, under domestic law, are required to perform military service. The central point for which each of these countries uses their conscript-term is to assign a legal duty of military service to specific individuals, typically based on age, sex, and citizenship, in order to support national defence. However, each country defines its conscript-term with a local constraint that enables it to apply only to its own citizens and defines citizenship in a way that excludes dual citizenship. As a result, no individual can meet the definition for more than one conscript-term, making their extensions mutually exclusive.

The setup in both cases is such that, despite there being no extensional overlap between the given counterpart terms, it still seems plausible that at least some advisory judgements can be reliably extrapolated across them. This is because, in each case, although the terms diverge in their semantic values, the central effects associated with their usage are the same. In the first case, although 'proposition' refers to different entities depending on which group of philosophers uses it, all these entities serve the same generic role in their respective groups: to enables their members to communicate about the entity serving as the content of sentences and the bearer of truth value. Presumably, there is a correct, invariant answer as to which of the potential referents of 'proposition' best fulfils this role. If one group correctly identifies it, that judgement seems reliably extrapolable to 'proposition', as it's used by other groups.

In the second case, the usage of conscript-terms serves the common end for each country, which is to construct an institutional kind whose members contribute to the defence and security of the state through compulsory military service. While how conscripts must operate for national defence to be effective might vary to some extent depending on the country's military structure, strategic priorities, and geopolitical context, it would be surprising if there were no reliably extrapolable judgements about the norms conscript-terms should promote. Effective conscription plausibly always requires lawful obedience, adequate training, discipline, restraint, and respect for

[61] A similar example can be found in Brun's discussion of the extensional overlap interpretation of topic continuity (2016, p.1221).

international humanitarian standards.

Two key lessons can be drawn from these cases. First, the effects associated with the usage of expressions can be type-identical even if there is no extensional overlap between them. Second, these effects can be at least as important for the reliable extrapolation of advisory judgements about them as their semantic similarity. In fact, I believe that the second point can be reinforced: strong evidence about semantic similarities between two expressions makes extrapolation more reliable only to the extent that it also provides strong evidence that the usages of these expressions are associated with some relevantly similar effects. That is, if we have strong evidence that two expressions have identical semantic values but know their users associate them with very different effects, advisory judgements about one expression cannot be reliably extrapolated to the other expression.

To illustrate, consider the following thought experiment. Most human languages have a term referring to H_2O — call these 'water-terms'. It's undeniable that human linguistic communities can often reliably extrapolate correct advisory judgements from one water-term to another. For instance, if one community judges that their water-term will be a more useful expression for them if they perceive as its prototypical instances those samples of H_2O that they can safely drink, this likely holds for others too. On its face, reliability might be thought to be a consequence of human communities having strong shared evidence that their water-terms co-refer to H_2O.

Suppose, however, that humans discover alien communities whose water-terms also refer to H_2O but who use it for entirely different purposes, unrelated to human needs such as drinking, sanitation, or agriculture. For these aliens, H_2O serves practical roles unknown to humans, and their water-terms reflect that. While humans apply water-terms to H_2O in order to verbally communicate about the substance that plays the above practical roles for them, aliens apply their water-terms to H_2O to communicate about the substance that, for them, plays a completely different practical role.

In such a case, humans cannot reliably extrapolate advisory judgements about their own water-terms to the alien water-terms, since those judgements would likely overlook the terms' practical relevance in the alien context. A better approach is to assess those terms separately. This suggests that what ultimately matters for reliable extrapolation isn't evidence of co-reference, but

evidence of similarity in the effects for which terms are used. Evidence about co-reference matters only insofar as it indicates such similarity.

It's no surprise that similarity in expressions' associated effects weighs more heavily in reliable extrapolation than similarity in their semantic content. After all, even if expressions serve referential functions, reference is rarely the final end for which they are used. Instead, reference is merely a means instrumental to producing further extra-representational effects of the kind discussed in Chapter 2. Naturally, these effects are of great practical importance to users. Yet RE3 recommends that we ignore these effects and focus solely on semantic similarities when determining which judgements are reliably extrapolable across expressions. Any extrapolation criterion focused only on semantic content, like RE2, offers the same recommendation. That's a flawed approach, however. It leads to conceptual assessments that are as incomplete as evaluating a house by its insulation, furnishing, or stability alone, while ignoring their contributions to its final purpose: habitability. The remaining question for the next section is how to formulate a reliable extrapolation criterion that takes into account relevant similarities between expressions' associated effects.

3.3. The Effect-based Approach to Reliable Extrapolation

3.3.1. Which Effects Bear on Reliable Extrapolation?

So far, I have described expressions as associated with various effects that, while related to their semantic features, go beyond them. There are several ways to interpret what it means for an effect to be associated with an expression in the sense relevant to reliable extrapolation. The most straightforward is that such effects are those an expression's usage tends to produce. This interpretation guides us to RE4.

> *RE4*: An assessor A can, under normal conditions, reliably extrapolate a correct advisory judgement J made about what features an expression X should have when used by one community to an expression Y used by a different community iff A has strong evidence E that there is a relevant overlap between some effects that the usages of X and Y tend to produce.

The effects an expression's usage tends to produce are certainly important for conceptual ethicists to consider, as they offer clues about the expression's impact on users' lives and why it

matters to them — insights that can inform decisions about its amelioration. Nonetheless, I don't think that E4 is a satisfactory reliable extrapolation criterion. After all, as discussed in Chapter 2, many expressions are used in ways that tend to produce multifarious effects.

To illustrate, consider the wide range of effects the term 'money' can have. Most obviously, it's used to discuss practical aspects of economic life — transactions, income, savings. For some, it symbolises success and opportunity, motivating professional performance; for others, it evokes stress and anxiety due to associations with debt, financial pressure, or inequality. The term also shapes the social reality of finance: banks, monetary policies, interest rates, and stock markets couldn't exist in a society that lacks the term 'money' or its semantic corollary in their language. Or else, the term enables assessments of creditworthiness and eligibility for loans, rent, benefits, or scholarships. The list could go on.

RE4 would be too demanding if we interpreted 'relevant overlap' as requiring that all or most effects produced by an expression's usage be the same. If that were the case, correct advisory judgements about 'money' would rarely be reliably extrapolable to other expressions. It's rare to find an expression used by a different community that not only translates as 'money' but also tends to produce all or most of the countless effects 'money' tends to produce when used by English speakers. We can easily imagine a society whose financial institutions are structured such that their closest equivalent to 'money' lacks some of these effects. It seems too restrictive to deny the reliable extrapolation of our assessments of 'money' to these societies as impossible on that basis.

A better reading of 'relevant overlap' in RE4 is that advisory judgements are reliably extrapolable only if *some* of the effects of each expression's usage are the same. Still, something needs to be added to this interpretation, since not just any overlap in effects supports reliable extrapolation. For example, the contemporary usage of 'critical thinking' arguably tends to produce some negative effects: individuals who distrust mainstream views and gravitate towards conspiracy theories sometimes justify their views by claiming that they are simply engaging in 'critical thinking'. More broadly, 'critical thinking' is often used arbitrarily as a catch-all phrase for preferred reasoning styles. But imagine there is a community in whose language the same negative effects tend to be produced by the usage of an expression Y that refers to what 'understanding' refers to in English. Despite this overlap, advisory judgements about 'critical thinking' don't seem reliably extrapolable to Y. After all, the two expressions still differ

significantly in other effects, which poses a risk that advisory judgements about 'critical thinking' are unsuitable for Y. Thus, we need a criterion to identify which overlapping effects are relevant to reliable extrapolation.

What I want to propose is that the relevant overlap must be not just between any effects an expression tends to produce, but between the effects that are explanatorily tied to the fact that the expression is used by a particular group. Specifically, these are the effects that define an expression's *function* for a group, where 'function' is understood according to the motivation-based interpretation defended in Chapters 1 and 2. This leads us to the reliable extrapolation criterion RE5.

> *RE5*: An assessor A can, under normal conditions, reliably extrapolate a correct advisory judgement J made about what features an expression X should have when used by one community to an expression Y used by a different community only iff A has strong evidence E that there is an overlap between X's and Y's functions for their respective communities, i.e., the effects central to their motivations for keeping X and Y in their conceptual repertoires.

To explain the rationale behind RE5, consider any pre-existing expression in a conceptual repertoire. Assessing whether such an expression is a valuable representational tool can be justified only if it's guided by certain beliefs about what centrally motivates its presence in speakers' repertoire, i.e., its function. This is clearest in advisory judgements about what features an expression should have to fulfil its function. Without beliefs about that function, one wouldn't even be in a position to form such judgements, as one would lack even an approximate idea of which alternative designs to compare in terms of how they contribute to its proper functioning.[62]

Furthermore, even if less obviously, such beliefs must also underpin other kinds of justified conceptual assessment. For example, it's difficult to see how one could justify their judgement about what function a pre-existing expression *should* have without relying on some beliefs about its current function. After all, such a judgement would require either a conservative justification,

[62] A congenial point is made by Richardson (2024), who argues that although the question of which concepts we should use is partly independent of our subjective values, those values can still delineate which conceptual alternatives are relevant to that question. In a similar vein, we can say that understanding an expression's function — that is, the effects for which a group of its users finds it valuable — helps us determine which of its alternative designs we should consider as relevant when assessing what enables its proper functioning.

which would involve arguing that the expression already serves the recommended function and explaining why it should retain it, or a *comparative justification*, which would involve arguing that the expression's recommended function is a preferable alternative to its current one.

What about judgements concerning what features an expression should have *regardless* of its function? A brief reflection on these judgements reveals that even they require the assessor to attend to the expression's function. To see this, let's consider what might motivate these judgements. One possibility is that the assessor believes the recommended features would enable the expression to fulfil a function it should have. If so, they must be able to justify why their recommended function is a preferable alternative to its current one, which — as shown above — requires them to have beliefs about the latter. Alternatively, assessors might think an expression should have certain features for reasons unrelated to function. But even then, it seems rationally incumbent on the assessor to check whether those features are compatible with the expression's actual or recommended function. It would be premature to recommend that an expression possess features without first ensuring that they aren't in conflict with the central motivation behind its use.

Thus, beliefs about an expression's function play a crucial role in justifying advisory judgements about pre-existing expressions. This shouldn't be surprising: an expression's function captures the central point motivating its use, so any attempt to assess what features it should have to be a valuable tool for its users without understanding this point seems to be mere guesswork. Since extrapolated conceptual assessments also concern pre-existing expressions, they are no exception here. Extrapolating an advisory judgement from one expression to another without first confirming overlap in their functions is unlikely to yield correct assessments. It risks error by missing some key differences in the motivations for which these expressions are included in their respective users' conceptual repertoires. RE5 reflects this by requiring strong evidence of functional similarity for reliable extrapolation.

RE5 allows us to explain why advisory judgements about 'critical thinking' and Y aren't reliably extrapolable in the example above. For Y, the negative effects likely don't contribute to its function. As a translation of 'understanding', Y refers to a gradable epistemic state consisting of

beliefs that represent explanatory relations between pieces of information relevant to the topic.[63] In contrast, it's less clear what effects centrally motivate the use of 'critical thinking'. A pessimistic diagnosis sees its use as driven primarily by the given negative effects; an optimistic diagnosis holds that it is used to produce some other effects related to the kind of thinking it refers to. Either way, Y and 'critical thinking' serve very different functions, making extrapolation of advisory judgements between them unreliable.

Additionally, RE5 no longer requires that X and Y be actually capable of fulfilling their functions. This is appropriate; otherwise, we would have to reject reliable extrapolation in cases involving expressions with a *phantom function*, i.e., a purported but unrealisable function discussed in Chapters 1 and 2. This would be an ad hoc restriction. For example, suppose speakers use 'global justice' to refer to an ideal socioeconomic state that is, unbeknownst to them, impossible to realise. Even so, the term may still guide them towards approximating that ideal. In such cases, we can still make valid judgements about how 'global justice' should be interpreted to fulfil its function partially. It seems unmotivated to claim that extrapolating such judgements is unreliable by default merely because the function is phantom.

3.3.2. Further Refinements to the Reliable Extrapolation Criterion

We aren't yet done with the reliable extrapolation criterion because contrary to RE5, not all correct advisory judgements can be reliably extrapolated across functionally similar expressions. This is because many correct judgements depend on idiosyncratic factors specific to a user group and its environment, which may not apply to another functionally similar expression. To see how this can happen, recall that earlier I divided the advisory judgements one can make about an expression into three types, depending on how they relate to its function.

First, there are judgements about what features an expression should have to fulfil its function for a particular user group — to produce the effects that primarily motivate the group to use it. Let's call these judgements 'judgements about proper functioning'. Second, one can take a step back and assess what function the expression should have for a user group. These judgements question whether an expression should be used to produce the effects central to the group's present motivation for using it, or to produce some alternative effects. As such, they address the

[63] This characterisation of understanding is widely accepted among epistemologists (e.g., Salmon, 1997, pp.79-92; Kvanvig, 2003, p.192; Grimm, 2010; Khalifa, 2013; and Hills, 2016, pp.661-668).

fundamental question of why the group should use the expression in the first place. We can, therefore, call them 'foundational judgements'. Third, one can judge whether an expression should have some additional features for reasons independent of whether possessing them contributes to the fulfilment of its function relative to a user group. We can call these judgements 'supplementary judgements'.

It would be naïve to think that any of these three types of advisory judgements guarantees reliable extrapolation between functionally similar expressions. This is because none of them is immune to cases where a judgement is correct for one expression due to idiosyncratic factors that don't apply to another functionally similar expression. Let me illustrate how such cases can arise for each of the three judgements, starting with judgements about proper functioning.

Sometimes, different features enable expressions to fulfil the same function. For example, in a community where the term 'afterlife' serves the function of mitigating death anxiety, what associations with the term enable it to fulfil this function likely depends on the community's religious beliefs. Catholics might find associating the concept of Purgatory with 'afterlife' effective in overcoming fear of death, while this association would be less effective for Protestants — who emphasise salvation through faith — and irrelevant for Buddhists or Hindus, whose conceptions of afterlife centre on reincarnation. Thus, even assuming mitigating death anxiety is the shared function that 'afterlife' serves for Catholic, Protestant, Buddhist, and Hindu communities, the judgements about the features it should have to fulfil this function for Catholics cannot be reliably extrapolated to other communities.

Next, consider how idiosyncratic factors can feed into foundational judgements. Imagine a community C that is primarily motivated to use the term 'piety' to track a moral virtue involving reverence and devotion and positively evaluate it. Hence, C's members use 'piety' as a thick concept whose function is both descriptive and evaluative. Suppose we assess that 'piety' shouldn't be an evaluative expression in C because the evaluative aspect of its function fosters a toxic and oppressive atmosphere in which C's members are assessed as morally worse or better people based on how 'pious' they are to the community's dominant religion. Accordingly, we conclude that 'piety' should be repurposed in C as a purely descriptive term that neutrally refers to any behaviour involving reverence and devotion. Even if this is the right judgement for C, it would be unreliable to extrapolate it to other communities using a functionally similar term. Those communities may

have a religious environment healthy enough to support evaluative uses of 'piety'-like terms without causing harm. After all, the judgement rests on considerations that are idiosyncratic to C and unlikely to generalise. Many other communities may have religious environments healthy enough to promote a moral virtue of reverence and devotion without fostering oppression.

Even some supplementary judgements about what additional features 'piety' should have when used by C may be too tied to the particularities of C to be reliably extrapolable to functionally similar expressions used by other communities. Suppose, for instance, we offer a less radical proposal to attenuate the toxic effects of 'piety' in C. Specifically, we judge that, instead of revising the function of 'piety' so that it's no longer a thick concept, it may suffice to shift its paradigmatic associations from religious to non-religious behaviours, such as filial or civic piety. Suppose this judgement is correct, as such associations make 'piety' a more profane term that is less susceptible to religious exploitation. Still, if the tendency to produce harmful effects isn't shared by functionally similar terms in other communities, extrapolating this judgement to them wouldn't be truth-conducive.

What bars the judgements in the above three examples from being reliably extrapolable is that there is a non-negligible possibility that what makes them true for the original expression are some idiosyncratic factors that aren't generalisable to other functionally similar expressions. Therefore, reliable extrapolation requires that the assessor firmly disconfirms this possibility as negligible. To incorporate this observation, I propose to revise RE5 into RE6.

> *RE6*: An assessor A can, under normal conditions, reliably extrapolate a correct advisory judgement J made about what features an expression X should have when used by one community to an expression Y used by a different community iff
>
> 1. A has strong evidence E_1 that there is an overlap O between X's and Y's functions, i.e., the effects central to their motivations for keeping X and Y in their conceptual repertoires.
>
> 2. A has strong evidence, E_2, confirming either (i) that the features J recommends for X are desirable for X independently of any idiosyncratic factors apart from a part of its function within O, or (ii) that the features are desirable due to such idiosyncratic factors, but similar idiosyncratic factors are also at play for Y.

While the condition 2 in RE6 is long-winded, its core idea follows directly from the setup of the reliable extrapolation criterion: E_2 captures the only type of evidence that can firmly disconfirm the possibility that the factors making J correct for X don't extend to Y, while still being obtainable without the assessor separately assessing Y or deferring to someone else who does. I will now demonstrate what this means for each of the three types of judgements under consideration.

Let's start with judgements about proper functioning. Suppose a conceptual ethicist judges that an expression X should have certain features to fulfil its function. She then considers whether this judgement would still hold for a functionally similar expression Y. The key consideration here is whether she has good reason to believe that the component of X's function enabled by these features isn't multiply realisable. If so, the features are likely desirable for other expressions sharing that functional component, and, as per the sub-condition (i) in 2, her judgement about X can be reliably extrapolated to Y.

But how can a conceptual ethicist gain good reason to believe that a function (or part of it) isn't multiply realisable? This can often be done by reflecting on whether its realisation is constrained by one of two kinds of facts. First, it may be constrained by invariant descriptive facts about humans and their environment. Consider the term 'knowledge'. Suppose the best epistemic theorising supports Williamson's (2002) view that knowledge is an explanatorily fundamental epistemic attitude all humans and cognitively similar animals can have, and that part of the function of 'knowledge' is to track this attitude. Then, if we correctly determine what kind of attitude 'knowledge' should refer to in order to serve this function, we can reliably extrapolate it to any expression with the same tracking function. On Williamson's account, there is a single epistemic attitude that is fundamental for all humans and whose nature is grounded in invariant facts about the human cognitive system and the environment in which it evolved.

Second, functional realisation may be constrained by universal normative facts. Take the expression 'war criminal'. One of the central purposes for which 'war criminal' finds its place in our language is presumably that it enables us to hold individuals who have committed morally impermissible acts during warfare legally accountable for them. Our best moral theorising suggests there are some universal moral facts about which acts these are. Based on these facts, we can then single out a unique kind 'war criminal' should pick out in order to serve this function. Thus, if we

correctly determine what 'war criminal' should refer to, our assessment rests on generalisable moral considerations. This allows us to reliably extrapolate the judgement to other expressions serving the same purpose across different societies.[64]

Suppose, instead, that a conceptual ethicist has no reason to believe that the part of X's function that J concerns is constrained in a way that precludes multiple realisability. Then the possibility that the features enabling X's proper functioning in this respect depend on idiosyncratic factors must be taken seriously. Still, there may be strong evidence satisfying the sub-condition (ii) in 2, accessible without separately assessing Y. Such evidence would strongly support the claim that X's and Y's proper functioning depends on similar idiosyncratic factors. Then, the possibility that X and Y differ in what enables their proper functioning becomes negligible. The example of 'afterlife' helps illustrate the type of evidence in question.

Imagine a conceptual ethicist who correctly judges which features 'afterlife' must have to help a Catholic community mitigate death anxiety. She also knows of another community where a term, Y, serves the same function. From a reliable source, she learns this community is also Catholic and comes across several credible psychological and anthropological studies showing that what conception of afterlife helps people mitigate death anxiety generally depends on their broader religious outlook. These findings strongly suggest the features desirable for 'afterlife' in the first community are also desirable for Y in the second community. Moreover, she can access these findings without separately assessing Y. Hence, she can reliably extrapolate her judgement about 'afterlife' to Y.

One important qualification: the findings in the above example have sufficient evidential strength to qualify as E_2 in RE6 only because the ethicist already knows 'afterlife' and Y are functionally equivalent in their respective communities. Without this knowledge, there is a salient possibility that the two expressions' functions only partially overlap, with Y's function including an additional component whose fulfilment is in tension with the mitigation of death anxiety. Given this possibility, what's desirable for Y's proper functioning is likely to differ from what's desirable

[64] Accordingly, this might justify treating other societies as committed to accepting such a definition of 'war criminal' that enables the term to fulfil its function. If so, this could help explain why we tend to export the social kind constructed by this definition to other societies, even if they don't accept it (see Epstein, 2015, p.124, 2019; Schaffer, 2019, pp.762-766; and Pagano, 2024 for discussion).

for ‘afterlife’. While ‘afterlife’ functions properly to the extent that it effectively alleviates death anxiety, Y’s proper functioning may involve trade-offs between this and fulfilling its other functional component.

For example, the second community’s central motivation for using Y might be not only that it helps mitigate death anxiety, but also that it allows them to cherish the hope that injustices in life will be rectified after death. While ‘afterlife’ might also produce the second effect in the first community, it’s peripheral to their motivation for using the term. Unless this difference can be ruled out as negligible, the findings about the second community’s religious outlook won’t provide strong enough evidence to justify extrapolating the judgement about the proper functioning of ‘afterlife’, as used by the first community, to Y, as used by the second community. Doing so would likely mischaracterise Y, as it would give undue priority to its fear-alleviating function at the expense of its other functional component. After all, each functional component of Y likely recommends that users associate the term with different conceptions of the afterlife — the first favouring forgiveness, the second favouring posthumous punishment. Thus, the proper functioning of Y for the second community cannot mean fully satisfying both components at once. More plausibly, Y functions properly to the extent that it achieves an optimal balance between the two components, proportionate to their relative importance for the community.

What this example demonstrates is that what kind of evidence is strong enough to qualify as E_2 in RE6 depends on what evidence matching the profile of E_1 is available to the assessor. Specifically, it depends on whether the assessor’s evidence about the functional overlap between the originally assessed expression and the target expression suffices to rule out, as negligible, the possibility that — despite the overlap — their conditions for proper functioning differ significantly in other respects, such that the judgement about the proper functioning of the former expression wouldn’t hold for the latter.

Let’s now turn to a different point. Evidence with the E_2 profile in RE6 can sometimes be available even when extrapolating supplementary judgements. As before, this evidence may satisfy either (i) or (ii) in 2. Regarding the sub-condition (i), reflection on an expression’s function can reveal that some supplementary features are generally desirable for it as well as functionally similar expressions. For example, those most responsible for wartime atrocities are often not the individuals directly involved in committing them, but remote decision-makers who issue

commands, orchestrate events, or enable them through poor leadership or policy. To raise awareness of this, it seems generally desirable for ordinary speakers to internalise a conception of 'war criminal' that casts such figures as paradigms. Regardless of whether this conception in any way contributes to the proper functioning of 'war criminal' *qua* legal term by affecting how 'war criminal' is used in international courtrooms, it plausibly helps ordinary people grasp the banality of evil often present in war crimes. Thus, if we make the supplementary judgement that 'war criminal' should paradigmatically refer to figures like Adolf Eichmann, Vladimir Putin, or Benjamin Netanyahu, it seems safe to extrapolate it to functionally similar terms.

For the sub-condition (ii), consider the earlier 'piety' example. There, conceptual ethicists rightly judged that 'piety' should be paradigmatically associated with non-religious behaviour to avoid religious exploitation. While this may not apply to all functionally similar terms, especially those whose religious use is benign, reliable extrapolation may still be possible with the right kind of evidence. Suppose conceptual ethicists identify a different community's expression Y, which is both functionally similar to 'piety' in C and causes similar toxic effects. Suppose also that strong evidence from psycholinguistics — accessible without separately assessing Y — indicates that the objects speakers perceive as an expression's paradigmatic instances tend to significantly influence their evaluative associations with it. As per RE6, such evidence satisfies the sub-condition (ii) in 2, thereby enabling reliable extrapolation.

As we can see, RE6 can guide conceptual ethicists in recognising when their judgements about an expression's proper functioning and supplementary judgements can be reliably extrapolated to other functionally similar expressions. However, I suspect RE6's implications regarding the reliable extrapolation of foundational judgements are more pessimistic. My concern is that evidence with the profile of E_2 in RE6 is rarely available in such cases.

I will first illustrate why evidence satisfying the sub-condition (i) in 2 of RE6 is generally inaccessible. Consider the following case. In many modern societies, the term 'child' and its translations (henceforth 'child-terms') function as protective legal terms, granting individuals classified under them special rights not afforded to adults, such as access to education, protection from labour exploitation, and adequate healthcare. Historically, however, this wasn't always so. In many past societies, child-terms simply denoted an age category without any special deontic status. Children often lacked access to education and healthcare and were subjected to exploitative labour

— often without such conditions being regarded as unlawful or problematic (Stearns, 2017, chs.1 & 4; Cunningham, 2021, ch.4).

Now imagine conceptual ethicists examine a child-term from one such past society, S, and make the foundational judgement that it should function there as a protective legal term. At first blush, it might seem obvious that general truths about children's vulnerabilities support extrapolating this judgement to child-terms in other past societies. But a closer look suggests a weaker claim: generally speaking, human societies should *ideally* have in their language a protective legal term for conceptualising children that ensures the protection of their special rights.

It would be too quick to conclude from this claim that the same foundational judgement that is true about the child-term in S is also true of child-terms in other past societies. This is because the claim is compatible with several salient options for how a child-term should be treated in a past society. One is to repurpose the child-term into a legal term with a formal protective role. But other options include using it as a non-legal, action-guiding term that highlights children's needs, replacing it with a new term serving a legal or informal protective role, or introducing such a term alongside the original one.

Which of these options is most appropriate for any child-term depends on various considerations about the society using it, such as the society's existing treatment of children; whether the child-term is associated with problematic roles or practices; and the strength of such associations. Furthermore, one must consider the state of the society's legal system — particularly its capacity to enforce children's rights, which depends in part on whether supporting legal terms are already in place. It's also important to assess the child-term's role within the broader conceptual framework and how feasible it is to repurpose it without disrupting that network, as opposed to introducing a new, coextensive term that serves the protective function. These are just a few of many relevant considerations.

Consequently, determining the right course for any child-term is more complex than it initially seems. I doubt one can reliably do so without thoroughly examining and weighing the relevant considerations for each society. That is, conceptual ethicists cannot cut corners by extrapolating their foundational judgement about the child-term in S to other child-terms. The only reliable way to form foundational judgements about these terms is through their separate assessment.

For similar reasons, evidence satisfying the sub-condition (ii) in 2 also seems inaccessible for foundational judgements. A variation of the 'piety' example helps illustrate this. Suppose the best response to the harmful religious usage of 'piety' in C is to repurpose it from a thick to a purely descriptive concept. Now imagine a functionally similar term, Y, used in another community where its religious use is equally harmful. At first, this may seem to justify extrapolating the foundational judgement about 'piety' to Y.

However, once alternative responses are considered, it becomes unclear whether this is the best solution for Y simply because it was for 'piety' in C. Obvious alternatives include abandoning Y and either replacing it with a purely descriptive term for reverence and devotion, or not replacing it at all. Other options involve repurposing Y differently — for instance, as a negative evaluative term — or introducing a new term with such a function. So, at least five options are on the table. Which is best depends on several factors: how difficult would it be for users to shift their central motivation for using Y so that it becomes a descriptive or negative term? How feasible is it for them to stop using Y? Does a suitable replacement term already exist in their conceptual repertoire? If not, how difficult would its propagation be?

These questions hinge on complex idiosyncratic facts about Y's users. It seems highly unlikely that we could obtain strong evidence about which option is preferable for Y without separately assessing the expression. Thus, even here, extrapolating the foundational judgement about 'piety' isn't a reliable substitute for separately assessing Y.

The key takeaway from RE6 is that only certain judgements about proper functioning and supplementary judgements, but not foundational judgements, can be reliably extrapolated in conceptual ethics. This conclusion, on reflection, shouldn't be surprising. Judgements about proper functioning and many supplementary ones rest on the background assumption that an expression should retain its current function and place in a user group's repertoire. They ask what features the expression should have, given that function — either to fulfil it properly or for independent reasons. Even if an expression shouldn't, all things considered, be used for its current function or used at all, this doesn't affect what the correct answer to these questions is. By contrast, foundational judgements address a more fundamental question: why should the expression be used at all? This focus makes the truth value of these judgements sensitive to the possibilities that the expression should, all things considered, serve a different function — or none at all, because it should be

replaced or simply abandoned. These possibilities can vary even among functionally similar expressions due to subtle differences in user-related considerations and their relative weight. It's no surprise, then, that reliably disconfirming these possibilities without separately assessing expressions in light of these considerations isn't feasible.

3.4. Objections and Replies

In this section, I address two objections to RE6 — one concerning its usefulness and the other its extensional adequacy. To explain the first objection, recall that the guiding objective behind our search for a reliable extrapolation criterion was to identify a method that is both epistemically accessible and truth-conducive. However, one might question whether RE6 adequately meets this objective. The criterion incorporates two types of evidence — E_1 and E_2 — whose acquisition may still be epistemically challenging for an assessor. Specifically, obtaining E_1 requires gaining insights into how functionally similar Y is to X, while obtaining E_2 requires first identifying whether there are any generally desirable features for expressions with a given function, and, if not, whether Y is subject to similar idiosyncratic factors as X. Even if this process is less epistemically demanding than separately assessing Y, one could argue that it still requires the assessor to acquire a substantive understanding of the community using Y. If so, then the usefulness of RE6 for cross-linguistic conceptual ethics may be called into question.

In response, I would like to revisit the plea made in Chapter 2 for an interdisciplinary and empirically informed approach to conceptual ethics. There, I argued that while identifying an expression's function often requires empirical research, conceptual ethicists needn't conduct such research themselves. Instead, they can collaborate with or consult the work of researchers from other disciplines who are better trained to conduct it. The same applies here: conceptual ethicists aren't required to gather empirical information about the community that uses the less familiar expression on their own. Rather, they can draw on the work of relevant experts.

Some of these experts may be anthropologists, historians, or linguists specialising in the community and its use of the expression, making them well qualified to provide E_1 — evidence about the expression's function. Others might be specialists on topics broadly related to that function, such as psychologists and anthropologists studying the impact of religious beliefs and fear of death in the 'afterlife' case. The work of both groups can help conceptual ethicists obtain E_2, which concerns either generally desirable features for expressions with a given function or

community-specific idiosyncratic factors affecting its proper functioning. While it would be problematic to outsource independent normative assessments to such experts — since this requires conceptual expertise they typically lack — outsourcing the empirical work needed for reliable extrapolation is appropriate, given their training. RE6 thus supports an interdisciplinary model of conceptual ethics that benefits from a division of labour between normatively trained philosophers and empirically trained researchers.

Let's now consider the second objection to RE6, which takes the form of a potential counterexample. Consider the following scenario:

> *Sexual Harassment*: There is a community C_1 that includes the expression 'sexual harassment' in their conceptual repertoire. Their central motivation behind the usage of this expression is that using it serves to raise awareness of unacceptable sexual behaviours in various contexts and guides them in establishing appropriate boundaries around what constitutes acceptable interpersonal conduct with respect to bodily integrity. There is also another community, C_2, that possesses a term Y, defined so that its extension includes exactly the same sexual behaviours as are in the extension of 'sexual harassment' in C_1. However, C_2's central motivation for using Y is radically different from C_1's motivation for using 'sexual harassment': C_2 views the behaviours included in the extension of Y as desirable and erotic, and wants Y to encourage people to engage in such behaviours more extensively. [65]

The functions of 'sexual harassment' in C_1 and Y in C_2 are clearly almost opposites. This results in significant differences in the features each expression must possess to fulfil its respective function. For example, while 'sexual harassment' must have a negative valence and be associated with criminality to fulfil its function, Y must have a positive valence and be associated with perverse thoughts to fulfil its function. Therefore, extrapolating judgements about their proper functioning from one to the other wouldn't be reliable.

That said, the two expressions share their extensions: they are correctly applied to the same behaviours. One might argue that strong evidence of this extensional equivalence suffices to reliably extrapolate the foundational judgement about 'sexual harassment' in C_1 to Y. The

[65] The possibility of a similar expression has recently been discussed in Bell (2025, pp.382-384).

reasoning might go as follows: 'the correct judgement about "sexual harassment" in C_1 is likely conservative — the expression should keep its existing function. After all, this function is so important that any community arguably needs an expression fulfilling it. If C_1 already has such an expression, it should remain unchanged. Furthermore, we can arguably reliably extrapolate this judgement to Y, concluding that Y should serve the same function in C_2 as "sexual harassment" does in C_1. After all, this function is clearly superior to Y's current function, as it promotes correct moral responses to the behaviours in the shared extension of the two expressions. Since RE6 requires functional similarity and thus cannot accommodate this intuition, it should be rejected.'

My response to this objection draws upon my earlier discussion of extrapolating foundational judgements. I can endorse the following three claims:

> *Claim 1*: Ideally, C_2 should have an expression with the same function as 'sexual harassment' in C_1.
>
> *Claim 2*: Y's current function in C_2 is morally reprehensible and unfit for any expression in any language.
>
> *Claim 3*: Claim 1 and Claim 2 are general judgements that can be reliably made without separately assessing Y.

However, these three claims are compatible with at least three alternative foundational judgements about Y being true. First, Y in C_2 could be repurposed to serve the same function as 'sexual harassment' in C_1. Second, Y in C_2 could be replaced with a new expression serving that function. Third, Y in C_2 could simply be abandoned — temporarily, without replacement. Determining which judgement is correct likely requires separately assessing Y, since it depends on specific factors about C_2: the strength of their erotic associations; the extent to which C_2's members already understand the moral wrongness of behaviours in Y's extension; how difficult it is for them to cease using Y; and whether, after abandoning Y, they would be more likely to suppress their erotic desires by diverting attention from the acts within its extension or by adopting a new term that foregrounds their moral wrongness. Simply extrapolating from 'sexual harassment' in C_1 to conclude that the first judgement is correct seems unreliable compared with carefully evaluating these and other relevant considerations. Thus, RE6 remains unchallenged.

3.5. Conclusion

This chapter aimed to guide conceptual ethicists on when they can reliably extrapolate advisory judgements from an expression used by one community to a similar expression used by another. My guidance was relatively restrictive, as I argued that such extrapolation is generally possible only under the specific conditions outlined in RE6.

To illustrate how my guidance works in practice, imagine a conceptual ethicist who makes advisory judgements about the features 'humility' should have when used by Christian ethicists and seeks to identify which of these judgements can be reliably extrapolated to 'qiānxū' as used by Confucian ethicists. If my argument for RE6 is correct, she must ask two key questions. First, is there strong evidence — obtainable without separately assessing 'qiānxū' — that the central motivations for using 'humility' and 'qiānxū' at least partially overlap? Second, if so, is there strong evidence — also obtainable without separately assessing 'qiānxū' — that some features are desirable for both expressions, either to fulfil these overlapping motivations or for related reasons?

It isn't difficult to imagine how the ethicist can determine which judgements about 'humility' can be reliably extrapolated to 'qiānxū' through inquiry into these questions. For example, a central motivation for both Christian and Confucian ethicists might be to guide users towards cultivating a moral virtue that helps them avoid the vice of pride. Moreover, there may be universal moral facts about what this virtue is. If so, a correct advisory judgement about how 'humility' should be interpreted by Christian ethicists to fulfil this motivation can be reliably extrapolated to 'qiānxū' for Confucian ethicists.

Overall, I believe any theorist practising conceptual ethics cross-linguistically should keep my proposed reliable extrapolation criterion, RE6, in mind. As noted in §3.1, it's often unrealistic to expect theorists to be in a good epistemic position to conduct separate assessments of unfamiliar expressions in other languages. Consequently, such theorists have no choice but to rely on the method of extrapolation when assessing expressions across linguistic boundaries. By ensuring that their extrapolations are guided by the evidence specified in RE6, these theorists can avoid being ignorant of the relevance of their conceptual assessments to other communities, while also minimising the risk of overestimating the applicability of evaluative standards tied to their own

community — thus avoiding faulty generalisations about what features make unfamiliar expressions valuable for their users.

Chapter 4. A Case Study of Thick Concepts

4.1. Introduction

Thus far in this dissertation, I have developed the motivation-based notion of conceptual function and demonstrated its value as a theoretical resource in cross-linguistic conceptual ethics. This chapter concludes the portion of the dissertation focused on function-sensitive conceptual ethics by presenting one further application of the proposed notion of conceptual function. I argue that this notion helps philosophers deepen their understanding of thick ethical concepts.

Thick ethical concepts are ethical concepts whose conveyed content includes both evaluative and descriptive components. In its current state, the debate about these concepts doesn't directly address their assessment and thus lacks immediate relevance to conceptual ethics, unlike the other topics covered in this dissertation. Nevertheless, it matters for conceptual ethics indirectly. To explain: it's no accident that, in the preceding chapters, I discussed several thick ethical concepts such as 'humility', 'piety', 'war criminal', and 'sexual harassment'. Thick concepts cry out for assessment. They have a profound impact on how users interpret and engage with the world, making it important to ask which thick concepts are worth relying on.[66] Such inquiry naturally raises further questions: Is a thick concept's descriptive content apt for moral evaluation? If so, what evaluation does it merit? To address these questions, we must first gain a better understanding of the relationship between the evaluative and descriptive components of thick concepts. While this relationship is a central point of contention in the current literature on thick concepts, I argue that my notion of conceptual function offers a way to clarify it.

The chapter is structured as follows. In §4.2, I first show that the motivation-based notion of conceptual function allows us to recognise an instructive but neglected distinction between two kinds of characterisations of the properties that expressions track: their significance-explaining characterisations and their real definitions. In §4.3, I set the stage by explaining what thick ethical concepts are and introduce a long-standing dispute between separabilists and inseparabilists over how the evaluative and descriptive components relate to each other. In §4.4, I introduce the most common argument against separabilism, known as the 'Disentangling Argument', and the most

[66] The first gestures towards this inquiry can be found in Eklund (2017, pp.192-203), Ohlhorst (2023), Brey (2024), and Queloz (2025, chs.1-2)

influential counterargument to it, advanced by Elstein and Hurka (2009). In §4.5 and §4.6, I employ the distinction between the real definitions and significance-explaining characterisations of the properties that thick concepts track to raise doubts about Elstein and Hurka's counterargument, while arguing that some aspects of their proposed analysis can still be useful for analysing the function of thick concepts.

4.2. A Neglected Distinction

Recall the Value-based Schema for specifying conceptual function in accordance with the motivation-based interpretation.

> *Value-based Schema*: The function of an expression 'A', relative to a group of its users G, consists of a representational effect R of tracking a property that best satisfies desiderata D_i-D_n, and of extra-representational effect(s) E_i-E_n, such that both R and E_i-E_n are central to explaining why G's members regard 'A' as a valuable enough expression to include in their conceptual repertoire.

The schema highlights two distinct questions that we can ask about a property P that an expression functions to represent for a user group:

> *Question 1*: Why is the user group disposed to consider P significant enough to be conceptualised by a separate expression in their language?
>
> *Question 2*: What is it to be P?

According to the Value-based Schema, answering Question 1 is crucial for successfully identifying an expression's function. After all, the schema states that this function includes a representational effect: tracking the property that best satisfies certain desiderata. These desiderata articulate the reasons for considering P worth conceptualising — precisely what Question 1 addresses. Thus, we can say Question 1 concerns P's *significance-explaining characterisation*: the characterisation that explains why P is significant enough a property to merit conceptualisation.

In contrast, Question 2 concerns what is commonly known as the *real definition* of P. This type of definition applies to objects and properties themselves, rather than to expressions. A property's real definition provides a set of conditions in virtue of satisfying which objects

instantiate it.[67] In neo-Aristotelian terms, these conditions articulate the very nature — or essence — of the given property that uniquely individuates it from all the other properties.[68] The construal of real definition in terms of essence has a long history in philosophy. It originates in Aristotle (c. 350 BCE/1960, 102a3) and can also be found in the writings of Spinoza (1677/1997, §95), Locke (1689/1975, pp.414-420), and Moore (1942, pp.664-665). Its recent defenders include Fine (1994a), Koslicki (2012), and Dasgupta (2014).[69]

What real definition a property has undoubtedly affects what significance-explaining characterisations it satisfies. Nevertheless, properties' significance-explaining characterisations should be distinguished from their real definitions because they don't have to be identical. This can be appreciated through the following observation: since a property's significance-explaining characterisation explains why a user group deems it worth conceptualising, it identifies it relative to the group's evaluative outlook and circumstances. By contrast, since the property's real definition captures what is essentially, and therefore necessarily, involved in its instantiation, it identifies it invariably. This observation has two noteworthy implications.

The first implication is that a single significance-explaining characterisation may be satisfied by one property in one set of circumstances but by a different property — or none — in another. That is, some such characterisations may identify properties that satisfy them only contingently. For example, imagine a community using the term 'salt' to track and communicate about the kind that satisfies a culinary desideratum (enhancing flavour) and a nutritional one (providing necessary minerals). In their environment, this characterisation is satisfied by sodium chloride (NaCl). But it could just as well be satisfied by a different substance — call it XYZ — that is chemically distinct but tastes the same and serves the same dietary functions. Suppose all naturally occurring NaCl were replaced by a compound with a different microstructure but identical taste and nutritional profile. In that case, the community would presumably begin using

[67] 'In virtue of' is commonly understood in the context of real definitions as expressing a metaphysical explanatory relation known as 'grounding'. See especially Fine (2015, pp.306-308), Rosen (2015, pp.197-200), and Correia (2017, pp.60-61) for various ground-theoretic accounts of real definition.

[68] That said, as Passinsky (2025, p.1032) notes, Neo-Aristotelians aren't committed to the view that all properties have fully individuating real definitions, but only to the view that they have partially individuating ones.

[69] Note that cashing out real definition in terms of essences doesn't commit us to demanding accounts of essence, such as those according to which essential properties must be non-gerrymandered (Mallon, 2007), manifest unity (Oderberg, 2011), be metaphysically sparse (Wildman, 2013), or intrinsic to their instances (Denby, 2014). See Passinsky (2025), who defends this point regarding the essences of social kinds.

XYZ for seasoning and mineral intake. The original characterisation would then be satisfied not by NaCl but by XYZ.

To clarify, some significance-explaining characterisations can identify properties that satisfy them necessarily. This happens, for example, when a community deems a property worth conceptualising solely because of its defining features. Suppose, for example, that the sole reason why some speakers consider the kind *hammer* worth conceptualising is that it's a handheld tool used to deliver physical impact to small areas. The example seems highly realistic because this significance-explaining characterisation arguably captures why many people primarily care about hammers. Moreover, many would argue that it explains what it is to be a hammer and thus amounts to the real definition of *hammer*. Accordingly, significance-explaining characterisations and real definitions may sometimes converge.

That said, just because a significance-explaining characterisation identifies a property that necessarily satisfies it doesn't mean this characterisation amounts to its real definition. To explain, while all essential features of an item are necessary features, the reverse — that all necessary features are essential — has been widely challenged since Fine's (1994a; 1994b, pp.56-58) influential critique. He offers examples of properties that are necessary but intuitively inessential, showing that our pre-theoretical grasp of essence can't be captured purely in terms of metaphysical necessity. For instance, being water and such that that 2+2=4 is necessarily true of any water sample but isn't essential to water, since it doesn't reflect what it is to be water. This leaves room for a property's significance-explaining characterisation to differ from its real definition, even if the two are necessarily co-extensive.

The following scenario serves to illustrate this possibility. Imagine a peculiar community that generally shows indifference towards geometry, resulting in the lack of conceptualisation of geometric kinds in their language. However, there is a single exception: when a mathematical theorem revolves around a specific geometric kind. In such cases, the community is incentivised to invent an expression standing for the given geometric kind. However, what makes the community consider the kind worth conceptualising is only its association with a specific theorem. The community disregards all other characteristics of it, including its defining characteristics.

While the given community uses the term 'triangle' to represent the kind *triangle*, they regard this kind as worth conceptualising solely because it satisfies the characterisation 'the kind to which the Pythagorean theorem applies whenever it has a 90-degree angle'. This characterisation is necessarily satisfied by an object iff it's a triangle. Still, it seems implausible to treat it as the real definition of *triangle*. What explains what it is to be a triangle — and thus qualifies as its real definition — is presumably a characterisation that appeals to more explanatorily basic and intrinsic features of triangles, such as being a polygon with three sides and three angles. Although this description is necessarily coextensive with the above significance-explaining characterisation, it clearly differs from it in conveying distinct information about triangles. To account for this, we must individuate property characterisations at a hyperintensional level fine-grained enough to distinguish necessarily coextensive characterisations that differ in informational value.[70]

Let's now turn to the second implication, which is a mirror image of the first: a single property can satisfy multiple distinct significance-explaining characterisations. Take the property of being true, instantiated by entities like beliefs or sentences. This property can plausibly satisfy various significance-explaining characterisations across different communities — for example, 'the property that serves as the ultimate goal of scientific inquiry', 'the property whose ascription signals acceptance of a belief or sentence', or 'the property identifying beliefs a community relies on in interacting with the world'. Each could count as the property's significance-explaining characterisation because we can imagine a user group situated in circumstances where this property satisfies it and is worth conceptualising for that reason. Thus, there appears to be no strict limit on how many significance-explaining characterisations a property can satisfy.

However, the same cannot be said about real definitions. To be sure, it has been argued that a single property may have more than one correct real definition whenever there are at least two equally strong candidates for accounts of what it is to instantiate it (Rosen, 2015, p.201). Nevertheless, it seems an impossible task to identify a property for which we can list countless

[70] If we allow that, for any characterisation of a property, there is a corresponding property defined by the given characterisation — an option that isn't indispensable (Rosen 2015, p.205) — we must adopt a hyperintensional individuation of properties to ensure the coherence of our account. Otherwise, we would have to claim that the property defined as 'polygon with three sides and three angles' is identical to that defined by the Pythagorean-theorem-involving characterisation of *triangle*, which would contradict our claim that the latter isn't its real definition.

alternative explanations that are equally strong in capturing what is essentially involved in instantiating it. This suggests real definitions are scarcer than significance-explaining characterisations.

4.3. Thick Concepts: Separabilism and Inseparabilism

The above discussion was meant to shed light on the distinction between significance-explaining characterisations and real definitions of properties. n the remainder of this chapter, I aim to demonstrate the usefulness of this distinction for theorising about *thick ethical concepts*.

Let's first clarify what thick concepts are. The distinction between thick and thin ethical concepts stems from an intuitive contrast between expressions such as 'just', 'cruel', 'courageous', 'benevolent', or 'selfish' and expressions such as 'good', 'bad', 'right', or 'wrong'. The former expressions seem more complex in what they convey because, unlike the latter expressions, they don't only convey that the objects to which they are applied merit a positive or negative moral evaluative response, but also that they satisfy some further at least partially non-evaluative condition(s) in virtue of which they merit it. Accordingly, while the content conveyed by the latter expressions is exhausted by the evaluative component, the content conveyed by the former expressions is 'thicker' in that it consists of both evaluative and non-evaluative components. This motivates calling the former expressions 'thick concepts' and the latter expressions 'thin concepts'.[71]

To illustrate, describing someone as 'cruel' conveys not only that they merit negative moral evaluation but also that they do so in virtue of satisfying some further condition, such as intentionally inflicting disproportionate harm upon others. Not all terms in this condition are evaluative, and determining whether the condition is satisfied clearly involves more than merely negatively judging the individual. Accordingly, the condition can be characterised as at least partially non-evaluative. By contrast, describing someone as 'bad' conveys that they merit negative evaluation but doesn't impose any more specific constraints on what makes them merit it. They can merit it not only in virtue of intentionally inflicting disproportionate harm upon others but also in virtue of deceiving others, manipulating others, betraying others, among other reasons.

[71] There have also been other attempts to define the distinction between thin and thick concepts, but all of them are contested. See, e.g., Hare (1963, pp.24-25), Williams (1985/2006, pp.140-141), and Gibbard (1992, pp.268-269).

'Bad' doesn't convey anything about which of these options is true, which is the sense in which it's more purely evaluative than 'cruel'.[72]

The most discussed question surrounding thick ethical concepts is how their evaluative and non-evaluative components relate to each other. In general, there are two competing answers to it:

> *Separabilism*: The evaluative and non-evaluative components of thick concepts are discrete components that can be disentangled from each other.[73]
>
> *Inseparabilism*: The evaluative and non-evaluative components of thick concepts present an irreducible fusion, and thus they cannot be disentangled from each other.[74]

Separabilism is a *reductivist* position because it implies that it's possible to reductively analyse thick concepts in terms of non-evaluative descriptions and thin concepts. That is, to create a thick concept, it's enough to add some descriptive content to the purely evaluative content of thin concepts. This would mean there is nothing distinctive about the evaluation conveyed by thick concepts that goes beyond the evaluation conveyed by thin concepts. In that case, all moral judgements can be expressed using basic thin concepts such as 'good', 'bad', 'right', and 'wrong' (Elstein & Hurka, 2009, pp.515-516).

Inseparabilists reject such a reductivist analysis as simplistic, as they believe the evaluative and descriptive components of thick concepts are inextricably tied to each other. In other words, no matter how much we attempt to separate out descriptive and evaluative components during a reductivist analysis of a thick concept, we will ultimately discover that we are in fact helping ourselves to a thick concept in such an analysis, because either its seemingly purely descriptive component is at least partially evaluative, or its seemingly purely evaluative component is partially descriptive.[75]

---

[72] It might be better to view the distinction between thick and thin as a gradable continuum and thus treat ethical concepts as more or less thick rather than categorically thin or thick (cf. Scheffler, 1987, pp.417-418; Williams, 1995, p.234; Väyrynen, 2013, p.7). For example, many would argue that 'ought' falls closer to the thin end of the continuum, but it still involves some modal constraints.

[73] See Stevenson (1944, pp.206-207), Hare (1952, pp.121-122), Gibbard (1992), Elstein and Hurka (2009), Blackburn (1992), and Kyle (2020).

[74] See Williams (1985/2006), Putnam (2002, pp.34-43), Kirchin (2017), Roberts (2013a, 2013b, 2017), Dancy (1995), Chappell (2013), and Harcourt and Thomas (2013).

[75] Some inseparabilists would even argue that any correct analysis of a thick concept is circular because it relies on the very same thick concept that is its analysandum (Harcourt & Thomas, 2013).

To illustrate, suppose separabilists offer the following analysis of the content conveyed by the thick concept 'kind': 'being morally good in virtue of treating people with respect, being helpful, and being considerate of their well-being'. For the sake of argument, let's assume this analysis is correct. Separabilists might also view it as reductivist because the evaluative component is exhausted by the thin concept 'morally good', and the remaining terms are purely descriptive. Inseparabilists, however, would challenge this by arguing that the seemingly descriptive component is partly evaluative, as it includes 'helpful' and 'considerate', which are themselves thick concepts conveying moral evaluation. They would further argue that the same problem arises even when trying to reductively analyse these concepts. Consequently, it becomes impossible to reach a base-level analysis of thick concepts in which evaluative and descriptive components are neatly separated. This leads inseparabilists to argue that the evaluation conveyed by thick concepts is irreducibly thick, and so any attempt to analyse it reductively in terms of thin concepts fails to capture its complexity and distinctiveness.

4.4. The Disentangling Argument and Elstein & Hurka's Counterargument

In this section, I explain the most common argument against separabilism, known as the 'Disentangling Argument', along with the most prominent counterargument to it, presented by Elstein and Hurka (2009). I will later use these arguments to show how the distinction between the significance-explaining characterisations of properties and their real definitions can enrich the debate between separabilists and inseparabilists.

4.4.1. The Disentangling Argument

Here is what I consider to be the strongest formulation of the Disentangling Argument:

The Disentangling Argument:

(Premise 1) Separabilism entails that the extension of thick concepts is determined only by their descriptive component.

(Premise 2) There are many thick concepts whose extension is partially determined also by their evaluative component.

(Conclusion) Separabilism is false.

Before I explain what supports the argument's premises, I first need to say something about the argument's historical background. The argument was originally introduced by Jonathan McDowell (1981), primarily as an argument against metaethical non-cognitivism, which is the view that moral judgements express non-cognitive attitudes and therefore their content isn't apt for assessment as true or false. The argument was recast as targeting separabilism about thick concepts only later by Bernard Williams (1985/2006, pp.130, 141-142). Williams thinks this reorientation is warranted because he assumes that separabilism presupposes non-cognitivism, which is also reflected in his justification of Premise 1. On this assumption, Premise 1 appears evidently true. After all, non-cognitivism, by its definition, rules out the possibility that the evaluative component contributes to the truth-conditional meaning of thick concepts, thereby playing a part in determining their extension.[76] In that case, the task of governing the application of thick concepts by determining their extension falls entirely upon the truth-conditional part of their content, specifically their descriptive component.

It's, to some extent, understandable that Williams made the above assumption, considering that some prominent defenders of separabilism were non-cognitivists, such as Stevenson (1944) and Hare (1952). These non-cognitivists had good reason to defend separabilism. After all, given the truth-conditionality of the descriptive component, if the evaluative and descriptive components of thick concepts were inextricable, as inseparabilism claims, evaluative judgements conveyed by thick concepts would present a counterexample to the non-cognitivist thesis that moral judgements aren't truth-apt.

However, as Elstein and Hurka (2009, p.517) note, Williams' assumption is false. Although non-cognitivists are committed to separabilism, some separabilists can consistently be cognitivists — arguing that evaluative judgements conveyed by thick concepts are truth-conditional yet reductively analysable in terms of thin concepts. Therefore, for the Disentangling Argument to present a genuine challenge to any form of separabilism, Premise 1 cannot be taken as self-evident; it requires a defence addressing both its non-cognitivist and cognitivist interpretations.

One possible defence of Premise 1 runs as follows: while separabilists see the evaluative and descriptive components of thick concepts as disentanglable, they acknowledge that the two

[76] See, notably, Stevenson (1944), Hare (1952), Blackburn (1998), and Gibbard (2003) for details on non-cognitivism.

components are explanatorily dependent, as they treat the former as the response to the latter. This treatment seems plausible. For example, if, as separabilism suggests, 'rude' is analysed as thin negative moral evaluation plus non-evaluative conditions related to disrespect, then describing behaviour as 'rude' plausibly conveys that it's morally bad *precisely because* it meets these conditions. Thus, the descriptive component explains the evaluative component, i.e., why the conveyed evaluation is correct.[77] And so, the argument goes, separabilists view the evaluative component as a mere explanandum of the descriptive one, with no independent role in determining extension.

Let's now consider how Premise 2 can be defended. What lends support to this premise is the observation that we have a poor track record in identifying the extension of thick concepts independently of grasping their evaluative component. To illustrate, think of the wide variety of actions that uncontroversially belong to the extension of thick concepts such as 'cruel' or 'courageous'. The extension of 'cruel' ranges through such actions as torturing someone, letting a student fail final exams, or ghosting someone. The extension of 'courageous' ranges through such actions as running into a burning house to rescue a child, going on a hunger strike when calling for a political reform, or coming out. In both cases, the respective instances of cruel and courageous actions seem to be so diverse in their descriptive features that it looks like an impossible task to extract some generalisable *purely descriptive conditions* that all of them share, on the basis of which we can reliably anticipate the correct usage of 'courageous' and 'cruel' in new cases to the extent that we might be said to master their extension. This seems to be strong evidence that the extension of many thick concepts is non-evaluatively shapeless, which is to say, it cannot be intelligibly explained in purely descriptive terms.[78] [79] Since thick concepts are assumed to group

[77] This observation also aligns with contemporary non-cognitivism, which typically holds that, even if moral judgements aren't truth-apt, they are systematically connected to descriptive facts and can be minimally assessed for correctness (cf. Dreier 2004; Bedke 2017, pp.297-298).

[78] For discussion, see Putnam (2002, pp.39-40), Kirchin (2010, 2017, pp.80-111), and Roberts (2011, 2013b). Of course, this isn't to say that thick concepts whose extension has a non-evaluative shape are categorically impossible. The point, rather, is that many actual thick concepts have a broader and more heterogeneous scope of application, making purely descriptive explanations of their extension inadequate.

[79] Even moral naturalists struggle to explain the extensions of many thick concepts in purely descriptive terms, as their accounts of moral goodness still rely on evaluative concepts, though framed naturalistically. For instance, Boyd (1988, pp.195-203) sees moral goodness as a homeostatic property cluster promoting *important* human *needs*, while Railton (1986, pp.189-191) defines it as what fully *rational* individuals would desire from an *impartial* standpoint. Terms like 'important', 'needs', 'rational' and 'impartial' clearly retain evaluative content.

together objects in a non-arbitrary, and therefore at least pro tanto intelligible, way, it follows that the evaluative component is often involved in determining their extension, just as Premise 2 says.

This summarises the motivation behind the premises of the Disentangling Argument. In a nutshell, the argument criticises separabilism on the grounds that it treats the evaluative component of thick concepts as being semantically inert despite evidence to the contrary.

4.4.2. Elstein & Hurka's Counterargument

Elstein and Hurka (2009) defend separabilism against the Disentangling Argument by rejecting Premise 1. Contrary to Premise 1, they argue that separabilism can, in principle, offer a reductivist analysis of thick concepts that accommodates Premise 2, which holds that the evaluative component partly determines their extension. To this end, they propose two analyses of thick concepts that are reductivist but compatible with Premise 2. I limit my focus here primarily to the first analysis, as I believe this is the analysis in relation to which the distinction between significance-explaining characterisations and real definitions proves most useful. But I also touch upon the second analysis along the way.

The blueprint of Elstein and Hurka's first analysis of thick concepts is the following pattern (2009, p.521):

> *Underspecified Pattern*: To fall under thick concept 'C' is to merit thin evaluation and have properties X, Y, and Z (not specified) of general profile P (specified) that make anything that has them merit it.

The underlying idea behind the Underspecified Pattern is that Premise 1 overlooks the possibility that the descriptive component of thick concepts isn't fully specified. Instead of being fully specified, it may just place some general constraints on what can fall under a thick concept, which alone are insufficient to fully determine its extension without the aid of the evaluative component.

Elstein and Hurka illustrate the Underspecified Pattern by applying it to the thick concept 'distributively just' (2009, pp.521-522). They suggest analysing the semantic content of 'distributively just' as UDJ:

> *UDJ*: To be good and have properties X, Y, and Z, whatever they are, that distributions have as distributions, or in virtue of their distributive shape, and that make any distribution that has them good.

UDJ is a reductivist analysis in that it only contains descriptive concepts and the thin concept 'good'. Elstein and Hurka call the evaluative component conveyed by 'good' in UDJ 'global thin evaluation' because it governs the whole concept, serving as an evaluative response to the entire descriptive component rather than to its part (2009, p.526). Moreover, the analysis only presents a very general characterisation of the properties X, Y, Z in the descriptive component, without specifying what exactly these properties are. Consequently, the descriptive component alone underdetermines the concept's extension. However, the global thin evaluation complements it by identifying X, Y, Z as *good-making properties*. Accordingly, the analysis allows for the following division of semantic labour between the descriptive component and the global thin evaluation in determining the extension of 'distributively just': the descriptive component generally delineates the concept's extension by stating that it only contains distributions that exhibit certain properties as distributions, or in virtue of their distributive shape, and the global thin evaluation further specifies that all of these properties are good-making.

To clarify, Elstein and Hurka don't argue that there is no intelligible way to further specify X, Y, Z (2009, p.523). Why then do they think this specification may not be built into the semantic content of 'distributively just'? To my understanding, they think that this may be due to the epistemic limits of the concept's users, who either struggle to entertain the correct specification of X, Y, Z, or are able to entertain it but widely disagree among themselves over whether it's correct.

For example, users of 'distributively just' may consider several candidate specifications for X, Y, Z — such as 'distribution where goods are allocated equally', 'distribution proportioned to one's merits', and 'a distribution maximising a society's overall happiness' — but disagree on which is correct (2009, p.521-522). Moreover, the true specification might be an alternative they haven't yet considered. Consequently, the correct specification may be inaccessible, leaving them no choice but to collectively identify the extension of 'distributively just' via UDJ.

Elstein and Hurka take the above example to shed light on a possible way in which the descriptive component of thick concepts can partially 'outsource' its role in determining extension

to global thin evaluation, which would then allow for the reductivist analysis of thick concepts through the Underspecified Pattern (2009, p.524). This leads them to conclude that the Disentangling Argument fails because it disregards this possibility in Premise 1.

4.5. Assessing Elstein & Hurka's Counterargument

I agree with Elstein and Hurka that the Underspecified Pattern illustrates what a reductivist analysis allowing the evaluative component of thick concepts to partly determine their extension might look like. However, I have two concerns about their counterargument. First, their counterargument doesn't convincingly show that such an analysis is applicable to thick concepts as they *actually* function in human languages. Second, even if realistic, the analysis fails to fulfil its core theoretical motivation of showing that there is nothing distinctive about the evaluative component of thick concepts beyond thin evaluation. Both concerns reflect my scepticism about the Underspecified Pattern's suitability for the role Elstein and Hurka assign it regarding the Disentangling Argument.

I will now elaborate on these concerns while bringing to the stage my distinction between real definitions and significance-explaining characterisations. In brief, addressing these concerns requires global thin evaluation to be part of the real definition of the properties tracked by thick concepts — which seems unlikely. Still, I argue that global thin evaluation is crucial for analysing thick concepts, as it helps articulate their significance-explaining characterisations.

4.5.1. Concern 1: The Semantic Role of the Underspecified Analysis

Let's start with the first concern. In order for global thin evaluation to play an extension-determining role for thick concepts, the Underspecified Pattern must offer the correct analysis of their *meaning*. That is, the pattern must capture their *actual semantic content*. As previously discussed, Elstein and Hurka seem to infer that this condition obtains from the fact that the descriptions the users fall back on to collectively identify the extension of thick concepts are commonly underspecified. I find this inference problematic, as it presupposes that thick concepts are governed by the descriptivist theory of meaning (and reference). Descriptivism, however, isn't a metasemantic theory that can be just presupposed in one's argument, given that it has been seriously challenged by the advent of the causal-historical theory of meaning (and reference).

The difference between the two theories can be articulated as follows. On the descriptivist theory, an expression's meaning is identified with a description (or some cluster of descriptions) that the concept's users collectively associate with it. The expression then refers to the item that best satisfies this description (or the given cluster).[80] By contrast, on the causal-historical theory, an expression's meaning is identified with an item whose initial baptism (or another appropriate initial defining event directed towards the object) is the causal origin of the present use of the expression. The expression also picks out this item as its referent.[81]

Now, on the causal-historical theory, the mere fact that an expression's users rely upon a description to identify its extension doesn't suffice to establish that the description also captures its semantic content. The following variation of the well-known scenario from Putnam (1975) will drive this point home: imagine a past linguistic community whose understanding of chemistry isn't advanced enough to know the microstructure of chemical substances. In this community, a member introduces the word 'water' to their language by pointing to a sample of the substance with the chemical composition H_2O and stipulating the following ostensive definition of 'water':

> *Ostensive Definition*: 'Water' refers to the substance with the same chemical composition as this sample has.

This demonstrative act successfully baptises 'water' and initiates a causal-historical chain of communication within the community regarding the given substance using 'water'. Consequently, all the subsequent uses of 'water' in the community become causally linked to the initial act. Moreover, the H_2O sample that is referred to as 'this sample' in the Ostensive Definition is collected and displayed to the community as a paradigmatic instance of what they call 'water'. As a result, the Ostensive Definition encompasses the description that the community's members dominantly associate with 'water'. Additionally, since the community doesn't have any specific beliefs about what the chemical composition of the substance they call 'water' is, the Ostensive Definition also becomes the description that all its members employ to identify the H_2O substance and, consequently, the extension of 'water'.

---

[80] Descriptivism has been mainly defended in relation to proper names by Frege (1893/1980), Russell (1905), and Searle (1958). Its defence in relation to theoretical terms can be found in Lewis (1972) and Jackson (1998).
[81] Various versions of the causal-historical theory have been developed by Donnellan (1970), Kripke (1980), and Putnam (1975).

Now, causal-historical theorists would acknowledge that the Ostensive Definition serves a heuristic role for the community in this scenario: it's metasemantically employed to fix the reference of 'water' and guides the community's members in their usage of the expression by enabling them to identify indirectly what substance 'water' picks out, even before they have enough information about its chemical composition.[82] However, they would deny the Ostensive Definition expresses the semantic content of 'water' (Kripke, 1980, pp.135-136; Putnam, 1975, p.150). Instead, they would argue that its semantic content is expressed by the real definition of the chemical substance indirectly identified by the Ostensive Definition, whose baptism is the causal origin of their use of 'water' — namely, the substance with the chemical composition H_2O.

By now, it's likely that the reader has recognised why I brought up the aforementioned scenario. The scenario serves to illustrate the parallel role played by the Ostensive Definition for the users of 'water' and UDJ for the users of 'distributively just'. UDJ, much like the Ostensive Definition, serves as an underspecified description that enables the users of 'distributively just' to delineate the expression's extension even before they know more specific details about what makes objects fall within it. It likewise seems plausible that this identificatory role of UDJ might have been metasemantically utilised by those who introduced 'distributively just' into the language to fix its reference. Nevertheless, causal-historical theorists would deny that UDJ captures the expression's semantic content. Instead, they would view UDJ as merely a heuristic tool for indirectly identifying the property whose real definition amounts to the expression's semantic content. This is the property whose initial baptism is the causal origin of our 'distributively just'-involving communication, i.e., the property of being distributively just. We would presumably arrive at this property's real definition by replacing UDJ's generic specification of X, Y, Z as 'good-making properties that distributions have as distributions, or in virtue of their distributive shape' with a specification offering a more detailed explanation of the extension of 'distributively just'. Unlike UDJ, such a definition would leave no room for any separable thin evaluation that could partly determine the extension of 'distributively just'.

[82] The Ostensive Definition isn't the only description the community can use to indirectly fix the referent of 'water.' As Putnam (1975, pp.169-170) and Salmon (1981, pp.154-155) note, the community can also rely on an operational definition based on stereotypical features of water — e.g., '"Water" refers to the odourless, colourless, thirst-quenching liquid that fills lakes and oceans around here'.

The above case of 'distributively just' effectively demonstrates why, under the causal-historical theory, the Underspecified Pattern fails to capture the meaning of thick concepts, thereby failing as a counterexample to Premise 1. To be sure, if Elstein and Hurka manage to establish that descriptivism is true of thick concepts, they can avoid this outcome. After all, descriptivism identifies an expression's meaning with the descriptions its users collectively associate with it. The descriptions derived from the Underspecified Pattern appear to be strong candidates for these associated descriptions in the case of thick concepts. Nonetheless, their argument, in its present form, lacks any defence of descriptivism with respect to thick concepts, and thus the authors still owe us such a defence.

What makes the incompleteness of Elstein and Hurka's argument more pressing is that, unlike the causal-historical theory, descriptivism about thick concepts is vulnerable to the objection that it cannot accommodate the plausible possibility of large-scale collective error about their referents.[83] For example, suppose our linguistic community, influenced by neoliberal views, uniformly believes that morally good distributions allocate goods on the basis of individual desert. This view, however, may be false, overlooking how luck-related factors — such as upbringing, natural talent, and access to resources — shape one's capacity to achieve something. Suppose we later recognise this error and ask how we would then respond to it.

Intuitively, we would say we were wrong about what it is to be distributively just, in the sense that we have come to see that 'distributively just' refers to something other than we previously thought. But under descriptivism, this conclusion is unavailable. Descriptivism holds that the referent of 'distributively just' is fixed by the description (or description cluster) we predominantly associate with it. In this scenario, the dominant description would be something like 'a distribution that is morally good in virtue of allocating goods on the basis of individual desert'. Descriptivism thus yields the implausible prediction that, in light of our realisation, we would conclude that nothing is distributively just, as we use the term, because it has an empty extension. By contrast, the causal-historical theory can preserve the intuitive verdict: the referent was fixed when the term was first introduced, likely via UDJ, and may diverge from more specific

[83] This argument parallels arguments from error against descriptivism about proper names (Kripke, 1980, pp.83-92), natural kind terms (Kripke, 1980, pp.118-123; Putnam, 1975, p.153; Strevens, 2019, pp.150-154), and artefactual kind terms (Putnam, 1975, pp.160-163; Nelson, 1982).

descriptions that the community later dominantly associates with the expression. Even if not decisive, this argument at least shows that the causal-historical theory is a serious alternative to descriptivism about thick concepts.

One might object, however, that causal metasemantics struggles with its own problems, which also extend to thick concepts. For one thing, the causal-historical theory originally proposed by Kripke and Putnam for the metasemantics of proper names and natural kind terms is often criticised on the grounds that it cannot account for the plausible phenomenon of a term changing its reference over time. This is because it presents the term's reference as being stably grounded in how it was initially defined when first introduced into the language, rather than in how it's used at particular times in its history (Evans, 1973; Fine, 1975; Devitt, 1981). This problem prompted Evans (1973, pp.197-207) and Devitt (1981, chs.5&7) to propose an alternative *causal source theory of reference*, according to which what a term refers to at a time t is the item that is the dominant causal source of the thoughts that speakers associate with the term at t. To clarify, this needn't be the item that the term's initial defining event can be traced back to, nor the one that best satisfies the descriptive content of the associated thoughts. Rather, it's the item to which (or to whose tokens) speakers predominantly apply the term in their thinking and communication at t, and thus the item from which their associated thoughts are causally derived. Hence, the theory shifts the focus from the term's historical origin to how its contemporary usage is causally linked to the world. This allows it to accommodate the possibility of reference change by holding that a term changes its reference whenever there is a shift in what things its users predominantly apply it to.

Accordingly, to avoid the implausible conclusion that thick concepts cannot change their reference over time, causal theorists must interpret them as governed by the causal source theory rather than the causal-historical theory. However, this move may invite a different challenge, analogous to the Moral Twin Earth problem raised by Horgan and Timmons (1991, 1992) for the causal theory of reference as applied to thin ethical concepts.[84] Suppose there are two causally isolated linguistic communities: one on Earth, the other on Twin Earth. Earthlings and Twin Earthlings are similar in most respects but differ in how they predominantly use the term

[84] More specifically, Horgan and Timmons (1991, p.453; 1992, pp.158-159) raise this challenge against Boyd's realist moral naturalism (1988), which is based on his causal source theory of reference for moral terms.

‘distributively just’. Earthlings, being egalitarians, predominantly apply the term to equal distributions, while Twin Earthlings, as neo-liberal meritocrats, predominantly apply it to distributions based on desert. Thus, the dominant causal source of Earthlings’ use of ‘distributively just’ is egalitarian distribution, while for Twin Earthlings it is desert-based distribution. If thick concepts are governed by the causal source theory, it follows that ‘distributively just’ refers to different things in the two communities. But this might be seen as problematic, as it counterintuitively suggests Earthlings and Twin Earthlings aren’t substantively disagreeing about what makes distributions just, but are merely talking past each other.

Consequently, it might be argued that if we aim to explain the metasemantics of thick concepts in causal terms, we face an unpleasant dilemma: either we adopt the causal-historical theory, in which case we cannot accommodate the plausible possibility of thick concepts changing their reference over time; or we adopt the causal source theory, in which case we are forced to predict — counterintuitively — that communities who predominantly apply thick concepts to different kinds aren’t engaged in genuine moral disagreement. Descriptivists about thick concepts might argue that each horn of this dilemma presents a more serious problem than the inability of their own theory to account for large-scale errors about what thick concepts refer to in certain scenarios.

In reply, I agree that a causal metasemantics for thick concepts can only get off the ground if it addresses the presented dilemma. However, I believe that both horns of the dilemma can be adequately addressed. Let’s begin with the first horn. It seems that all plausible cases of reference change involving thick concepts can be accounted for without appealing to the causal source theory. This can be done by tweaking the causal-historical theory as follows: what an expression means and refers to is either the object to which its initial defining event was directed or the object to which the most recent redefining event in the chain was directed, depending on whether the causal-historical chain of communication involving the expression includes such a redefining event. Importantly, the redefining event needn’t be explicit and instantaneous; it may instead consist in a community’s gradual and implicit rejection of the original defining description. For example, the reference of ‘distributively just’ in our community would change if we ceased to associate the term with UDJ, thereby no longer intending it to refer to a good-making property characteristic of distributions.

Admittedly, unlike the causal source theory, the revised causal-historical theory doesn't allow a thick concept's reference to change merely because users adopt a new usage — unless this shift reflects a collective rejection of the original definition. But this is no flaw, since reference change in such cases lacks intuitive plausibility. For example, consider a community that predominantly applies 'distributively just' to a specific kind of distribution, making it the dominant causal source of their thoughts. Suppose they do so only because they mistakenly believe such distributions are morally good. They still accept UDJ and would stop using the term this way if they discovered otherwise. In such a case, it's more plausible to say that the community is misusing the term than that they have changed its reference. Far from undermining the revised causal-historical theory, this example supports it by illustrating why it should resist reference change in such scenarios

The second horn of the dilemma also seems surmountable. Even if the causal source theory governs thick concepts, it's premature to conclude that there is no substantive disagreement in the moral Twin Earth scenario. That would follow only if such disagreement required co-reference — a claim Plunkett and Sundell (2013) challenge. On their view, speakers can disagree metalinguistically over how an expression should be used to fulfil a shared functional role even if they refer to different things.[85] It's natural to interpret the Earthlings and Twin Earthlings in the above scenario as divided by a metalinguistic disagreement. That is, both groups associate 'distributively just' with UDJ, expecting the expression to fulfil the role of tracking a good-making kind characteristic of distributions. However, they hold different beliefs about what makes distributions morally good, which translates into a substantive metalinguistic disagreement about what kind 'distributively just' should be applied to in order to fulfil this role. This reading is available under both the causal-historical and causal source theories of 'distributively just'. Importantly, under the latter theory, it breaks the alleged link between co-reference and substantive disagreement on which the second horn of the dilemma relies. Therefore, the causal metasemantics of thick concepts withstands the dilemma.

[85] Notably, Plunkett and Sundell (2013, pp.5-6; 19-22) raise this point specifically in their critique of Horgan and Timmons' original Moral Twin Earth argument.

4.5.2. Concern 2: The Explanatory Role of the Underspecified Analysis

My second concern about Elstein and Hurka's counterargument goes one step deeper. Even assuming that, contrary to my argument above, descriptivism is the correct metasemantic theory about thick concepts, Elstein and Hurka still fail to refute the Disentangling Argument under its strongest possible interpretation. This is the interpretation under which the gist of Premise 1 isn't that separabilism cannot provide a reductivist analysis of thick concepts that allows their extension to be determined also by their evaluative component. Its gist is instead that separabilism cannot provide such an analysis while still fulfilling its core theoretical motivation: to establish that there is nothing distinctive about the evaluative component of thick concepts because it's reducible to the evaluative component of thin concepts.

The reason why I wish to question whether analysing thick concepts via the Underspecified Pattern fulfils its reductivist promise under descriptivism is this: even if descriptivism is true and underspecified descriptions capture thick concepts' semantic content, they do so only due to users' contingent epistemic limits. That is, speakers associate these descriptions with thick concepts not because they best explain what falls into their extension, but because they offer the most epistemically accessible explanation. Yet, to show that there is nothing distinctive about thick concepts' evaluative component, one must show that even the best explanation of their extension is exhausted by descriptive and thin evaluative concepts.

To see this point, consider once again the underspecified description UDJ associated with 'distributively just'. Elstein and Hurka themselves argue that there is a unique property of distributive justice that UDJ enables us to indirectly single out, even when we don't know how to specify X, Y, and Z (2009, p.523). Assume, for the sake of argument, that Rawls's theory of justice (1999) is correct, and thus that the real definition of distributive justice — reached by correctly specifying X, Y, and Z in UDJ — is RDJ.

> *RDJ*: To be a distribution that grants each person the most extensive scheme of basic rights and liberties compatible with the same scheme for everyone, and where social and economic inequalities are arranged under conditions of fair equality of opportunity, and so that they are to the greatest benefit of the least advantaged members of society.

Given the above assumption, RDJ is necessarily coextensive with UDJ. Nonetheless, they are hyperintensionally distinct characterisations of distributive justice, each with different informational content. RDJ provides users with more specific guidance on what it is to be distributively just, offering clearer cues about which distributions to look for. This enables RDJ-guided users to apply the term more consistently and systematically, unlike UDJ-guided users, whose applications vary depending on their interpretation of the good-making properties X, Y, and Z. As a result, RDJ-guided users are more reliable in anticipating correct uses and withholdings of the term, and thus more accurately identify its extension. They therefore qualify as more competent users of 'distributively just'. This suggests RDJ is a better explanation of the expression's extension than UDJ. Even if, under descriptivism, RDJ isn't the definition of the referent of 'distributively just', it still seems to be the real definition of the moral category the expression's users ultimately seek to track.

Notice, however, that RDJ incorporates several thick concepts: 'basic rights', 'liberties', 'fair', and 'benefit'. Thus, while users who associate 'distributively just' only with a generic description like UDJ may qualify as minimally competent users of the expression, *full* competence requires more than grasping its descriptive content and thin evaluation; it requires engagement with the relevant thick concepts. More generally, although underspecified analyses like UDJ help explain thick concepts' extension at a basic level, they may still fail to capture evaluative similarity relations that characterise their extension. It's in this important sense that, contrary to the reductivist rationale behind separabilism, the evaluative component of thick concepts like UDJ may not be fully derivative of thin evaluation, even under descriptivism.

One might attempt to reject the hypothesis that there are better explanations for the extension of thick concepts than those offered by underspecified descriptions. But how could such a rejection be justified? It seems to me that doing so would require a compelling reason to believe there is no intelligible answer to what X, Y, and Z in these underspecified descriptions are. To make this belief plausible, one would likely have to adopt a radical form of *moral particularism* concerning the explanation of the extension of thick concepts — the view that there are no generalisable truths about what makes the objects falling under thick concepts merit global thin evaluation.

For a moral particularist, whatever specific properties we would plug in for X, Y, and Z, for example, in UDJ, these properties wouldn't generally make distributions good. At most, they would make *some* distributions good in *some* circumstances. Consequently, the only way to correctly specify X, Y, and Z in UDJ would be through an immensely lengthy disjunctive description, whose disjuncts would survey various features that can make distributions morally good, along with a variety of particular circumstances in which they are good-making. Such a disjunction wouldn't only be exceedingly difficult, if not impossible, to cognitively entertain, but it would also fail to exhibit any intelligible explanatory pattern that would allow us to understand what makes distributions just. Hence, it would hardly give us a better explanation of what it is to be distributively just than UDJ does.

However, Elstein and Hurka don't want to commit their argument to moral particularism, and thus they allow that there is a generalisable way to correctly specify X, Y, and Z in underspecified descriptions associated with thick concepts, even if it isn't epistemically accessible to an entire community of their users (2009, p.523).[86] I consider this to be a reasonable decision on their part because moral particularism strikes me as a position that could be well-motivated about thick concepts only if one insisted the correct specification of X, Y, and Z must be *purely* descriptive. As discussed earlier, many thick concepts tend to apply to so many diverse things that, for all we know, their extension doesn't seem to be unified under *purely* descriptive similarity relations. This is what, after all, motivates Premise 2, which Elstein and Hurka themselves accept.

It seems unfounded to limit the possibilities of specifying X, Y, Z to purely descriptive terms, considering that, for many thick concepts, serious candidates for an evaluative specification of X, Y, Z seem available. To illustrate, consider two main candidates for filling out X, Y, Z in UDJ, which Elstein and Hurka mention (2009, p.521): the desert-based specification that characterises just distributions as those in which goods are awarded proportionally to one's merit, and the egalitarian specification that characterises just distributions as those in which goods are allocated equally. The first specification is clearly partially evaluative because 'merit' is an evaluative concept *par excellence*. And even while the second specification is purely descriptive,

[86] The most explicit defence of moral particularism regarding thick concepts can be found in Burton (1992, pp.28-32), who advances this position to argue that the semantic content of thick concepts is partially indeterminate (see also Tappolet, 2004, for a structurally similar account in which thin concepts are replaced by affective concepts). Elstein and Hurka (2009, p.523) distance themselves from this view.

this seems to be so only because it's in fact incomplete. After all, no theorist of justice would think the distribution of just any goods can be assessed along the dimension of justice; rather, only the distribution of those goods that are valuable for the quality of one's well-being and life opportunities — and possibly meet some further condition — can be so assessed. Therefore, both of the above specifications can be extensionally adequate only if they further specify the relevant goods — something that seems impossible without invoking an evaluative concept (cf. Kirchin 2017, pp.140-141, who makes a similar observation).

The above example was just one illustration. However, I encourage the reader to try to identify a thick concept whose extension clearly resists not only purely descriptive explanations but also any partially evaluative explanation other than an underspecified one involving global thin evaluation. I struggle to think of such cases and suspect that, even if some exist, they are unrepresentative. In fact, many evaluative explanations of thick concept extensions seem plausible precisely because we can often generalise about whether the relevant features are good-making or bad-making. This disconfirms moral particularism about thick concepts. Returning to our earlier example, one plausible partially evaluative explanation of the extension of 'cruel' is that something counts as cruel iff it involves intentionally inflicting disproportionate harm upon others. One reason it's plausible is that it's hard to imagine something involves intentionally inflicting disproportionate harm without thereby being, at least *pro tanto*, bad or wrong, and the explanation reflects that.

Additionally, the lack of a unified explanation for why instances of a thick concept merit global thin evaluation doesn't by itself show that their extension is best explained by an underspecified description involving global thin evaluation. For example, distributive justice could be a multiply realisable property, with different ways a distribution might be good depending on societal circumstances. In such a case, the correct specification of X, Y, Z in UDJ might have a disjunctive form. However, even such a specification can still provide valuable information that helps explain the extension of 'distributively just', so long as the selection and variation mechanisms behind the alternative explanations are non-arbitrary and intelligible to its users.[87]

[87] If a non-arbitrary mechanism governs variation among different specifications, this opens the theoretical possibility of interpreting 'distributively just' as a context-sensitive term whose extension varies across circumstances. See Väyrynen (2013, pp.169-178) for discussion of this view.

Hence, finding a compelling example of a thick concept whose extension is best explained by an underspecified description is more challenging than it might initially appear. Merely demonstrating that the alternatives to such a description exhibit some degree of complexity and disunity is insufficient. One would also need to show that this complexity and disunity preclude any intelligible explanatory pattern.

For these reasons, I find it justified to doubt that descriptions derived from the Underspecified Pattern better explain the extensions of thick concepts than their thicker evaluative alternatives. Accordingly, my second concern with Elstein and Hurka's counterargument holds: there is good reason to think the moral categories speakers aim to track with many thick concepts may be defined not by underspecified descriptions, but in a more informative, thicker way. Admittedly, this doesn't mean the Disentangling Argument prevails over separabilists, as Elstein and Hurka offer another reductivist analysis of thick concepts alongside the Underspecified Pattern (2009, pp.526-531).

Their second analysis is compatible with my above claim that a plausible-looking partially evaluative specification of X, Y, Z seems available for many thick concepts. The key idea behind the analysis is, however, that the evaluative component embedded in this specification can be thin. Elstein and Hurka call this thin evaluative component 'embedded evaluation' to differentiate it from global thin evaluation (2009, p.526). Unlike global thin evaluation, embedded evaluation doesn't govern the whole concept but only a specific part of it.[88] So, for example, they suggest that what groups together the acts in the extension of 'cruel' is that they 'involve causing some evil from desire or with indifference' (2009, p.528). This is an evaluative explanation, but it's nevertheless reductivist, since its evaluative component is expressed solely through the embedded thin concept 'evil' and thus involves no thick evaluation.

While the second analysis strikes me as more compelling than the first, I have doubts about its scope of application. I can think of several thick concepts for which it seems highly questionable whether a thin evaluative specification of X, Y, Z provides a better explanation of their extension

[88] Elstein and Hurka incorporate their first analysis in the second, arguing that the specification of X, Y, and Z — within which thin evaluation is embedded — still involves some residual underspecification, and therefore that global thin evaluation takes part in determining the concept's extension (2009, p.530). However, since this move causes their second analysis to inherit the flaws of the first, I set it aside here.

than a thicker alternative. For instance, Elstein and Hurka's reductivist analysis of 'cruel' seems too permissive, as 'cruel' cannot be applied to just any act that involves causing some evil from desire or with indifference; it applies only to acts involving the infliction of disproportionate harm — where 'disproportionate harm' is itself a thick description. Similarly, I fail to see how X, Y, Z in UDJ could be correctly specified solely with descriptive and thin concepts.

That said, the second analysis appears more promising when applied to certain thick concepts that refer to moral virtues, such as 'integrity', which Elstein and Hurka analyse as involving 'an agent's sticking to a significantly good goal despite temptations and distractions' (2009, p.528). Hence, it may be more appropriate to view thick concepts as a heterogeneous category — some of which are irreducibly thick, while others aren't — rather than attempting to subject all of them to a uniform analysis. I'm open to this approach. My objective here wasn't to argue against *any* defence of separabilism against the Disentangling Argument, but rather only against Elstein and Hurka's specific defence that appeals to the Underspecified Pattern.

4.6. An Alternative Diagnosis

So far, I have raised doubts about the Underspecified Pattern's potential to challenge Premise 1 of the Disentangling Argument. My argument centres on the idea that thick concepts may often group together objects not only on the basis of shared properties defined by underspecified descriptions, but also on the basis of properties whose real definitions are evaluatively thicker and better suited to perform the theoretical work of explaining — and possibly determining — their extensions. The evaluative component of such thick concepts cannot be reduced to global thin evaluation. That said, I don't mean to suggest that we should discard global thin evaluation as theoretically useless for analysing these thick concepts. On the contrary, I now want to show that global thin evaluation is generally useful for analysing the significance-explaining characterisations of the properties that the given thick concepts are used to track. To see this, notice that underspecified descriptions containing global thin evaluation often behave just as one would expect significance-explaining characterisations to behave.

Recall from Chapter 2 that a property's significance-explaining characterisation explains why a community deems the property worth conceptualising through an expression in their language. In doing so, it also specifies the representational component of the expression's function:

to represent the property that best matches the given characterisation. Accordingly, this characterisation is the description that the community's members collectively associate with the expression and that primarily motivates their use of it. Significance-explaining characterisations can thus serve as a shared touchstone that guides speakers in resolving disagreements over how to interpret an expression. In such cases, speakers rely on a significance-explaining characterisation by asking under which of its competing interpretations the expression refers to the property that best matches the characterisation — and is thereby worth conceptualising.

Underspecified descriptions containing global thin evaluation behave similarly whenever it's contested which property a thick concept functions to represent. As Elstein and Hurka themselves observe (2009, pp.521-522), these descriptions capture the common ground on which various sides of a disagreement over how thick concepts should be interpreted can agree. For instance, all minimally competent users of 'distributively just' agree that the expression serves to represent the property that meets the generic profile outlined by UDJ. Moreover, as already hinted at in the earlier discussion of the Moral Twin Earth scenario, when arguing about whether 'distributively just' should be interpreted in egalitarian, desert-based, or Rawlsian terms, the decisive consideration is under which of these interpretations, the expression picks out a good-making category characteristic of distributions, i.e., the property that best satisfies UDJ. This suggests UDJ captures what primarily motivates many — if not all — users of 'distributively just' to employ the expression: they use it to track the property that renders distributions morally good and to communicate moral commendation of its instances. Accordingly, to identify the representational component of the function of 'distributively just', we must appeal to UDJ.

To give one more supporting example of a thick concept whose functional specification incorporates global thin evaluation, consider the following underspecified analysis that Elstein and Hurka propose for the thick concept 'lewd' (2009, p.524):

> *UL*: To be wrong and have properties X, Y, and Z, whatever they are, that involve somehow passing beyond limits on sexual display and that make any act that has them wrong.

UL can be plausibly interpreted as a significance-explaining characterisation of the property 'lewd' functions to represent. What speaks in favour of this interpretation isn't only that UL captures what even speakers who disagree about how 'lewd' should be interpreted tend to agree on. Additionally,

this interpretation of UL helps explain why many speakers regard 'lewd' as an objectionable term they are reluctant to use. Let me now unpack this point.

A first-pass explanation of the objectionability of 'lewd' might be that speakers refrain from using it because they reject any limits on sexual display, believing the expression refers to a property that is never instantiated and whose extension, therefore, is necessarily empty. However, this explanation isn't satisfactory because the users of 'lewd' are often unwilling to use it in any sentence, including those where the term is embedded under operators that block the semantic entailment that someone is lewd, such as negation (Väyrynen, 2013, pp.60-61). That is, they are reluctant to utter not only sentences such as 'this behaviour is lewd' but also sentences such as 'this behaviour isn't lewd'. This suggests that the objectionable aspect of 'lewd' projects beyond its truth-conditional meaning. In that case, there must be an additional reason why some speakers are reluctant to use 'lewd'.

If we treat UL as capturing the representational component of the function of 'lewd', this additional reason comes to light: some speakers are reluctant to use 'lewd' because they perceive the term's very existence in their language as reflecting a problematic evaluative outlook within their linguistic community regarding what is worthy of conceptualisation. After all, their community considers the property that 'lewd' functions to represent worth conceptualising because its instances are morally bad in virtue of somehow exceeding the limits of sexual display. Those who object to 'lewd' may have various reasons to dispute whether this provides a sufficient justification for conceptualising the property under consideration.

As noted above, some of them might think that nothing is apt for moral evaluation on the basis of how much sexual display it involves. But this isn't the only possible reason for refraining from using 'lewd'. Some objectors might think otherwise and yet still hold that the property 'lewd' functions to represent shouldn't be conceptualised in their language. For instance, they might argue that even if some things are *pro tanto* morally wrong in virtue of involving too much sexual display, the degree of their wrongness is much smaller than many people take it to be. Accordingly, they might conclude that lewdness isn't a significant enough property to justify having a separate expression to represent it. Additionally, they may fear such an expression could reinforce oppressive norms regulating people's fashion and lifestyle choices or be prone to misapplication,

leading to the condemnation of various activities that aren't morally wrong. Hence, the objectors' underlying motivation for avoiding the use of 'lewd' is to prevent the perpetuation of the problematic view that lewdness is worth conceptualising. This view is reinforced by any use of the term, even in sentences that block the semantic entailment that someone or something is lewd.

To avoid confusion: although my two examples above involve underspecified descriptions, my point isn't that the significance-explaining characterisations of the properties thick concepts function to represent are always underspecified, but rather that they *generally* incorporate global thin evaluation. Some thick concepts may function to represent properties whose real definitions are widely recognised within a community and also explain why the community collectively regards these concepts as worth having in their language. The significance-explaining characterisations of such properties may not be underspecified.

For example, speakers of our language may generally treat 'cruel' as a valuable term insofar as it refers to the property of intentionally inflicting disproportionate harm. Yet it seems false to say we regard this property as worth conceptualising using 'cruel' independently of whether its instances merit negative global thin evaluation. This evaluation still figures in the property's significance-explaining characterisation. To illustrate, imagine speakers who correctly apply 'cruel' to acts involving disproportionate harm but mistakenly think it's a descriptive or even positive thick concept. Despite mastering its extension, they would fail to understand the function 'cruel' serves in our language — namely, to express moral condemnation of such acts, which is what primarily motivates our use of the term.

What's more, treating thick concepts as having the function of conveying global thin evaluation of what they are applied to — rather than viewing this as merely an accidental or secondary effect — is useful even for the purpose of criticising their objectionable aspects. To see this, consider Blackburn's example of the term 'cute', which is often objectionably applied to young women (1998, pp.101-105; 2013, p.124). According to Blackburn, to explain why this practice is problematic, we must disentangle the term's positive global thin evaluation from its other conveyed content and subject it to normative criticism. I'm sympathetic to this line of thought. In my own terms, this involves explaining why the evaluative function of 'cute' is problematic when the term is applied to women. However, in line with Chappell (2013, pp.123-124), I disagree

with Blackburn's claim (2013a, p.123) that thinking in terms of thick concepts does a disservice to ethics by discouraging critique. In fact, as Chappell (2013, p.194) points out, Blackburn's own critique of 'cute' relies on thick concepts such as 'infantile', 'subservient', and 'patronized':

> 'We may want to say that there is something wrong with them, along the lines of this: they admire and respond excitedly . . . to the non-threatening, infantile, subservient self-presentations that some women consciously or unconsciously adopt. Theirs is a group amongst whom women are successful by presenting themselves as there to be patronized like pets or babies (frequent terms of endearment). And that, we say, is bad.' (Blackburn, 2013a, p.123)

As I see it, the criticism of 'cute' that Blackburn points to is most compelling when framed as follows: the term functions to appraise what it's applied to for behaving in a non-threatening, infantile, and subservient way. This isn't merely an accidental or secondary effect; it's what primarily motivates speakers to use the term. Consequently, applying 'cute' to women perpetuates a patronising evaluative outlook in which women are admired for displaying such behaviour. Therefore, we should avoid applying 'cute' to women. This demonstrates why, in critically assessing a thick concept and its use, it's crucial to recognise that conveying global thin evaluation is part of its function — even if its functional specification also features thick concepts.

Two final clarifications are in order. First, my claim that thick concepts can often be analysed as having the function of tracking a property that merits a global thin evaluation and of communicating this evaluation is consistent with what Kirchin (2017) advances as the 'liberal view of thick concepts'. This view holds that some concepts can be categorised as evaluatively thick, even if, in some instances of their use, no negative or positive global evaluation is conveyed (Kirchin, 2017, pp.5, 111). To illustrate, some authors argue that utterances which suspend the global evaluation typically conveyed by a thick concept — such as 'whether or not Madonna's show is lewd, it's not bad in any way distinctive' — can be felicitously uttered in some situations (Väyrynen, 2013, pp.70-72; Bergström, 2002, p.5).[89]

[89] See also Willemsen and Reuter (2020), whose experimental studies indicate that suspending global evaluation is easier for positive thick concepts than for negative ones, possibly owing to social norms governing evaluative language.

M-Function allows us to treat 'lewd' as having a globally evaluative function despite such uses. After all, that an effect is central to speakers' motivation for using an expression — and thereby part of its function — simply means that their disposition to use the expression heavily depends on their interest, need, or desire to see the effect realised. This doesn't mean they must intend to produce the effect each time they use the term. As discussed in Chapter 2, speakers' disposition to use 'family' may heavily depend on their general interest in regulating the domestic lives of those they classify as such, even if they don't intend to regulate anyone's domestic life on each occasion. Similarly, speakers' disposition to use 'lewd' may heavily depend on their general interest in negatively evaluating acts involving excessive sexual display, even if they don't intend to convey such evaluation each time they call something 'lewd'.

Secondly, I don't claim that *all* thick concepts involve a clear positive or negative global thin evaluation. Some may be evaluatively flexible, regularly used to convey both positive and negative global evaluations. For instance, Kirchin (2017, p.130) argues that 'macabre' is sometimes used to praise what it's applied to, and at other times to criticise it. Such expressions might be best described as having the function of conveying some form of evaluation — whether positive or negative. Also, I don't rule out that some expressions may count as thick even if they convey only embedded, not global, evaluations. For example, it could be argued that 'friend' is a thick concept because, in applying the term to someone, a speaker evaluates that the person stands in a relationship of mutual trust and support with them — without necessarily evaluating the person as morally good because of that. Global evaluation may be absent in the function of such terms. Nothing I have said contradicts the possibility of such thick concepts; my point is only that the appeal to global thin evaluation can be useful for analysing *many* thick concepts.

4.7. Conclusion

The aim of this chapter was to show that the distinction between real definitions and significance-explaining characterisations of conceptualised properties — foregrounded by the motivation-based interpretation of conceptual function — can illuminate the debate between separabilism and inseparabilism about thick concepts. I argued that, for Elstein and Hurka's defence of separabilism against the Disentangling Argument to succeed, their proposed analysis of thick concepts via the Underspecified Pattern would need to capture the real definitions of the properties that thick concepts track. However, this doesn't seem to be the case. Still, I suggested that the true theoretical

value of Elstein and Hurka's analysis lies elsewhere: the appeal to global thin evaluation it incorporates is often useful for identifying significance-explaining characterisations of the properties that thick concepts are used to track, thereby helping us understand what primarily motivates us to use these concepts.

Part II– The Ethics of Conceptual Abandonment

Chapter 5. Referential Mismatch and Conceptual Abandonment

5.1. Introduction

It isn't uncommon among philosophers to argue that a certain lexical item should stop being used and, therefore, should be abandoned, either entirely or at least within a specific subdomain of language. Such conceptual abandonment proposals have been made about expressions such as 'race' (Appiah, 1996; Zack, 2002), 'fake news' and 'post-truth' (Habgood-Coote, 2019a, 2022), 'gender identity' (Gheaus, 2023), 'democracy' (Cappelen, 2023), 'echo chamber' and 'filter bubble' (Coady, 2024), 'conspiracy theory' (Shields, 2023), and 'concept' (Machery, 2009). These proposals point us towards a fruitful but underexplored question in conceptual ethics: when and why should we abandon a piece of terminology? This question is the focus of the next two chapters of this dissertation.

In this chapter, I wish to explore a specific argumentative strategy that has been invoked in support of some conceptual abandonment proposals. The strategy consists in pointing out that what an expression actually refers to — if anything — is very different from, and disappointing compared to what its users, on reflection, want it to refer to. Following Mallon (2006, p.533) and Cappelen (2023, pp.23-32), we can call this strategy the 'mismatch argument', as it identifies a mismatch between an expression's actual reference and its users' referential expectations of it. While such a mismatch may appear to be a critical defect justifying abandoning the expression, I argue that this needn't always be so.

Let me first make mismatch arguments more tangible for the reader. How can an expression's actual reference diverge from users' expectations? Cappelen and Mallon primarily associate mismatch arguments with what they call an 'extensional mismatch'. This mismatch arises when the things in an expression's actual extension don't match the profile of what speakers want to talk about when using it. In other words, its actual extension either overgenerates or undergenerates relative to speakers' expectations (Cappelen, 2023, p.30). Moreover, since an expression's extension is determined by its reference (i.e., its intension), an extensional mismatch is often a consequence of what Mallon (2006, p.533) calls an 'import mismatch' — a mismatch that arises when an expression's actual referent is an entity that isn't as significant as speakers envision it to be.

Perhaps the best-known example of a mismatch argument appealing to extensional and import mismatches is Appiah's (1996) argument for abandoning the term 'race'.[90] As previously noted in Chapter 1, Appiah begins his argument by examining the history of 'race'-talk among the intellectual and political elites of the Anglosphere, to whom he believes ordinary speakers have deferred on how to use 'race' (1996, pp.64-66). Using this method, he identifies two conceptions that have historically shaped what speakers expect 'race' to refer to. First, on the racialist conception of 'race', the term is supposed to pick out an essentialist kind realised by human groups sharing fundamental biological characteristics ('racial essences') that explain their distinctive moral, intellectual, and cultural traits (1996, pp.79-91). Second, on the biological conception of 'race', the term is supposed to refer to a taxonomic kind that divides the human population into subspecies (1996, pp.91-98).

Our current understanding of human biology reveals, however, that neither of these two kinds exists in reality. This, according to Appiah, leaves us with two possibilities regarding what 'race' refers to, depending on what metasemantics governs it (1996, pp.98-101). One possibility is that it refers to a putative category that merely purports to be a real human kind, even though no human group actually realises it. If so, 'race' is subject to an extensional mismatch because its extension is empty, contrary to its users' expectations. Alternatively, its extension may consist of some real-world entities that best causally explain 'race'-talk. Appiah suggests these might either be reproductively isolated local communities (e.g., the Amish) or discrete groups of people who share visible physical traits characteristic of major subcontinental regions. If so, 'race' is subject to both extensional and import mismatches because these groups don't align with ordinary racial classifications and aren't explanatorily or taxonomically significant in the envisioned way. Therefore, either way, 'race' fails to deliver on its referential promises, which makes Appiah conclude that it's a useless term that should be abandoned.[91]

In addition to extensional and import mismatches, there is another way in which speakers' referential expectations of an expression can go unfulfilled. Specifically, it can happen that while an expression's users want it to have a stable referent, no stable referent has been fixed for it, rendering it semantically indeterminate. In this chapter, I treat such cases as instances of referential

[90] Appiah's mismatch argument is also discussed by Mallon (2006, pp.528-534) and Cappelen (2023, pp.30-31).
[91] See also Zack (2002), who makes several points similar to Appiah's in her case for abandoning 'race'.

mismatch, as they similarly involve a significant divergence between what speakers expect from an expression's reference and what they actually get. We can call these cases 'indeterminacy mismatches'.

For an example of a mismatch argument that appeals to an indeterminacy mismatch, consider Habgood-Coote's argument that academics and journalists should stop using 'fake news' (2019a, 2022). His proposal can be interpreted as relying on the following mismatch argument (2019a, pp.1038-1041): We want 'fake news' to refer to some distinctive phenomenon. However, when we observe how 'fake news' is used in practice, we see it's used in multiple incompatible ways to talk about a host of different phenomena. Both ordinary and expert users of 'fake news' widely disagree about which of these alternative usages is correct. This makes it likely that none of them is. If so, 'fake news' involves a referential mismatch: even though we expect it to be a referentially settled expression, its reference has in fact never been successfully settled. Habgood-Coote thinks this alone shows that 'fake news' is a semantically defective expression that academics and journalists should refrain from using (2019a, p.1034). Additionally, he reinforces his abandonment case by arguing that, as a referentially unsettled expression, 'fake news' not only fails to add any useful descriptive resources to our language but is also vulnerable to exploitation for various demagogic purposes (Habgood-Coote, 2019a, pp.1040-1041, pp.1047-1054). In light of these considerations, he concludes that we should stop using the expression.

Now, that an expression has mismatched reference might seem like a strong reason to abandon it, as this basically means it fails to fulfil the referential purposes for which it's deployed by its users. Still, my objective is to defend a more moderate perspective on the argumentative force of referential mismatches in the debate about conceptual abandonment. I argue that abandoning an expression is an appropriate response to its suffering from a referential mismatch only under relatively demanding conditions, and thus mismatch arguments are significantly less potent than they might initially seem.

One clarification before we begin. Perhaps the most straightforward strategy for resisting a mismatch argument is to deny that the expression in question suffers from a referential mismatch at all. After all, all mismatch arguments depend on assumptions about what speakers want the expression to refer to, or what it actually refers to. If one can show that these assumptions are mistaken, and thus the expression's reference isn't mismatched, then the argument collapses.

While this strategy is certainly viable, it isn't the approach I take here. Instead, I aim to explore whether retaining an expression might still be justified *even if* it genuinely suffers from a referential mismatch. Accordingly, whenever I discuss a specific mismatch argument, I simply grant the initial diagnosis and focus on its implications.

The chapter is organised as follows. In §5.2, I argue that abandonment risks being an overly radical response to a referential mismatch unless it's clearly a more feasible strategy than semantically adapting the afflicted expression to fit users' referential expectations. In §5.3, I shift my focus to a type of referential mismatch that cannot be addressed through semantic adaptation because it arises from the afflicted expression being associated with categorically unfulfillable referential expectations. However, I demonstrate with the example of 'gender identity' that even an expression suffering from this type of mismatch can sometimes be useful enough to retain. Overall, I aim to show that, upon careful examination of all the relevant considerations surrounding a referential mismatch, it often becomes questionable how strongly it supports the afflicted expression's abandonment. In §5.4, I conclude by summarising how the insights presented in the chapter can be utilised to develop more compelling mismatch arguments.

5.2. Addressing Referential Mismatches Through Semantic Adaptation

When deciding whether to abandon an expression with mismatched reference, one natural consideration is whether it might not be better to retain it but try to adapt it so that its reference fits users' expectations. We can call this way of approaching referential mismatches the 'semantic adaptation strategy'. In this section, I explore whether this strategy may be a preferable alternative to the abandonment strategy.

5.2.1. The Possibility of Semantic Adaptation

For the semantic adaptation strategy to be feasible, speakers must be able to revise what expressions refer to. Whether this is possible — and, if so, what such a revision involves — depends on what factors determine reference. Various metasemantic theories offer different answers to this question. According to some of these theories, an expression's reference is determined by external factors detached from its present usage, such as historical facts about its initial dubbing and past transmission (e.g., Kripke, 1980; Putnam, 1975), or facts about which of its candidate referents are more metaphysically natural than others (e.g., Lewis, 1984). Since these

factors lie beyond the control of the expression's present users, if they fully determine reference, it's unclear how those users could change it.

At the same time, there are metasemantic theories under which revising an expression's reference appears at least theoretically possible by altering some aspects of its present usage. These are theories on which an expression's reference is determined by some factors related to how speakers are presently disposed to use it. According to the causal source theory of reference, these factors may include facts about which object is the dominant causal source of the expression's present usage, given what things speakers (or some relevant subset of them) tend to apply it to in their thought and speech (e.g., Evans, 1973; Devitt, 1981). Alternatively, according to the descriptivist theory of reference, these factors may include facts about which object best satisfies the descriptions speakers (or some relevant subset) tend to associate with the expression when using it (e.g., Jackson, 1998). Additionally, according to social externalism, facts about whom in their linguistic community, if anyone, speakers are disposed to defer to regarding how to use the expression may also be relevant (e.g., Burge, 1979; Putnam, 1975).

There is little consensus in the literature about which of these theories is correct. Notice, however, that mismatch arguments can get off the ground only if what expressions refer to somehow depends on how they are presently used, as suggested by the latter theories. This is because whether an expression fulfils a linguistic community's referential expectations seems relevant to its usefulness only insofar as its referent corresponds to what community members track — either descriptively or causally — when using it.[92]

To illustrate, imagine the following toy example: there is a linguistic community that, *pace* Craig's (1990) diagnosis discussed in Chapter 2, wants 'knowledge' to refer to *the belief kind whose instantiation is indicative of good informants*. Let's call this kind 'K'. The members of this community predominantly apply 'knowledge' to beliefs that instantiate K and tend to associate the expression with descriptions that accurately capture the defining conditions for being K. Yet, it turns out that 'knowledge' is governed by externalist metasemantic processes that prevent it from referring to K — for example, because K isn't the kind originally dubbed as 'knowledge' by its initial users, or the most metaphysically natural kind 'knowledge' might pick out. Consequently,

[92] A similar point is made by Riggs (2019), who argues that, insofar as meaning and reference are determined by externalist factors beyond speakers' control, they aren't what we aim to change when engineering expressions.

‘knowledge’ suffers from a referential mismatch. Nonetheless, this mismatch seems powerless as a reason for abandoning the expression. After all, even if what ‘knowledge’ in fact refers to is significantly less useful and more disappointing than K, this doesn’t in any way affect the community’s present linguistic practices with the expression. On the contrary, ‘knowledge’ is still an extremely useful expression for them, as it allows them to track K both descriptively and causally, and to reap all the information-pooling benefits of doing so.

This example demonstrates that mismatch arguments are dialectically committed to reference being closely tied to the present usage of expressions. Hence, when engaging with mismatch arguments, it’s theoretically appropriate to ask whether there are cases in which, even though an expression suffers from a referential mismatch, it’s possible to eliminate the mismatch by altering some aspect of its present usage.

To see what such cases might look like, imagine a linguistic community that, just like the previously considered community, wants ‘knowledge’ to refer to the kind K, whose instantiation is indicative of good informants. However, this community performs very poorly at both causally tracking K and descriptively identifying it. Its members, including its most authoritative speakers, are disposed to apply ‘knowledge’ to beliefs that aren’t K because they are false or luckily true, and hence K isn’t the dominant causal source of their usage of ‘knowledge’. Also, they associate ‘knowledge’ with descriptions that lead them astray in identifying K, such as ‘knowledge can be acquired by luck’ or ‘knowledge doesn’t have to be justified’. Hence, according to both the descriptivist and the causal source theory of reference, the community fails to use ‘knowledge’ so that it refers to K.

Assuming that either descriptivism or the causal source theory of reference governs ‘knowledge’ in the above example, the expression suffers from a referential mismatch in it. Moreover, it seems at least theoretically possible for the community to eliminate this mismatch by properly changing either what descriptions they associate with ‘knowledge’ or what they dominantly apply it to. Still, it seems difficult to assess whether this mismatch justifies abandoning ‘knowledge’ without knowing more details about the community’s future potential to realise the given possibility.

On the one hand, this potential could be very low. Imagine, for example, that making any progress in thinking about K and tracking it is nearly an impossible task for the community’s

members because doing so is just beyond their cognitive limits, or because they have been hopelessly brainwashed about what makes someone a good informant. In such a scenario, a case for abandonment is more appealing, for it seems plausible to think that operating with the expression 'knowledge' for such a community might be a waste of time. Its members should perhaps spend their time trying to learn how to navigate the world without pooling true information, rather than trying to learn how to recognise good informants.

On the other hand, we can also think of such elaboration of the scenario in which the community has a good prospect of revising the reference of 'knowledge' in accordance with their expectations. For example, imagine the community's failure to track K isn't caused by the kind being cognitively inaccessible to its members, or by its members being brainwashed, but rather by some corrigible epistemic error. If that's so, it seems too early to conclude that they should abandon 'knowledge'. After all, these speakers clearly have an interest in using 'knowledge' to refer to K, but they simply haven't yet figured out how to do so. Perhaps 'knowledge' is currently a useless expression for them. Still, as long as there is a realistic chance that they will transform it into a useful expression in the future by making progress in their attempts to track K, abandoning it would deprive them of the future opportunity to fulfil their referential expectations. Doing so seems as implausible as throwing away a bicycle belonging to a person who has a strong interest in cycling but hasn't yet learnt to ride it.

These two versions of the scenario show that when examining mismatch arguments, we must consider not only whether it's theoretically possible for a linguistic community to address a referential mismatch through semantic adaptation, but also whether doing so is a feasible option for them. To the extent that we can recognise that the community has a reasonable prospect of revising an expression's reference so that it no longer suffers from a referential mismatch, the mismatch is best interpreted as its temporary defect. In such cases, abandoning the expression seems to be a premature response. Therefore, to explore mismatch arguments in their strongest form, we must focus on the cases in which the prospect of addressing a referential mismatch through semantic adaptation is uncertain.

5.2.2 Is Abandoning Expressions Easier than Semantically Adapting Them?

Admittedly, it might often be practically difficult for a linguistic community to adapt an expression to their referential expectations. This holds true even if its reference is determined by some aspects of its present usage. Even in that case, reference shift requires that many members of the community change what they apply the expression to, or what descriptions they associate with it. There is ongoing debate in recent conceptual engineering literature regarding how difficult it's to implement such community-wide sociolinguistic changes. Some authors raise concerns that this might be practically unfeasible (Cappelen, 2018, pp.72-79; Deutsch, 2020; Jorem, 2021). Others offer various reasons for optimism. For instance, they suggest speakers can effect these changes through their collective long-large control over the relevant metasemantic processes (Koch, 2021b), through some revisions of social norms governing their behaviour (Nimtz, 2024a, 2024b), or through localised initiatives targeting smaller linguistic subcommunities (Matsui, 2024). What seems true, however, is that even if sociolinguistic changes enabling reference shift are implementable, their successful implementation is far from guaranteed — it's something we can hope for but not take for granted.

But here is an important point: if one seeks to justify abandoning an expression on the grounds that it suffers from a referential mismatch, it doesn't seem enough to provide some general reasons to think that addressing it through semantic adaptation might be difficult. In addition, she must also show that addressing it through abandonment is a more feasible option. Yet, this might often be questionable.

As an illustration, consider the referential mismatch that Habgood-Coote (2019a) identifies as affecting 'fake news'. As explained above, Habgood-Coote's concern is that while users of 'fake news' expect it to have a settled referent, it's in fact a referentially unsettled expression due to its usage being too varied and disparate. To eliminate this mismatch, users would need to collectively agree on a unified usage of 'fake news', thereby settling its reference. However, reaching such consensus appears significantly more challenging than settling the reference of a newly coined term, considering how widespread and entrenched the disagreements over how to use 'fake news' are. The situation is further complicated by these disagreements taking place within a highly divided political landscape, resulting in the expression being surrounded by heavy

ideological baggage that might be difficult to dissociate from it. Hence, the possibility that it might be very difficult for our linguistic community to semantically adapt 'fake news' should be taken seriously.

At the same time, once we consider what it would look like if academics and journalists followed Habgood-Coote's advice and stopped using 'fake news', it's unclear whether the abandonment strategy would be more likely to succeed in achieving its objectives than the semantic adaptation strategy. For one thing, Habgood-Coote (2022, p.490) frames abandoning 'fake news' as 'an act of political resistance' against its exploitation for various manipulative purposes, such as slurring mainstream media reports, eliciting various emotional responses, spreading bad ideology, or disparaging fact-based objections. Hence, the hope seems to be that, by ceasing to use 'fake news', academics and journalists will block its exploitation in public discourse. Yet, there is a salient risk that they will instead achieve the right opposite. Not only might the exploitative usage of 'fake news' not get blocked after the abandonment, but it might even become more prevalent. This is because demagogues might gain more opportunities to exploit the expression once they no longer face pressure from its non-exploitative usage. If that happens, the abandonment act will be better described as a surrender to exploitation rather than as resistance against it. To prevent this from happening, academics and journalists must not only stop using the expression themselves but also encourage the public to do the same, thereby challenging the expression's normalisation in public discourse. It's far from obvious that this would be an easier task than settling the expression's reference, as both options involve implementing large-scale sociolinguistic changes across our linguistic community.

Moreover, there are reasons to think that abandoning 'fake news' faces even further challenges. As Brown (2019, pp.146-147) points out, there seems to be a growing convergence among academics that 'fake news' should be interpreted as referring to some kind of mimicry of real news stories.[93] This naturally raises the question of whether our linguistic community couldn't fix one of these kinds or their disjunction as the referent of 'fake news'. In fact, Habgood-Coote (2022, pp.499-500) is himself sympathetic to the idea that real news mimicries are interesting phenomena we need to talk about. Still, he thinks we shouldn't aim to turn any of them into the

[93] See, e.g., Levy (2017), Rini (2017), and Gelfert (2018).

referent of ‘fake news’ — first, because doing so might be too difficult given the confusion surrounding the expression, and, second, because none of them deserves the disproportionate public attention it would receive as the referent of ‘fake news’. But in that case, the abandonment strategy doesn’t only require convincing the public to stop using ‘fake news’ but also ensuring that they will still be incentivised to talk about real news mimicries even without it. This translates into additional success conditions for the abandonment strategy.

To explain, one potential risk that the strategy must address is that those speakers who will be responsive to our calls for abandoning ‘fake news’ might easily misinterpret these calls as signalling that the phenomenon the expression is associated with — real news mimicries — isn’t worth discussing. Hence, we must spend extra effort to make the reasons for abandonment sufficiently clear to them. Relatedly, we must also equip the public with alternative conceptual resources that will enable them to discuss real news mimicries even after abandoning ‘fake news’. This might involve either introducing new terms into language or finding efficient ways to discuss them using existing terms. Additionally, as Pepp et al. (2022, p.484) warn, alternative conceptual resources for discussing ‘fake news’ also risk being exploited by demagogues. Therefore, we might also need to develop strategies to pre-empt such exploitation. None of these tasks seems like a small undertaking.

This example demonstrates that, when deciding how to address a referential mismatch, we cannot simply assume that the choice is between the straightforward abandonment strategy and the demanding semantic adaptation strategy. Even the abandonment strategy may encounter significant complications on the way to achieving its objectives. Moreover, the strategy might also be risky, as failing to resolve these complications might result in costs to our conceptual resources, or even leave them in a worse state than before. When such risks are at stake, it seems prudent to first make every effort to address the mismatch through semantic adaptation and resort to abandonment only if these efforts fail.[94]

[94] I refrain from drawing specific conclusions about the ‘fake news’ case based on this general advice. Whether our community’s best efforts to settle the reference of ‘fake news’ by promoting a reasonable interpretation of the term in public discourse have already failed is a complex empirical question that I cannot address here.

As we can see, the justificatory burden on those advocating the abandonment strategy in response to a referential mismatch is substantive: they must convince us not only that addressing the mismatch through semantic adaptation is difficult but also either that (1) addressing it through abandonment is the easier and safer option, or that (2) our best efforts to semantically adapt the expression have already been exhausted. Without such justification, abandonment risks being an overly radical response.

5.3. Categorical Mismatch: A Decisive Reason for Abandonment?

So far, I have only discussed the referential mismatches that are, at least theoretically speaking, fixable through semantic adaptation. Yet, there also seems to be referential mismatches that are clearly immune to semantic adaptation. What I have in mind are the mismatches that arise because of the afflicted expression being associated with referential expectations that have unsatisfiable content. In other words, the desiderata specifying what speakers expect the expression to refer to may be such that no entity could possibly satisfy them. This renders it *categorically impossible* for the expression to refer to an entity envisioned by these desiderata. I will call these mismatches 'categorical mismatches'.

An illustrative example of a categorical mismatch is the one Appiah (1996) identifies as affecting 'race'. As a reminder, Appiah contends that many speakers have expected 'race' to refer either to an explanatorily important essentialist human kind or to a taxonomically important human kind akin to subspecies. However, no such kinds exist, and thus there is nothing by referring to which 'race' can fulfil the given referential expectations.

In this section, I demonstrate that even categorical mismatches don't always justify abandoning the affected expressions. This claim may initially appear controversial. After all, it might be wondered what speakers might gain from using an expression that consistently falls short of satisfying their referential expectations for principled reasons. However, the devil lies in the details. The key point to recognise is that the effects of using an expression affected by a categorical mismatch may not, upon closer examination, be solely or predominantly negative. I will illustrate this point using the example of 'gender identity', as analysed by Anca Gheaus (2023) in her recent argument for abandoning the expression.

5.3.1. A Categorical Mismatch Affecting 'Gender Identity'

Gheaus' case for abandoning 'gender identity' has all the earmarks of an argument that advocates abandoning an expression on the basis that it's associated with categorically unfulfillable referential expectations. Gheaus first identifies five desiderata commonly associated with 'gender identity' by those who promote the expression in trans-debates (2023, pp.34-35). These desiderata are meant to capture what trans writers and activists expect the expression to do when promoting its usage in public discussions.[95] Whether 'gender identity' satisfies these desiderata depends on what it picks out. Thus, the desiderata can be plausibly interpreted as specifying what gender kind the users of 'gender identity' expect it to refer to. On this reading of Gheaus' argument, the five desiderata can be spelled out as follows:

> *Desideratum 1*: 'Gender identity' should refer to a gender kind that vindicates trans as well as non-trans people's beliefs about how they instantiate it.
>
> *Desideratum 2*: 'Gender identity' should refer to a gender kind such that people have privileged first-person access to their own ways of instantiating it.
>
> *Desideratum 3*: 'Gender identity' should refer to a gender kind such that everyone can self-identify with some way of instantiating it.
>
> *Desideratum 4*: 'Gender identity' should refer to a gender kind that sits well with the idea that misgendering is a serious harm, because people have a right to be treated according to their gender identity.[96]
>
> *Desideratum 5*: 'Gender identity' should refer to a gender kind that sits well with the idea that it's sometimes permissible for institutions to require information about one's gender identity.

[95] It's a separate question how accurately these desiderata capture their expectations. Nevertheless, I bracket this question here, as I'm interested in how compelling Gheaus' argument is, assuming its initial diagnosis of mismatch is correct.

[96] See Bettcher (2007), Kapusta (2016), and Jenkins (2016) for discussions of the harms associated with misgendering.

The first three desiderata can be called 'personal desiderata', for they concern what personal role the referent of 'gender identity' plays for people. By comparison, the last two desiderata can be called 'normative desiderata' because they concern how its referent *should* be treated in public life.

Next, Gheaus explores six candidate referents for 'gender identity' that have been suggested in the literature (2023, pp.39-46). She argues, though, that none of them satisfies more than three of the five desiderata. Gheaus' discussion suggests this is largely due to a tension between personal and normative desiderata. On the one hand, 'gender identity' is supposed to refer to gender roles that carry some distinctive personal significance for individuals. On the other hand, these roles are supposed to deserve special recognition in public life. Yet, Gheaus (2023, pp.37-39) thinks gender roles can never deserve such recognition because they are constituted by what she calls 'sui generis gender norms' — norms that arbitrarily categorise people into gender roles based on their sexual characteristics and assess them on how well they perform them. Some examples of sui generis gender norms are the norm that women should be nurturing or that men should be great leaders. Feminists convincingly criticise these norms as illegitimate.

However, if 'gender identity' refers to a kind constituted by illegitimate norms, it's difficult to see why people have a right to be treated according to their gender identity or what could justify a requirement to disclose one's gender identity. Accordingly, Gheaus finds it unlikely that we can identify a candidate referent for 'gender identity' that satisfies both personal and normative desiderata (2023, pp.45-46). In that case, 'gender identity' suffers from a categorical mismatch, as it cannot fulfil its users' referential expectations. Since this renders the expression ill-suited for the role it's meant to play in public discourse, Gheaus proposes abandoning it.

5.3.2. Should 'Gender Identity' Be Abandoned?

For the sake of discussion, let's assume there is no kind satisfying all five of the above desiderata. While Gheaus infers from this that 'gender identity' isn't worth using, I consider this conclusion too radical for two reasons. Firstly, 'gender identity' can still be a useful expression even if its referent satisfies only personal desiderata, which seems possible. Secondly, even if normative desiderata remain unsatisfied, they still reflect understandable concerns of transgender people that shouldn't be ignored. The usage of 'gender identity' helps enhance our awareness of these concerns. Let me elaborate on the two reasons in turn.

For the first reason, consider Jenkins's interpretation of 'gender identity' as referring to a gender role someone has internalised, in the sense that they experience its characteristic norms as relevant to themselves (2018). Gheaus (2023, pp.43-44) argues that although such internalised gender roles meet all three personal desiderata, the illegitimacy of gender norms prevents them from meeting normative desiderata.[97] Yet, even if that's so, this doesn't necessarily make internalised gender roles unworthy of conceptualisation. After all, they reflect the reality of human inner life, as people *do* experience some gender norms as more relevant to themselves than others. Attending to these roles allows us to distinguish between trans and cis individuals and renders the inner life of both groups more intelligible. This suggests we need a conceptual resource, such as 'gender identity', which spotlights internalised gender roles in our thinking and communication.

In reply, it might be countered that even if we need to talk about internalised gender roles, we should do so without relying on 'gender identity'. The reasoning might proceed as follows: 'The normative expectations that what "gender identity" refers to deserves special public recognition have become very widespread among its users. As a result, the usage of "gender identity" serves to reproduce them. Given that these expectations are unjustified yet difficult to disentangle from "gender identity", it might be better to abandon the expression and discuss internalised gender roles in different terms.'

What this response fails to appreciate is that the given normative expectations don't arise out of nowhere; rather, they reflect the genuine concerns of transgender people about being publicly treated in ways that are insensitive to the gender roles with which they self-identify. These concerns existed even before 'gender identity' entered our repertoire; the expression only foregrounded them. If we abandoned 'gender identity', this wouldn't mean that these concerns disappeared, but only that we decided to ignore them. Moreover, even if the given concerns conflict with the feminist interest in eliminating sui generis gender norms, they still offer valuable insights

[97] Some theorists have questioned whether internalised gender roles, as defined by Jenkins, meet the third desideratum, arguing her definitions of a woman's gender identity and non-binary gender identities aren't sufficiently inclusive (Andler, 2017; Dembroff, 2020, p.9, respectively). I set these concerns aside, as Jenkins' definitions could likely be tweaked to address them, and her account is only illustrative here.

into transgender people's experiences and are understandable on a fundamental human level. Therefore, they deserve our attention, at the very least.[98]

This leads us to the second reason why 'gender identity' is a useful expression: the talk of 'gender identity' raises public awareness of transgender people's concerns. This is an important contribution, if only because acknowledging these concerns marks a crucial first step towards fostering respect and empathy for transgender individuals. Moreover, it's only when we acknowledge these concerns that we can start thinking about how to at least partially reconcile them with our criticism of sui generis gender norms.

To explain, suppose that transgender people don't have a right to be treated according to their self-identified gender roles due to their illegitimacy. Still, as Gheaus (2023, p.43) herself admits, they may have a special claim against being subjected to the gender role with which they most disidentify. Alternatively, suppose that it's indeed impermissible for institutions to require information about people's self-identified gender roles. Nonetheless, encouraging people to voluntarily disclose this information for the purpose of making public spaces more inclusive might still be permissible. It seems fruitful to engage in such reflections on how transgender people's concerns can be at least partially addressed in a legitimate manner. Doing so not only encourages a more charitable interpretation of these concerns but also helps us envision how far we can progress towards a society that embraces both trans-inclusive and feminist values.

Accordingly, it would be a mistake to view the fact that 'gender identity' is associated with normative desiderata for its referent as an entirely negative feature. If we abandoned 'gender identity' and instead discussed internalised gender roles using different terms, transgender individuals' concerns about their treatment in public life would likely become harder to articulate and easier to ignore in discourse. This wouldn't only disrespect transgender people but also present a missed opportunity to reflect on how their concerns might be at least partially addressed.

My opponent may argue that the identified benefit of using 'gender identity' is outweighed by the harms its usage causes by perpetuating sui generis gender norms. However, it's unclear to

[98] Cf. Queloz (2024a; 2025, chs.5-6), who similarly argues that concepts reflecting tensions between human concerns can be preferable to tidier but artificially tension-free ones.

me whether this is indeed the case. I agree that the usage of 'gender identity' legitimises sui generis gender norms to some extent by reproducing the expectation that once gender roles constituted by these norms are internalised, they deserve public recognition. For this reason, there might perhaps be no place for 'gender identity' in ideal societies where sui generis gender norms have either disappeared or are on the brink of disappearance.[99] Nevertheless, in less-than-ideal societies such as ours, sui generis gender norms remain deeply entrenched, and it seems unlikely that abandoning 'gender identity' would significantly change this situation. Arguably, even if we stopped using 'gender identity', our inclination to follow these norms in our behaviour and treatment of others would still remain strong. Moreover, the application of 'gender identity' to transgender people might even help subvert the objectionable double standard embodied by sui generis gender norms. It might do so by making salient the fact that some people prefer to be treated according to gender norms different from those society imposed on them based on their birth-assigned sex characteristics.

It's for the above reasons that I remain unconvinced that the overall balance of the positives and negatives of using 'gender identity' supports its abandonment. This illustrates how an expression associated with categorically unfulfillable referential expectations might still be worth using.

5.3.3. A Comparison with 'Race'

Of course, there might be other cases where a categorical mismatch presents a stronger reason for abandoning the afflicted expression than in the case of 'gender identity'. Consider, for example, a referential mismatch that Appiah identifies in relation to 'race'. 'Race' is categorically incapable of fulfilling, *even partially*, the racialist expectations that, according to Appiah's diagnosis, speakers have predominantly associated with the term throughout its history.[100] While there may be various ways to categorise the human population, none of these categorisations would even

[99] This is merely a tentative possibility. A relevant consideration here is whether such a genderless society could avoid the oppression of transgender people. See Cull (2019) and Weltman (2024), who disagree on this question.

[100] Appiah's diagnosis isn't without challenge. See Taylor (2000) and Glasgow (2009, pp.38-58), who respectively criticise the method through which Appiah arrives at it for disregarding non-racialist perspectives in the history of racial discourse and failing to track the folk concept of 'race'. While certainly important, I must once again set such problems aside here, since my focus is on whether to abandon 'race' under the assumption that Appiah's diagnosis is correct.

approximately track putative human groups distinguished by racial essences, as envisioned by racialism. Furthermore, racialist expectations don't reflect any understandable concerns worthy of consideration. Instead, they are rooted in pernicious racist attitudes, the proliferation of which we should strive to curb by all possible means. Insofar as the usage of 'race' involves the risk of promoting these beliefs, this seems to be a strong reason for its abandonment.

Admittedly, even strong reasons are defeasible. Various authors have argued that, under certain interpretations, 'race' can be a useful enough term to retain in our repertoire. For example, Haslanger (2000) defends an interpretation of 'race' that ties one's racial identity to their position within the social hierarchy, arguing that this interpretation is useful for critical theorising about racial oppression. Kitcher (2007) and Hardimon (2017, pp.150-169) argue that there are non-essentialist biological and social interpretations of 'race' that make it a useful term in medical research for studying and remedying health inequalities and medically relevant differences in the population caused by racial discrimination. Alternatively, Jeffers (2019) argues that 'race' is useful for talking about culturally significant collective identities that racially oppressed groups have developed.

Yet, it's questionable whether we should keep using 'race' for the aforementioned benefits, insofar as its usage produces them only under certain local interpretations, while its dominant interpretation in our linguistic community remains racialist. This concern is especially pressing, considering how profoundly harmful the racialist interpretation of 'race' has historically proven to be (cf. Marques, 2020, pp.269-270). Therefore, the potential benefits of using 'race' provide strong countervailing reasons against abandoning the term only if the non-racialist interpretations of 'race' don't remain confined to specific subcommunities but instead gain enough traction within the broader linguistic community to displace the racialist referential expectations.

Accordingly, we should perhaps strive to promote the non-racialist interpretations of 'race' as widely as possible within our linguistic community. However, if our efforts prove unsuccessful, we should also be prepared to adopt eliminativist responses. This could involve discouraging the broader linguistic community from using 'race' altogether, and reserving its non-racialist usage only for clearly delineated local purposes, such as critical theorising about racism, medical

research, or expressing intra-group solidarity.[101] Alternatively, if even these local uses become frequently conflated with racialist uses by ordinary speakers, we might also consider abandoning 'race' entirely and replacing it with an alternative term that has a clearer non-racialist interpretation, such as 'racialised group' (cf. Blum, 2010; Hochman, 2017).

Which of the above responses towards 'race' and its translations is most appropriate to adopt is likely to vary on a case-by-case basis, depending on how deeply these terms are tainted by racialist associations in different linguistic communities. My concern here, however, lies elsewhere and is more moderate. What I have tried to show is that, although both Appiah and Gheaus invoke a categorical mismatch in their respective cases for abandonment, there is a noteworthy difference between them: the racialist referential expectations that, according to Appiah, surround 'race' seem significantly more problematic than the mutually inconsistent expectations that, according to Gheaus, surround 'gender identity'. As a result, Appiah's case provides a more serious consideration for abandoning 'race' than Gheaus' case does for abandoning 'gender identity'.

The lesson to be drawn here is that a categorical mismatch cannot be just assumed to always provide a decisive reason for abandoning the afflicted expression. There might be a trade-off between the negatives and positives of using an expression associated with categorically unfulfillable referential expectations, with the latter potentially outweighing the former.

5.4. Conclusion

I have argued that developing a compelling mismatch argument for abandoning an expression isn't easy. However, to be clear, the takeaway of this chapter isn't that we should avoid employing these arguments altogether. Rather, my point is that we should employ them with moderation, carefully investigating relevant considerations regarding the targeted expression. I have identified several such considerations. We should first consider whether it's at least theoretically possible to fix the identified mismatch through semantic adaptation. If it is, we should then ask how difficult it would be to semantically adapt the afflicted expression compared to abandoning it, as well as whether our best attempt at semantic adaptation has already failed. If it isn't, the key consideration is

[101] See Wodak (2022) for discussion of this kind of moderate eliminativist response.

whether, despite being associated with categorically unfulfillable referential expectations, the afflicted expression can still serve a useful role for its users — such as raising awareness of important human concerns that shouldn't be ignored — or whether its usage merely promotes misconceptions that cause more harm than good. If the former, abandonment isn't necessary. If the latter, abandonment should be seriously considered, especially if separating the expression from the misconceptions in users' minds proves too difficult.

In sum, rather than an attack on mismatch arguments, the chapter is better interpreted as offering guidance on how to develop more nuanced versions of these arguments that take the above considerations into account.

Chapter 6. Why 'Democracy' Is Still a Word Worth Using

6.1. Introduction

This chapter is another contribution to the inquiry into the ethics of conceptual abandonment. While the previous chapter had a more general focus, addressing an argumentative strategy commonly employed in conceptual abandonment projects, this chapter examines a specific conceptual abandonment proposal. I focus on Herman Cappelen's (2023) recent argument for abandoning the words 'democracy' and 'democratic' (henceforth, D-words).

While D-words are deeply entrenched in political discourse, academic debates, and everyday conversations, Cappelen (2023) argues, however, that using them isn't a good practice to cultivate; rather, we should abandon these words altogether. To this end, Cappelen presents four considerations that he takes together to present a strong case for abandoning D-words (2023, pp.40-41, chs.7-8). These considerations are: (i) D-words fail us semantically because they are either significantly semantically unsettled or, if semantically settled, their meaning results in a massive mismatch between their actual extension and what speakers want to talk about when using them; (ii) the usage of D-words gives rise to pointless verbal disputes; (iii) D-words are often exploited in empty rhetoric for various bad purposes; and (iv) whatever we want to convey using D-words can be more effectively conveyed with alternative terminology.

In this chapter, I critically engage with Cappelen's argument, focusing primarily on what he presents as the most likely hypothesis for why D-words fail us semantically, as outlined in the first consideration. According to this hypothesis, which Cappelen (2023, p.108) calls 'the No-Meaning Hypothesis', while D-words give us the illusion of meaning, their reference — and thus their extension — is, in fact, unsettled. Hence, the hypothesis diagnoses D-words as being subject to *indeterminacy mismatch*. Against Cappelen, I argue for two claims. First, even if D-words aren't fully semantically settled, they are likely at least partially settled, and thus Cappelen overstates the extent of their semantic failure. Second, even if D-words are only partially semantically settled, they are still useful enough to retain in our conceptual repertoire.

The chapter is structured as follows. In §6.2, I explain Cappelen's reasoning behind the No-Meaning Hypothesis and why this hypothesis is important to his overall case for abandoning

D-words. In §6.3, I argue that even if D-words aren't fully semantically settled, speakers are likely to agree at least partially on how to interpret them. Namely, they are likely to agree that there is a close connection between how 'democratic' a decision-making system is and whether it enables the people to influence its outcomes through fair processes. In §6.4, I argue that, to the extent that this diagnosis is correct, the usage of D-words provides various substantive epistemic and practical benefits, which speaks in favour of retaining them in our conceptual repertoire. In §6.5, I address two foreseeable objections to my argument, exploring how the positive value of D-words holds up against Cappelen's broader case for abandoning them.

6.2. The No-Meaning Hypothesis

Cappelen (2023, p.108) formulates the No-Meaning Hypothesis as follows:

> *The No-Meaning Hypothesis*: The meaning-determining mechanisms (whatever they may be) failed to give meaning to 'democracy' and 'democratic'.

As should be clear from this formulation, Cappelen doesn't base his defence of the No-Meaning Hypothesis on any particular metasemantic theory concerning what meaning-determining mechanisms govern D-words. He thinks that such a *top-down* approach would be unsuitable here because there is a shortage of well-developed metasemantic frameworks for social and political terms such as 'democracy' (2023, pp.110-111). Instead, he defends the hypothesis through what he calls a *bottom-up* approach: he appeals to some general considerations about D-words that give us evidence that these words are significantly semantically unsettled, without building on a particular metasemantic theory.

The reasoning through which Cappelen motivates the No-Meaning Hypothesis can be summarised as follows: Cappelen first (2023, pp.77-79, p.111) observes that speakers can broadly agree on some paradigmatic examples of entities that are in the extension of D-words (e.g., Norway, Canada, the HKU Philosophy Department) as well as on those that clearly lie outside it (e.g., North Korea, Harvard, Tesla, *The New York Times*). It seems plausible to think that what makes the former entities but not the latter entities fall under D-words are some features that characterise how they make collective decisions. Cappelen, however, suspects that once we ask what exactly these features are, things get murkier. Specifically, he argues that the structure of collective decision-

making processes that entities must implement to count as 'democracy' and 'democratic', which he calls 'KQIE structure', is unsettled across the following four dimensions (2023, pp.107-108):

> *Kind*: What kind of collective decisions (and their stages) are relevant to whether a decision-making entity is in the extension of D-words?
>
> *Quantity*: How many decisions of a relevant kind must be made in a relevant way in order for a decision-making entity to be in the extension of D-words?
>
> *Input-Mechanism*: What do various input-mechanisms through which the entities in the extension of D-words make collective decisions have in common?
>
> *Embedding*: How must relevant input-mechanisms be embedded in a larger cultural, historical, and social setting for a decision-making entity to be in the extension of D-words?

Cappelen finds it very unlikely that there are any settled answers to these questions. That is, he doubts that our linguistic community has successfully converged on a particular configuration of the KQIE structure that is required for falling into the extension of D-words. His doubts begin with the following speculation: D-words may have originally entered the language to describe far simpler decisions than those that currently motivate our interest in them. If that's the case, there is a live possibility that the KQIE structure is unsettled. After all, it might well be the case that the semantics of D-words was never developed beyond the initial anchoring to the simple cases to encompass structurally complex decisions (2023, pp.111-112).

Cappelen then goes on to present other considerations about D-words that lend support to this possibility. For one thing, it's common for speakers to be descriptively indifferent about what D-words denote, instead using them primarily to evoke various associations, emotions, and memories in their audience (2023, pp.114-115). This tendency also leads to many speakers' wanting D-words to track whatever decision-making processes that they favour while widely disagreeing or lacking concrete views about their KQIE structure (2023, pp.115-117). In such circumstances, the KQIE structure is likely underdetermined by how D-words are used within our broader linguistic community.

Furthermore, Cappelen argues that we shouldn't expect much help in settling the KQIE structure from externalist metasemantic factors, such as deference to an expert consensus, as per

social externalism (Burge, 1979; Putnam, 1975), or considerations about which interpretations of D-words are more natural than others, as per reference magnetism (Lewis, 1984; Sider, 2011). Regarding the former option, Cappelen argues that not only is there disagreement among ordinary speakers about whom to defer to on how to use D-words, but even those individuals who might plausibly be considered experts on democracy — namely, political theorists — widely disagree about how to interpret these terms (pp.117-118). Regarding the latter option, all the eligible candidates for what D-words might refer to likely render the KQIE structure arbitrary. Consequently, considerations about which candidate is easier to refer to than the others, due to being more natural or joint-carving, don't seem applicable in this context (pp.118-119).

The above considerations summarise Cappelen's argument for the No-Meaning Hypothesis. While Cappelen regards the hypothesis as most likely true, he also acknowledges that his argument is inconclusive given our limited understanding of the nature of meaning (2023, p.106).[102] This leads him to also explore an alternative hypothesis: D-words are semantically settled enough to have a determinate KQIE structure, but their extension either significantly overgenerates or undergenerates (2023, pp.122-127). Nonetheless, the No-Meaning Hypothesis appears to play a significant role in the overall structure of Cappelen's argument. This is because, if D-words are semantically unsettled, they seem particularly susceptible to the other three problems Cappelen identifies as reasons for abandoning them, beyond their semantic defectiveness.

Consider, for example, Cappelen's concern that the usage of D-words is confusing because it's conducive to merely verbal disputes. These are disputes that arise not from substantive differences in people's beliefs but simply because they use D-words differently without noticing it. As Cappelen himself points out, even if D-words were semantically settled, this wouldn't make them immune to merely verbal disputes (2023, p.32, pp.127-128). Such disputes could still arise in virtue of speakers having diverse ideas (or 'speaker-meanings') in mind when using D-words without themselves realising it. That said, it seems plausible that if D-words are semantically unsettled, there is a greater potential for such disputes to arise. After all, the fact that an expression lacks a stable semantic foundation suggests that the linguistic community of its users hasn't developed consistent and widely accepted norms for its correct usage. In the absence of such norms,

[102] See McPherson (2025, pp.4-9) and Koch (2025, pp.11-14) for recent criticisms of Cappelen's argument for the No-Meaning Hypothesis.

the use of the expression becomes unregulated, making it easier for speakers to interpret it in various conflicting ways and ultimately talk past each other.

Furthermore, it's for similar reasons that the semantic instability of D-words exacerbates Cappelen's worry that they tend to be exploited by demagogues for various bad purposes. As noted earlier, Cappelen defends the No Meaning Hypothesis by making it 'at least a salient option that D-words are empty shells — deceptive pieces of terminological fluff exploited by speakers to trigger emotions' (2023, p.122). This suggests that the widespread exploitation of D-words among speakers might be what prevents them from acquiring stable meanings. However, it also seems plausible that the explanatory relation between the semantic instability of D-words and their exploitation is bidirectional, creating a vicious circle: not only is the exploitation of D-words an obstacle to their acquiring a stable meaning, but the lack of semantic norms regulating their usage also promotes their exploitation by creating conditions favourable to their arbitrary use in emotionally charged ways.

Finally, consider Cappelen's claim that there is nothing we use D-words to talk about that couldn't be conveyed more effectively without them (2023, pp.134-141). The No-Meaning Hypothesis makes this claim sound more compelling. After all, if D-words are semantically unsettled, their presence in the language might be naturally thought to be not only conducive to miscommunication and exploitation but also redundant, as they fail to provide any distinctive representational value to their users that would otherwise be unavailable to them. In that case, we only make our communication smoother by discussing the ideas associated with D-words in alternative terms — or so the argument goes.

As we can see, even if Cappelen's overall argument for abandoning D-words isn't dependent on the No-Meaning Hypothesis, the hypothesis still adds to its force by bolstering its underlying considerations. Still, in the next section, I argue that this hypothesis is overstated.

6.3. Possibilities and Limits of Agreement about D-words

For the No-Meaning Hypothesis to be compelling, it isn't enough to show, *pace* Cappelen, that D-words are surrounded by great variability in how speakers, including experts, interpret them. In addition, it must also be shown that speakers' interpretations of D-words are so irreconcilably divergent that they cannot reach *even a partial agreement on how to interpret them*. After all, if

such an agreement is possible, the potential utility of D-words might lie in that they are partially semantically settled expressions that help speakers track at least an approximate category that ties together their divergent interpretations. In this section, I argue that the possibility of such an agreement is underappreciated in Cappelen's discussion.

I agree with Cappelen that when theorising about D-words, we shouldn't only focus on the noun 'democracy' but also on the adjective 'democratic' (2023, pp.73-74). Accordingly, we should carefully consider Cappelen's observation that 'democratic' (2023, pp.76-77) is a gradable predicate. This observation suggests that what we generally interpret 'democratic' as referring to a property that can be realised to varying degrees. Also, it seems plausible that we generally interpret 'democracy' as referring to various decision-making social entities, such as states, governments, organisations, institutions, or platforms that realise this property to a high enough degree to instantiate it. We can then distinguish between three questions that one can ask about what 'democratic' refers to.

> *Relevance question*: What dimensions are relevant to the degree to which an entity realises the property that 'democratic' refers to?
>
> *Weighing question*: How do the relevant dimensions compare to each other in how much they contribute to the degree to which an entity realises the property that 'democratic' refers to?
>
> *Threshold question*: What are the conditions setting the minimum threshold that an entity must satisfy to instantiate the property that 'democratic' refers to, thereby counting as 'democracy'?

Now, it would be naïve to assume that our linguistic community has managed to agree on all the details of how to answer the three questions above, as widespread disagreements and indifference regarding the KQIE structure indeed disconfirm this. But this is consistent with the possibility that speakers can widely agree at least on *some aspects* of these questions. This possibility deserves more serious consideration than Cappelen affords it. While Cappelen (2023, pp.128-132, chs.10-11) provides examples of different ways in which theorists as well as laypeople interpret 'democracy', he doesn't examine whether there might be something that these

interpretations have in common. Yet, these interpretations don't appear entirely disunified — at least not in an obvious way. For instance, the following partial specification of what 'democratic' refers to strikes me as a reasonable candidate for a point of convergence among the various sides of the disagreement over how to interpret D-words:

> *D-specification*: The word 'democratic' refers to a property that decision-making systems — and the entities that implement them — realise, at least in part, to the extent that they enable the people to influence decision-making outcomes through fair processes.

While many aspects of D-words can be contested, the D-specification doesn't seem particularly controversial. For example, if we look at the entries for 'democracy' in many respectable English dictionaries, we find there such definitions as 'a system of government in which people choose their rulers by voting for them in elections' (HarperCollins Publishers, n.d.), 'government by the people; esp. a system of government in which all the people of a state or polity (or, esp. formerly, a subset of them meeting particular conditions) are involved in making decisions about its affairs…' (Oxford University Press, n.d.), 'a system of government in which power is held by elected representatives who are freely voted for by the people, or held directly by the people themselves' (Cambridge University Press, n.d.). These don't seem to be wildly different definitions of 'democracy', as all of them centre on the idea that democracy is about people-centred decision-making, i.e., decision-making that grants the people significant power to influence its outcomes. Of course, like all dictionary definitions, these definitions are somewhat half-baked; lacking theoretical nuance. Still, they tentatively indicate that the given idea holds significant currency in our linguistic community.

What's more, even when we consider some representative academic definitions of D-words that Cappelen discusses in his book, they don't sever the connection between democracy and people-centred decision-making. Some of them can be reasonably interpreted as expanding upon this connection by requiring, in line with the D-specification, that democratic decision-making must uphold some standards of fairness in the processes through which it enables the people to influence its outcomes.[103] The clearest example is Estlund (2008, p.38), who defines 'democracy'

[103] The fairness standards relevant to the D-specification concern *procedural fairness*, i.e., they assess the fairness of processes in people-centred decision-making systems. However, the D-specification doesn't imply that how democratic a people-centred decision-making system is depends solely on whether its outcomes are produced through

as 'the actual collective authorisation of laws and policies by the people subject to them'. As Cappelen (2023, pp.165-166) points out, Estlund (2008, p.66) holds that laws and policies are authorised through voting, and that voting has the normative power to do so in part because it's a *fair* procedure that provides people with an equal opportunity to participate in collective decision-making. Hence, Estlund's definition of 'democracy' accepts the D-specification, as it implies that the extent to which a decision-making system enables the people to influence its outcomes through fair processes partially determines how democratic it is.[104]

A similar point can also be extended to two other academic definitions of 'democracy' discussed by Cappelen, namely Christiano's and Dahl's definitions of 'democracy'. Christiano (2008, sect.1) defines 'democracy' as referring to 'a method of collective decision-making characterised by a kind of equality among the participants at an essential stage of the decision-making process'. Dahl (2006, pp.63-71), by contrast, formulates a very demanding definition of 'democracy' consisting of several conditions. What matters for our purposes is that, as one of the three main characteristics of democracy that these conditions are meant to capture, he mentions that 'in the process of choosing the alternative to be enforced as government policy, the preference of each member is assigned an equal value' (2006, p.64). That equality is important to Dahl's understanding of 'democracy' is also evident from his other work where he writes that 'a key characteristic of a democracy is the continuing responsiveness of the government to the preferences of its citizens, considered as political equals' (1971, p.1). Hence, both Christiano's and Dahl's definitions of 'democracy' centre on whether the people participating in a decision-making system are treated equally. Since this equality plausibly contributes (at least *pro tanto*) to the system's fairness, the given definitions seem to commit both authors to the D-specification.[105]

a fair procedure, irrespective of their content. It merely states that this depends at least in part on its procedural fairness, allowing for additional, procedure-independent constraints on the content of its outcomes. Moreover, a fair decision-making procedure can sometimes produce outcomes that undermine procedural fairness in future decisions. For example, a society might, through a fair decision-making process, adopt a discriminatory policy that harms minority groups' interests, thereby limiting their future participation and gradually eroding procedural fairness over time. Therefore, safeguarding procedural fairness plausibly requires attending not only to decision-making processes but also to the effects of their outcomes (cf. Dworkin, 2006, pp.144-145).

[104] For completeness, Estlund (2008) goes on to argue that the authoritativeness of voting doesn't only derive from its procedural fairness but also rests on its tending to make better decisions than random and better than all the generally acceptable alternatives.

[105] In this respect, Christiano's aim (2008, sect. 1) that his definition doesn't carry any normative weight seems overstated because equality *is* a normatively significant property. That said, the normative implications of his definition can still be weak enough for it to remain neutral on whether qualifying as democracy is desirable. This is so for two reasons. First, his definition doesn't specify what kind of equality, and thereby what degree of fairness, is

Admittedly, Cappelen also discusses approaches to conceptualising 'democracy' whose commitment to the D-specification is less straightforward than in the case of the three mentioned above. However, it seems to me that, upon closer examination, even these approaches turn out to be committed to the D-specification, as rejecting it would be too theoretically costly for them. The first approach to consider is Schumpeter's minimalist approach to defining 'democracy'. As Cappelen (2023, p.168) summarises Schumpeter's view, Schumpeter (1943/2003, pp.284-285) defines 'democracy' as referring to a political system in which the people have the opportunity of accepting or refusing the men who are to rule them, through free competition among would-be leaders for the vote of the electorate. This definition is minimalist because what characterises democracy, according to it, is simply free elections and the free competition preceding them, rather than further ends — such as equality or fairness — to which elections might be instrumental.

When one reads Schumpeter, it's easy to come away with the impression that he avoids any commitment whatsoever to the D-specification. This is because he explicitly writes that his definition of 'democracy' doesn't exclude the cases of 'unfair' or 'fraudulent' electoral competition, arguing that if it did, it would define 'a completely unrealistic ideal' (1943/2003, p.271). Yet, notice that this statement is compatible with the D-specification. This is because, while the D-Specification addresses the Relevance Question, Schumpeter's statement only addresses the Threshold Question. That is, he proposes that fairness in electoral competition shouldn't be a prerequisite for a system to be classified as 'democratic'. However, even then, fairness can still be a factor that contributes to the degree to which political systems are democratic, in line with the D-specification. Schumpeter would arguably have a hard time denying this, given that he writes, '…there is a continuous range of variation within which the democratic method of government shades off into the autocratic one by imperceptible steps' (1943/2003, p.271). This suggests he acknowledges that political systems classified as 'democratic' can be ordered on a continuum depending on how close they are to falling into the extension of 'autocratic'. But if that's so, whether a democratic system is based on fair electoral competition is presumably a factor that affects its position on this continuum. Otherwise, it would follow that democratic systems based

required for something to qualify as democracy. Additionally, even if the required degree of fairness is high, it's still consistent with the possibility that the fairness of a democratic decision-making system could be outweighed by its other negative features, rendering it undesirable.

on fair competition are just as close to becoming autocratic as those that aren't — other things being equal — a conclusion that sounds highly implausible.

Last but not least, consider the six conceptions of 'democracy' in political theory outlined by Coppedge et al. (2011), whose paper Cappelen (2023, pp.174-181) discusses in his book. These conceptions — electoral, liberal, majoritarian, deliberative, participatory, and egalitarian — are presented by Coppedge et al. (2011, p.253) as alternative interpretations of what 'rule by the people' means, such that the degree of democracy in a political unit can be measured in a multidimensional and disaggregated manner within each of them. What is important in the context of our discussion is that it seems highly uncharitable to interpret any of these conceptions as rejecting the D-specification. The egalitarian and electoral conceptions cannot reject it for the reasons that should already be familiar to us. Equal empowerment in people-centred decision-making, which lies at the heart of egalitarian definitions of democracy like those of Dahl and Christiano, is precisely aimed at enhancing fairness. And any electoral conception — including minimalist ones like Schumpeter's — must accept the D-specification to avoid the untenable conclusion that rigged elections are as democratic as those based on fair competition.

What about the other four conceptions? Prominent advocates of the deliberative conception of democracy clearly don't disregard fairness in people-centred decision-making, as they recognise the need to address social and economic inequalities to enable rational public deliberation (Habermas, 1998, p.308; Knight & Johnson, 1997, p.307). The participatory conception similarly holds that active citizen engagement reduces the risk of unfair outcomes driven by the parochial interests of a powerful few (Pateman, 1970; Barber, 1986). And, as Coppedge et al. (2011, p.253) characterise the majoritarian and liberal conceptions, their key tension lies here: the majoritarian conception favours centralised, party-driven governance that effectively responds to majority preferences and ensures electoral accountability, whereas the liberal conception promotes decentralised, pluralist systems that represent diverse groups and incorporate safeguards against domination.[106] This sounds more like a disagreement over which system promotes the people's interests more fairly, not over the importance of fairness itself. The latter interpretation would be

[106] The majoritarian conception originates with the works of Bagehot (1867/1963) and Wilson (1885/1956). For its overview, see Lijphart (1999, ch.2). The liberal conception can be traced back to the works of Mill (1865/1958) and Hamilton, Jay, and Madison (1787-1788/1961). For its overview, see Held (2006, ch.3).

highly uncharitable, as a conception disregarding fairness would clearly be inferior to its counterpart. Moreover, Coppedge et al. (2011, p.253) emphasise that both conceptions uphold values like civil liberties, due process, and transparency — all closely aligned with fairness. Therefore, insofar as all six conceptions are legitimate approaches to people-centred decision-making, none can afford to reject the D-specification. They are best interpreted as upholding fairness, albeit in different ways.

As we can see, the D-specification serves as a unifying thread across diverse academic definitions of D-words. But how widely do ordinary speakers associate D-words with fairness in people-centred decision-making? Fully answering this question would require empirical research beyond the chapter's scope. Still, it seems unlikely that people would generally describe a system as democratic regardless of how fair its decision-making processes seem to them. Even widespread exploitation of D-words doesn't suggest such a disconnect. For instance, many Trump supporters — despite being susceptible to manipulative uses of D-words — claimed the 2020 election was undemocratic due to alleged voter fraud (Cillizza, 2021; Chatelain, 2020). Similarly, regimes like Russia, Belarus, and Iran often rig elections yet still claim to be democracies and simulate fairness (Sauer, 2024; Karmanau, 2025; Golkar, 2024). If fairness is widely seen as irrelevant to their professed democratic status, why maintain this façade? These examples suggest the link between fairness and democracy is deeply rooted, even in discourses where D-words tend to be exploited.

Accordingly, we should take seriously the possibility that the D-specification captures a minimum common ground within our linguistic community on how to interpret D-words, thereby settling at least some semantic facts about them. What implications does this possibility have for the above three questions about what D-words refer to?

When it comes to the Relevance Question and the Weighing Question, the D-specification does seem to render them at least partially answerable. Regarding the Relevance Question, there are very plausibly facts about which features of decision-making systems are relevant to whether the processes through which these systems enable the people to influence decision-making outcomes are fair. For instance, these might include such things as whether the people have equal opportunities to participate in elections; whether they have access to diverse and reliable media to make informed decisions; whether they have freedom of speech and freedom of association,

ensuring that diverse political voices can be heard; whether there is a thriving independent civil society capable of voicing critical perspectives and holding those in power accountable; whether institutional decisions are periodically reviewed to reflect changes in people's preferences; or whether there are mechanisms in place to ensure government accountability to the public, such as transparent decision-making processes or regular elections.

The same reasoning also applies to the Weighing Question. There are very plausibly facts about how features like those listed above compare to one another in terms of their importance for fairness in people-centred decision-making — and, by extension, for democracy. Admittedly, even experts often disagree on how to weigh these features against one another. However, this is true for many other comparative moral questions as well. For example, experts in moral philosophy disagree over which distribution policies are more just, which courses of action in a moral dilemma are morally preferable, or which virtues contribute more to human flourishing. Yet, the widespread practice of asking such questions suggests a strong tendency to believe there are facts about the comparative relations that these questions aim to uncover. It seems *ad hoc* to treat the question of what matters for fairness in people-centred decision-making as exceptional in this regard.

To clarify, there might not be one generalisable answer to the Relevance Question and the Weighing Question. It's entirely possible that the dimensions relevant to achieving fairness in people-centred decision-making, as well as their relative importance, can vary depending on the nature of the decision-making entity — for example, whether it's a university, corporation, neighbourhood association, family, or political unit such as a state, municipality, or supranational union. However, this doesn't mean that the Relevance Question and the Weighing Question are unanswerable; it only means that they cannot be answered in a one-size-fits-all manner.

Nevertheless, it would be a stretch to think the D-specification imposes semantic constraints on D-words tight enough to fully determine their reference. For one thing, considerable variability may remain in how people interpret D-words beyond the D-specification, leaving the Relevance Question and the Weighing Question unsettled. Moreover, the D-specification doesn't address the Threshold Question. While it indicates that some degree of fairness in people-centred decision-making is required for inclusion in the extension of D-words, it leaves open what that degree is. As a result, the metaphysical landscape contains too many properties D-words could

pick out under interpretations extending beyond the D-specification. This not only means that the D-specification is vague in the sense of failing to establish sharp referential cutoff points. The instantiation conditions of the properties D-words would refer to under alternative interpretations vary significantly in how demanding they are regarding fairness and possibly other aspects of decision-making systems they admit as instances.

To be clear, some clues about the minimum threshold for falling under D-words can perhaps be drawn from our usage of the expression. As Cappelen points out, in many paradigmatic instances of D-words, a massive number of decisions of enormous importance are made in a way that strikes us as undemocratic (pp.81-82). For example, the systems in the countries such as Canada or Norway are standardly classified as paradigmatic democracies. However, a large number of important decisions in these countries are made not by the people, but from above by judges, central banks, or private corporations. This suggests that we interpret D-words as referring to a property the minimum threshold for instantiating which is undemanding enough to encompass these systems.

Also, we commonly use D-words in a way that suggests we interpret them as referring to a multiply realisable category. For example, we commonly talk about 'representative democracy', 'deliberative democracy', 'direct democracy', 'liberal democracy', or 'majoritarian democracy'. The fact that we attach all these modifiers to 'democracy' suggests that D-words refer to a category whose membership conditions are disjunctive, encompassing not just one but several alternative decision-making systems. Importantly, these systems aren't grouped together under D-words arbitrarily, as all of them can be understood, in line with the D-specification, as alternative proposals for how to enable the people to influence the outcomes of decision-making through fair processes.

Yet, even when combined with these additional constraints, the D-specification specifies only some aspects of the KQIE structure. It allows us to say that what the input mechanisms of the systems in the extension of 'democracy' have in common is that all of them contribute — albeit to varying degree — to fairness in people-centred decision-making. Nevertheless, considerable indeterminacy lingers on regarding Quantity, Kind, and Embedding. The D-specification remains

silent on what quantity of which kind of decisions, under what embedding conditions, constitutes the minimum threshold required for falling into the extension of D-words.

To summarise, the above discussion reveals two countervailing insights. The bad news is that we may lack a shared interpretation of D-words detailed enough to allow us to fully answer any of the three questions about what 'democratic' refers to. However, the good news is that our shared interpretation still seems sufficiently detailed to make the Relevance Question and the Weighing Question partially answerable, rendering it meaningful to ask them. The most plausible interpretation of D-words, in light of this outcome, is that while many aspects of their reference remain unsettled, they most likely function as *partially semantically settled expressions*.

6.4. The Epistemic and Practical Usefulness of D-words

It seems uncontroversial that expressions needn't be fully semantically settled to serve as useful representational devices. After all, many expressions are vague or open-textured, allowing for borderline cases or novel situations where their applicability is unclear, yet remain perfectly usable within their settled domains of application.[107] At the same time, the semantic indeterminacy of D-words runs deeper: it stems not only from their vagueness and open-texture but also from the underspecification of their application conditions beyond the D-specification. Nonetheless, I will now argue that even this deeper indeterminacy doesn't preclude D-words from being highly useful expressions to rely upon in our lives.

The benefits of using D-words can be divided into epistemic and practical. Let's start with the epistemic benefits. Generally speaking, the key epistemic difference that referential expressions make to their users is that they coordinate their attention to what they refer to. If I'm correct that users of D-words generally tend to agree on the D-specification, it follows that when they deliberate on whether to apply a D-word to a decision-making system, they are prompted to pay attention to *the category of fairness in people-centred decision-making*. That is, they are prompted to consider how fair the system is when it comes to the processes through which it enables the people to influence its outcomes. This, in turn, makes users sensitive to what

[107] There is also discussion about whether open-texture and vagueness cannot, in some cases, even enhance an expression's usefulness (see, e.g., Hart, 1961, pp.120-132; Lanius, 2021; Queloz, 2025, pp.205-208).

dimensions contribute to the realisation of fairness in people-centred decision-making as well as how these dimensions compare to each other in their relative importance.

In the previous section, I have already given some examples of the dimensions that speakers commonly treat as relevant to fairness in people-centred decision-making: these include equal electoral participation, access to diverse and reliable media, freedom of speech and association, independent civil society, periodic institutional review, and government accountability mechanisms such as transparency and regular elections. Other relevant factors may include the proportionality of representation in decision-making bodies; separation of powers with checks and balances; protections against electoral manipulation or coercion; the extent to which minority voices are safeguarded in the decision-making process; the impartiality and independence of the judiciary; the consistency of electoral laws; and the presence of avenues for public deliberation and civic engagement, among others.[108]

Now, Cappelen would likely object that, even if the above dimensions deserve our attention, each is such that speakers can direct their attention to it without using D-words. This is because they can simply discuss them in different terms, such as those I used above when listing them. This would, however, be to overlook the central epistemic import of D-words. The import is that D-words allow us to interpret the given dimensions from a distinctive, integrated perspective — one we would miss out on if we discussed them using different terminology.[109] Specifically, when we attend to the given dimensions as factors relevant to how 'democratic' various decision-making systems are, we interpret them not merely as isolated elements but within the holistic picture that brings to the fore their potential contribution to fairness in people-centred decision-making.

D-words can enrich us epistemically to the extent that they promote such an integrated perspective. By prompting us to interpret the given dimensions not in isolation but as components of a more complex decision-making system, they allow us to uncover various interesting connections and tensions between them. For instance, separation of powers and government

[108] That D-words draw our attention to such dimensions is evidenced by the fact that many of them are used as measures of democracy in various democracy indices, such as those produced by the Economist Intelligence Unit (2024), the V-Dem Institute (2024), or the Bertelsmann Transformation Index (2024).
[109] The relevant notion of perspective is close to Elisabeth Camp's, who understands perspectives as 'open-ended dispositions to interpret, and specifically to produce intuitive structures of thought about, or *characterisations* of, particular subjects' (2019, pp.18-19).

accountability appear complementary in achieving fairness in people-centred decision-making. Conversely, while both freedom of speech and rational public discourse are instrumental to achieving fairness in people-centred decision-making, they can potentially conflict with one another, as unlimited freedom of speech may undermine rational discussion. Recognising these relationships allows us to better discern when the given dimensions can be aggregated and when they cannot. Furthermore, it also allows us to engage in fruitful discussions about how to compare their relative importance and reconcile between them when conflicts arise.

To illustrate, consider the tension between majority rule and minority rights. While majority rule is often regarded as a paradigmatic example of a democratic decision-making method, it can be challenged on the grounds that it sometimes undermines the inclusiveness of decision-making by marginalising or harming minority groups. In such cases, discussions might focus on comparing the relative importance of majority rule and the protection of minority rights for the purpose of achieving fairness in people-centred decision-making, as well as on how to balance these competing priorities. This could involve implementing mechanisms such as supermajority requirements, constitutional safeguards, or deliberative forums that amplify underrepresented voices. Such discussions prompt us to think about how D-words relate to other concepts in their vicinity, thereby deepening our understanding of the complexities of the category they track.

Additionally, insofar as D-words promote discussions about fairness in people-centred decision-making, such as those described above, this opens the possibility that many disputes involving D-words aren't merely verbal. To explain, one standard procedure for diagnosing whether a dispute is merely verbal that Cappelen (2023, p.34) himself recommends is what Chalmers introduces as 'the method of elimination' (2011, pp.526-527). The method consists of temporarily barring the use of an expression suspected to be a source of a merely verbal dispute and checking whether any substantive disagreement remains between the parties of the original dispute, which can be articulated without using the original expression. If such a disagreement remains, the original dispute wasn't merely verbal; if not, it was.

However, it's far from clear how often disputes involving D-words qualify as merely verbal under this method. After all, we can now see that divergent interpretations of D-words might be more unified than initially assumed: they may reflect different attempts to cash out the conditions

under which a decision-making system enables the people to influence its outcomes through fair processes. The disagreement about these conditions would likely persist even if we temporarily barred the use of D-words. Moreover, this disagreement seems substantive because what underlies it are deeper moral disagreements over such questions as how people-centred decision-making systems must operate so that they don't give morally unjustified advantages to any of their participants, or what category 'fairness' should refer to in the first place, given the important moral role the term plays in our lives.[110] While there may still be other disputes involving D-words that are merely verbal, the presence of substantive questions about fairness in the vicinity of D-words gives us reason to be more cautious in diagnosing them as such.

Besides their epistemic utility, D-words can also enrich us *practically* as action-guiding expressions. Specifically, they guide us towards acting so that various components of our society — such as electoral systems, checks and balances mechanisms, public institutions and offices, various policies, media, NGOs, civil society — jointly promote fairness in people-centred decision-making. This translates into what norms we expect one another to follow in public life, what possibilities for public action we perceive in the political environment around us, as well as how we respond to the violations of the given norms and the obstruction of the given possibilities.

Firstly, even speakers whose conceptual repertoire doesn't contain D-words can perceive people in their society as inhabiting various public roles, such as citizens, politicians, governors, journalists, activists, bureaucrats, judges, legislators, executives, etc. However, if speakers possess D-words and see they are regularly applied to their society, this considerably alters the normative lens through which they perceive these roles. Namely, such speakers are disposed to frame these roles as involving various distinctive norms they view as conducive to a healthy democracy grounded in a culture of fair people-centred decision-making. Some examples of these norms might be the norm that politicians should be accountable to the people for their decisions and responsive to their interests, the norm that journalists should serve as public watchdogs, the norm

[110] As Chalmers (2011, p.542) himself would recognise, the latter disagreement is substantive despite being metalinguistic. This is because it revolves around an important normative question in conceptual ethics: which of the multiple moral categories that 'fairness' could pick out would make it the most useful concept for various tasks in our lives where we typically rely on it — such as everyday moral guidance, conflict resolution, collective decision-making, processing the experience of due or undue treatment, working towards social justice, etc.

that citizens should strive to actively participate in public affairs, or the norm that individuals working across various branches of government should maintain a system of mutual oversight and checks and balances.

Secondly, D-words also seem to be expressions whose usage can affect what affordances — that is, the possibilities for action — users perceive in the political environment around them. To the extent that users consider D-words to be applicable to a society's decision-making system, this disposes them to perceive the society as affording distinctive kinds of actions they regard as characteristic of how public affairs are organised in fair people-centred decision-making systems. For example, they might perceive the affordance to voice one's opinion in a public assembly, the affordance to participate in elections, the affordance to run for elected office, the affordance to organise public gatherings, the affordance to engage with those in power, or the affordance to demand transparency and responsibility in policy-making. Alternatively, if they consider 'democracy' not to be applicable to their society but think it should be, this likely disposes them to perceive the affordance to reform the society accordingly.

Thirdly, suppose a society's political representation seeks to undermine some of the norms in the first example or to obstruct some of the affordances in the second example. For example, it curtails the independence of major media outlets, disrupts the mechanisms for controlling power, or suppresses various forms of civic engagement. Admittedly, members of this society might recognise these acts as abuses of political power even if their conceptual repertoire doesn't include 'democracy'. However, having 'democracy' at their disposal enables them to vividly recognise that these actions aren't merely isolated abuses of power, but coordinated stages of a larger destructive process of 'democratic backsliding'. This makes speakers more alert to wider risks that these actions imply, which gives them stronger reasons to criticise their representation and collectively mobilise in protest against it. It's no coincidence that the slogan 'saving democracy' often emerges in protests against systemic abuses of political power in countries like Hungary, Slovakia, or Poland (Spike, 2025, Euronews, 2024; Al Jazeera, 2017).

The above three examples effectively show that D-words have significant practical import for their users. They steer them towards particular normative outlooks, affordance perception, and responsiveness. By doing so, they empower them by opening up to them new patterns of practical

reasoning through which they can coordinate their public affairs and collectively shape their society in distinctive ways oriented towards fairness in people-centred decision-making.

To summarise the second part of my argument, D-words foreground the category of fairness in people-centred decision-making in their users' minds, presenting it as deserving their attention. Attending to this category yields epistemic benefits, as it prompts users to interpret various relevant dimensions from a distinctive, integrated perspective. Moreover, it also yields various practical benefits. Most importantly, D-words needn't be fully semantically settled to foreground the given category. It's enough that their users widely agree on the D-specification, as argued in the previous section.

6.5. Objections and Replies

So far, I have argued that D-words are partially semantically settled expressions and, as such, still possess substantive epistemic and practical value for their users. However, one might worry that this alone is insufficient to justify retaining D-words in our language, particularly in the face of Cappelen's overall case for their abandonment. In this section, I address two foreseeable objections along these lines.

The first objection goes as follows: 'Let's assume D-words indeed coordinate our attention to the category of fairness in people-centred decision-making and, as a result, produce various epistemic and practical goods. However, the same goods can be easily produced even without D-words. I might be right that expressions that refer to various isolated components of fairness in people-centred decision-making cannot do so as effectively as D-words. Nonetheless, speakers seem capable of thinking about the category of fairness in people-centred decision-making even without relying on D-words. This suggests it's perfectly possible to reap all the epistemic and practical benefits of engaging with this category after dropping D-words. Hence, D-words are redundant expressions.'

I have two alternative replies to this objection. Firstly, there is a significant difference between speakers being able to think of a category and their being motivated to actively attend to it. As Koch and Lupyan (2024, p.14) point out, empirical evidence shows that if a category is lexicalised in a linguistic community — meaning that there is a compact verbal label for it — this

increases the likelihood that members of the community will pay attention to it and incorporate it into their categorising practices.[111] Hence, even if speakers could, in principle, bring the category of fairness in people-centred decision-making into focus in their thoughts and communication by using a lengthy description instead of D-words, D-words likely enable them to do so far more effectively.

Secondly, the objection is based on the background assumption that abandoning an expression simply means altering our conceptual repertoire to the original state in which it would *ceteris paribus* be if the expression hadn't been introduced into it. Yet, this assumption strikes me as simplistic. Abandoning an expression isn't merely an act of omission that undoes the past act of adding it to one's conceptual repertoire. It's rather an act of intervention that can potentially establish new dynamics within the repertoire. In this regard, there is a risk that abandoning D-words will trigger destructive dynamics, for it might diminish speakers' interest in the category of fairness in people-centred decision-making.

Let me elaborate on this worry. If a linguistic community begins actively promoting abandoning an expression, there is a risk that many ordinary speakers will interpret this as signalling that the category tracked by it no longer deserves their attention — even when the real motive behind the abandonment is different. Such an interpretation seems natural, especially when the real motive isn't something apparent — such as that the expression's usage is offensive, or that there are more effective alternative expressions for discussing the same category. This worry seems relevant in the case of D-words, as I doubt ordinary speakers would find Cappelen's reasons for abandoning them immediately obvious. The frequent usage of D-words suggests ordinary speakers generally see them as useful and unproblematic expressions to rely on. Hence, the proposal to abandon D-words would likely be initially met with confusion and surprise by many speakers. And, even if eventually accepted, its rationale might easily be misunderstood as implying that we should no longer pay attention to the aspects of decision-making that D-words foreground.

Consequently, what the first objection at most shows is that even if D-words had never been introduced into our language, we might still find ourselves attending to fairness in people-

[111] Koch and Lupyan (2024) primarily discuss studies conducted by Zettersten and Lupyan (2020) and Lupyan and Zettersten (2021).

centred decision-making. What it doesn't show is that we would be equally interested in the category even after abandoning D-words. Of course, we might get around this issue if we find a way to promote abandoning D-words without being misinterpreted by many speakers as signalling that the given category no longer merits our attention. My contention is, however, that there is a considerable risk that we will fail in this task, for the misinterpretation is very suggestive. This uncertainty makes retaining D-words the safer choice.

Let's now proceed to the second objection. One might be sceptical about whether D-words really yield the epistemic and practical goods I have identified. It might be suspected that I underappreciated the extent to which D-words are subject to normative exploitation. D-words tend to invoke various positive normative associations in their users' minds, leading them to assume that whatever is in their extension is morally good. There are several ways in which such an *appraisive* usage of D-words complicates my portrayal of them as useful expressions. Firstly, as seen in §6.2, Cappelen cautions that speakers often apply D-words uncritically to whichever decision-making system they favour (2023, pp.115-117). When they use D-words in this manner, they treat them almost as if their role in language were purely expressive**,** without asking what features make decision-making systems belong to their extension. This makes D-words significantly less effective for directing their users' attention to the category of fairness in people-centred decision-making.

Moreover, even when the normative exploitation of D-words doesn't prevent speakers from attending to the given category, it still prevents them from doing so critically and open-mindedly. It causes speakers to misidentify which features of decision-making systems contribute to how fair and, consequently, how democratic they are by focusing solely on what they perceive as good-making features while disregarding others. This diminishes the usefulness of D-words by making them conducive to misguided reasoning. For instance, it might lead some individuals to oppose a system that they perceive as producing bad outcomes on the grounds that it isn't democratic, even if it's, in fact, a paradigmatic example of a fair people-centred decision-making system. Alternatively, it might discourage people from expressing criticism of what they call 'democracy', as this might make them appear to hold deplorable normative views (Cappelen, 2023, p.191). Relatedly, the uncritical appraisal of D-words causes people to take for granted that they pick out the best possible decision-making systems. This fosters an unhealthy state of public

discourse where people allocate disproportionate attention to the decision-making systems classified as 'democratic' while neglecting alternatives (Cappelen, 2023, p.151).

In reply, I agree the normative exploitation of D-words is a genuine problem adversely affecting their usefulness. Nonetheless, it's unclear to me whether abandoning D-words is an effective strategy for resisting their exploitation. The rationale behind the abandonment strategy presumably is that the less D-words are used, the less they resonate with speakers, which might be thought to limit the opportunities for their exploitation. Yet I have two reasons to consider this a precarious idea, which point to risks similar to those I identified regarding the proposal to abandon 'fake news' in Chapter 5.

Firstly, if we stop using D-words, we effectively surrender the expressions to demagogues who actively try to manipulate public discourse by normatively exploiting them. That is, we give demagogues even more room to exploit D-words, unchallenged by more responsible uses of the terms. There is a risk that the result will be the exact opposite of what we aimed for: not only will D-words remain in use, but their exploitative usage may become even more widespread in our discourse.

Secondly, let's assume the abandonment strategy would indeed lead to demagogues having less opportunities to exploit D-words or even to D-words going out of use. Still, the demagogues' interest in achieving the rhetorical effects for which they originally exploited D-words would likely persist. Thus, there is a risk that they would easily find an alternative target of exploitation to serve their purposes. For instance, they might exploit related expressions such as 'self-government' or 'people's rule' in the same manner as they exploited D-words. If this happened, not only would we lose D-words without gaining any benefit in return, but we would also see another expression subject to exploitation in our language. This would arguably leave us worse off in terms of our conceptual resources than we were before.

Yet one might ask: if not abandonment, what alternative strategy should we adopt to counter the exploitation of D-words? I suggest we strive to make our broader linguistic community more resilient to the exploitation of D-words by encouraging a more critical attitude towards using these expressions. This doesn't seem impossible, as the appraisive usage of D-words among

ordinary speakers often appears to be a situational habit rather than a stable linguistic norm. As Cappelen (pp.88-89) notes, there are already contexts where speakers don't uncritically praise what they call 'democracy'. For example, most would agree that piloting a plane or performing surgery shouldn't fall under the extension of D-words. Moreover, as noted earlier, speakers commonly use D-words as referring to a multiply realisable category. This practice is often accompanied by criticisms of some forms of decision-making that are called 'democracy', such as 'direct democracy', and 'majoritarian democracy', or 'illiberal democracy'. So, the non-appraisive usage of D-words is already present in circulation; we simply need to promote its wider adoption.

But what specific steps can we take to help speakers unlearn the habit of using D-words in an appraisive manner? My proposed ameliorative strategy is to implement educational initiatives that help the public understand that the moral implications of fairness in people-centred decision-making (as foregrounded by D-words) are weaker than they initially appear. This will involve cautioning speakers against two specific reasoning fallacies when engaging with the category.

First, we should warn them against making overly strong assumptions about the explanatory links between fairness in people-centred decision-making and other moral goods, such as freedom or equality. This isn't to say no such links exist. For instance, it stands to reason that a decision-making system is fair to the extent that it doesn't privilege some individuals over others, whether in their formal right to participate in it, or in their de facto opportunities to do so. Similarly, one might argue that a decision-making system failing to protect freedom of the press, expression, or assembly is unfair — if only because it denies the public an environment where people can freely influence its outcomes through informed decisions or dissent against those in power. The problem arises, however, when speakers overestimate these links and infer, for example, that if a decision-making system allows people to influence its outcomes through fair processes (and is thus democratic), it necessarily promotes all forms of freedom and equality. Such inferences foster unduly positive attitudes towards D-words.

Secondly, fairness in people-centred decision-making is undeniably a normatively significant property, one whose possession makes its bearer morally better than it would otherwise be. However, we should caution speakers that this comes with an important caveat that can be easily overlooked. The caveat is that when an entity is fair, this can *pro tanto* improve its moral

standing without also making it morally good overall. For example, if a parent gives both their children equal treatment, this ensures that the treatment is morally better than it would otherwise be, without implying that it is therefore morally good overall. After all, they could be spoiling both children equally. Similarly, a democratic decision-making system can be fair without thereby being morally good overall.

It seems premature to conclude that we should abandon D-words before first attempting to raise public awareness of the two identified fallacies. The more successful our efforts in this regard, the less inclined the public will be to uncritically praise whatever they apply D-words to. This will likely also make them more immune to manipulation through the exploitative usage of D-words. Of course, the proposed ameliorative strategy isn't guaranteed to succeed. However, there seems to be no special reason to worry that its failure would make the exploitative usage of D-words more widespread than it currently is. Moreover, since the ameliorative strategy is based on making the public more cautious about the moral implications of fairness in people-centred decision-making, its success could also make them more resistant to potential manipulation through the exploitative usage of other expressions in the semantic vicinity of D-words, such as 'self-government' or 'people's rule'. Therefore, the ameliorative strategy promises to combat the exploitation of D-words while avoiding risks similar to those of the abandonment strategy.[112]

6.6. Conclusion

In this chapter, I raised doubts about Cappelen's diagnosis of D-words as meaningless expressions. Instead, I proposed an alternative diagnosis of D-words that advances two key claims. First, there is likely a partial convergence in how speakers interpret D-words. Given this convergence, D-words are better understood as partially semantically settled expressions that foreground the category of fairness in people-centred decision-making. Second, by foregrounding this category,

[112] This isn't to say that ceasing to use an expression can never be the preferable strategy to the ameliorative strategy because of the risks it involves. A case in point is Kazankov and Yi's (2024) analysis of how 'critical race theory' ('CRT') has been weaponised to stoke fears about race-conscious education in U.S. schools. They argue that this exploitation is based on portraying what 'CRT' picks out as a dangerous left-wing academic ideology imposed on schoolchildren (2024, p.1198). This creates a tricky situation where exploiters benefit even from CRT scholars correctly calling their own discipline 'CRT', as this makes the given portrayal more believable. It's worth considering whether they should adopt a different label for their discipline to undercut this exploitation. However, D-words face no such predicament. As long as they are used without positive associations, their exploitative use is challenged rather than reinforced.

D-words offer various epistemic and practical benefits to their users. My alternative diagnosis makes Cappelen's case for abandoning D-words significantly less convincing. Admittedly, D-words emerge even from my diagnosis as far from perfect expressions. However, perfect expressions may not be what we need in the first place. Instead, what we need are expressions that cognitively enrich us by enabling us to interpret the world through useful, albeit imprecise and complex categories that otherwise risk escaping our attention. I believe D-words meet this criterion under my diagnosis, which makes them worth using.

Part III – Conceptual Ethics at the Intersection of Epistemic and Moral Values

Chapter 7. Navigating Epistemic-Moral Conflicts in Conceptual Ethics

7.1. Introduction

In this chapter, which forms the final part of this dissertation, I examine a problem for conceptual ethics that emerges from a tension between two observations about what the project involves. The first observation is that conceptual ethics involves normative deliberations about how a concept should be designed. The expected outcome of such deliberations is that participants make an all-things-considered judgement about which of the alternative conceptual designs is supported by the greatest balance of considerations in its favour over those against it. To reach this judgement, one needs to weigh up various competing considerations in their relative importance to the concept's overall value.

The second observation is that the considerations relevant to a concept's value tend to be highly diverse. Many of them can be classified as either epistemic or moral. Epistemic considerations include how accurately a concept tracks the world, how informative it is, what inferential and explanatory power it possesses, and whether its definition is internally coherent. By contrast, moral considerations include what positive or negative effects a concept has on someone's well-being, the moral status of the norms it promotes, how it enables users to shape the world around them, what character traits and behavioural patterns it fosters, and what morally significant practical reasons that follow for its users from its application.[113]

Why do these two observations stand in tension? To explain, it seems compelling to think that epistemic and moral considerations often pull in opposite directions. We don't even need to look to conceptual ethics to find situations where epistemic and moral considerations come into conflict in an agent's practical reasoning. For example, a commonly discussed epistemic-moral dilemma in the ethics of clinical research concerns whether researchers should conduct epistemically valuable clinical trials that expose their subjects to potential moral harms (Jonas, 1969; Levine, 1988). Also, some authors in the ethics of belief literature argue that even true and epistemically rational beliefs can still morally wrong the agent if they involve racial or gender

[113] The epistemic-moral dichotomy isn't exhaustive, as other considerations — such as prudential or aesthetic ones — can also be relevant to a concept's value.

profiling (Fabre, 2021; Fritz, 2020).[114] This raises the question of whether an agent should form such beliefs. Last but not least, in political philosophy, Lepoutre (2024) discusses whether public speakers are permitted to use falsehoods in their speech to mobilise their audience in service of a weighty just cause. All these cases involve dilemmas in which epistemic considerations favouring an action come at the cost of moral considerations, or vice versa.

As discussed in the next section, even conceptual ethics isn't spared from such dilemmas: conflicts between epistemic and moral considerations arise even in the course of deliberating about how we should design representational devices, such as linguistic expressions. However, epistemic and moral considerations appear to be so different in kind that it's difficult to see how conceptual ethicists could non-arbitrarily address such conflicts using the standard method of weighing them on a balance — that is, by comparing their relative contributions to a concept's all-things-considered value.[115] One might worry, however, that if epistemic and moral considerations form two incomparable normative domains, this considerably diminishes the scope of conceptual ethics. We will see that conceptual ethics raises many interesting dilemmas that prompt us to engage in cross-domain comparisons of the relative weights of conflicting epistemic and moral considerations. It might be thought that no systematic progress on resolving these dilemmas can be made if the two domains cannot be non-arbitrarily compared.[116]

What I aim to show in this chapter is that, even if epistemic and moral considerations are incomparable, the outlook for conceptual ethics isn't as bleak as it may seem. I argue that although epistemic-moral conflicts are likely irresolvable through cross-domain comparative reasoning, there is an alternative and highly promising approach that conceptual ethicists can use to address them. Namely, they can adopt the *integrative approach* to these conflicts. This approach is based on the idea that epistemic-moral conflicts often arise from contingent and malleable factors. It aims

[114] See, however, Schroeder (2018) and Basu and Schroeder (2019), who argue that epistemically well-formed beliefs cannot morally wrong someone, since the epistemic status of a belief partly depends on moral considerations.

[115] The question of whether and how distinct kinds of considerations relevant to a concept's value can be compared has been previously raised at a general level by Burgess and Plunkett (2013, p.1106), Plunkett and McPherson (2016, pp.210-211), and Löhr and Veluvenkamp (2025, pp.3-5) as well as in relation to epistemic and moral considerations by Simion (2018a), Podosky (2018, 2022), Crisp (2025), and Queloz (2025, pp.293-295). Still, this question hasn't been sufficiently explored, and the chapter's objective is to fill that gap.

[116] While, as noted above, epistemic-moral conflicts aren't unique to conceptual deliberation, this is no reason for complacency. On the contrary, given that representational devices have wide-ranging practical and epistemic implications, conceptual ethics invites us to take such conflicts more seriously.

to examine the extent to which agents can integrate epistemic and moral standards of reasoning by altering these factors, making it easier for the standards to converge in their overall verdicts on what features an expression should have. I show that once we recognise what kind of tools linguistic expressions are, the potential for successfully integrating epistemic and moral standards in conceptual ethics proves greater than it initially appears.

Here is the chapter's roadmap. In §7.2, I introduce three types of epistemic-moral conflicts in conceptual ethics. In §7.3, I present a case for scepticism regarding the possibility of non-arbitrarily addressing these conflicts through the method of cross-domain comparison. In §7.4, I discuss three strategies the proponents of cross-domain comparison might adopt to counter this scepticism. I argue, however, that all of them ultimately fail. §7.5 is the most constructive part of the chapter, in which I show that the integrative approach presents a promising way of addressing epistemic-moral conflicts in conceptual ethics. I first identify two general points about linguistic expressions one should keep in mind when assessing their performance in both epistemic and moral tasks. I then use these points to demonstrate how successful epistemic-moral integration can be achieved in the examples of conflicts introduced in §7.2.

7.2. Epistemic-Moral Conflicts

There are several ways in which epistemic and moral considerations can pull in opposite directions. In the context of conceptual ethics, the following three conflicts are readily conceivable:

> *Conflict 1:* A representational device brings about a substantive moral good by promoting false beliefs, or brings about a substantive moral bad by promoting true beliefs.
>
> *Conflict 2:* A representational device brings about a substantive moral good by directing its users' excessive attention towards one phenomenon, thereby creating attention deficits with respect to other phenomena.
>
> *Conflict 3:* An increase in theoretical coherence and explanatory unity within a conceptual network brings about a substantive moral bad.

To give the reader a better sense of how these conflicts can arise, let me exemplify each of them.[117] As an example of Conflict 1, imagine a country whose regional definition of 'refugee' matches the international legal definition, which qualifies someone as a refugee only if they flee due to a well-founded fear of political persecution (United Nations, 1951). However, the country experiences an influx of people fleeing for other reasons, such as external aggression, internal conflict, climate disaster, or massive human rights violations. Suppose we are political authorities in this country who know three facts about its citizens: (1) they are poorly informed about the regional 'refugee' definition; (2) they tend to believe that 'refugee' is defined as their authorities use it; and (3) the word 'refugee' resonates with positive associations among them, so they respond with empathy and friendliness towards those newcomers they consider to fall under 'refugee' but behave in a very unwelcoming and discriminatory way toward those they consider excluded by the term 'refugee'.

Taken together, these three facts give rise to the following question: As political authorities, should we start deceptively applying 'refugee' to newcomers who flee their country for reasons other than persecution? Such usage of 'refugee' would promote among the country's citizens the false belief that, under the current regional definition, even these newcomers qualify as refugees. There seems to be a strong moral reason to do this, given that the likely effect would be that the newcomers receive more sensitive treatment from the citizens, significantly increasing their well-being.

At the same time, the strategy is based on deception, which seems to be an epistemic reason against it. This reason is also arguably very strong, for the deceptive application of 'refugee'

[117] Besides the examples discussed below, there is also the following well-known example of an epistemic-moral conflict presented by Simion (2018a, pp.920-921). The term 'deer' is taxonomically defined as referring to hoofed ruminant mammals. It's a perfectly epistemically functional term, well-suited for biological research. Suppose, however, one subspecies within the deer population — roe deer — is at risk of extinction because neither hunting legislation nor conservation legislation distinguishes between roe deer and less vulnerable deer populations. Consequently, if we don't revise 'deer' to artificially exclude roe deer — thereby making it less epistemically accurate — their population is likely to die out. While this example might sound more familiar to the reader than my own, I take issue with it on two grounds. First, it neglects that taxonomic definitions in biology are shaped not only by epistemic considerations but also by pragmatic interests and purposes (Dupré, 1981, pp.83-90). Second, the example fails to adequately motivate why our hunting and conservation laws couldn't simply be revised to account for the distinctive vulnerability of roe deer without excluding them from 'deer'. For these reasons, I base my discussion on other examples of epistemic-moral conflict.

doesn't involve merely producing a benign semantic lie about how 'refugee' is regionally defined. 'Refugee' is a world-making legal and social kind term: it not only represents but also shapes social and legal reality as its core extra-representational effect. Hence, how the term is defined has far-reaching worldly consequences. Whatever legal kind 'refugee' refers to, this kind is constituted by the definition of 'refugee' in the sense that the very fact that 'refugee' is defined to refer to it enables its instantiation. Therefore, the conditions newcomers must satisfy in order to enjoy various privileges that follow from being a refugee depend on how 'refugee' is defined in the country's jurisdiction. Lying to the citizens about the definition of 'refugee' amounts to promoting a host of false beliefs about the portion of social and legal reality built around this definition.[118]

Next, for an example of Conflict 2, consider an expression that stands for a positive political ideal, such as the term 'democracy' discussed in the previous chapter. I argued that 'democracy' can have a significant action-guiding force for its users. It can direct their attention towards relations between various features of decision-making systems, which enables them to collectively organise their public affairs in distinctive ways conducive to the fair reflection of popular will. Let's make the following two assumptions for the purposes of this example. The first assumption is that what considerably helps 'democracy' guide its users towards organising their public affairs in these ways is that it's a thick concept — one its users not only use to convey that the system they apply it to has certain descriptive features, but also that it deserves strong moral commendation for having them. That is, they appraise this system, cherishing it as the ultimate political ideal towards which their own society should strive. This assumption doesn't seem very contentious. Surely, the users that wouldn't appraise the system 'democracy' tracks would be significantly less motivated to shape their society in accordance with it. The second assumption is

[118] There are also structurally reversed conflicts in which a representational device causes a substantive moral bad by promoting a true belief. A well-known example is Jenkins' critique of Haslanger's (2000) definition of 'man' and 'woman' as social class categories based on presumed sex. Haslanger argues that her definitions are epistemically useful for analysing and raising awareness of sex-based oppression, but Jenkins (2016) faults them morally for marginalising trans people. What complicates this example, however, is that raising awareness about social oppression is clearly not only a substantive epistemic good but also a moral one, while marginalising trans people involves ignorance of their existence, which is a substantive epistemic loss. Hence, this clearly isn't merely an epistemic-moral conflict but also a case of internal conflicts between two substantive epistemic considerations and two substantive moral ones, respectively. By contrast, although deceiving people about the definition of 'refugee' is *pro tanto* morally wrong, here the moral harm is clearly outweighed by the substantive moral benefit of lying.

that the patterns of organising public affairs that 'democracy' promotes indeed present the morally best ways to organise political societies. This is a more controversial claim. Readers who find it unpalatable to tentatively accept this claim even as an assumption should feel welcome to replace 'democracy' in this example with a different term that tracks their favoured political system. These two assumptions together imply that, in our example, there is a strong moral reason for 'democracy' to be a thick concept associated with positive moral valence.

There is, however, a strong countervailing epistemic reason against 'democracy' being a thick concept. Recall from Chapter 6 one of the criticisms that Cappelen levels against 'democracy' (2023, p.151). Speakers who use 'democracy' as a positive thick concept, believing that whatever falls into its extension deserves strong moral appraisal, tend to single-mindedly focus on the extension of 'democracy' while ignoring the possibility of alternative decision-making systems. This attention deficit arguably deprives the speakers of various epistemic gains they could access if they were more attentive and open-minded towards alternative decision-making systems and engaged in inquiry about them. Some possible examples of such gains include learning new information about the alternatives, acquiring a deeper understanding and the ability to critically justify why democracy is preferable, or expanding their imagination about what ways of political organisation are possible.

Finally, to illustrate Conflict 3, we can use Queloz's (2024a) recent reconstruction of the debate between Dworkin and Williams over how to define the political concepts of liberty and equality. At the heart of this debate is the observation that, although many political societies claim to jointly uphold the values of liberty and equality, 'liberty' and 'equality' are, under their standard interpretations, what Queloz (2024a, p.10; 2025, pp.161-162) calls 'non-accidentally incongruent concepts', i.e., concepts such that the realisation of one systematically comes at the expense of the realisation of the other. Specifically, it seems impossible for a state to fully embrace the value of equality without sacrificing some individual liberties by regulating the distribution of resources among its citizens. Conversely, it seems impossible for a state to fully embrace the value of liberty without allowing its more powerful citizens to enjoy unrestricted personal freedoms, thereby perpetuating inequalities between them and less powerful citizens.

Dworkin (2011, p.4) addresses this tension by proposing we immunise 'liberty' and 'equality' against this conceptual tension by redefining 'liberty' so that it doesn't denote personal freedom from interference but instead denotes a political *right* that must be distributed in accordance with the principle of equality. On this redefinition, 'liberty' is no longer incongruent with 'equality' because it denotes a value that can be realised only if it's equally realised among all citizens. While, as Queloz (2025, p.159) notes, conceptual incongruence doesn't automatically imply incoherence, it can be argued that, given many political societies view themselves as jointly promoting the values of both 'liberty' and 'equality' — and even apply them to the same objects— such incongruence can easily lead speakers to hold incoherent and explanatorily disunified beliefs about the subject matter. Hence, to the extent that eliminating the tension between the two concepts renders these beliefs more coherent and explanatorily unified, this constitutes a strong epistemic reason in favour of doing so.[119]

Williams (2005, pp.83-86), however, presents a moral reason against Dworkin's redefinition of 'liberty'.[120] He argues that once the congruence between 'liberty' and 'equality' is secured, the two concepts no longer enable their users to develop a morally mature political community where fellow citizens can co-exist despite political disagreements. This is because they then fail to do justice to how life under a political order is experienced by some community members. To explain, when the state makes a decision that promotes greater equality, some citizens often feel they are on the winning side, while others feel they are on the losing side. The losing side consists of those who complain that the decision promoting equality has restricted their liberty. Williams thinks 'liberty' should remain a concept that renders these complaints intelligible by allowing for the possibility of conflict with 'equality'. Otherwise, it morally falters because it encourages those on the winning side to patronisingly dismiss these complaints as mere conceptual confusions (2005, pp.85-86). Such dismissals are morally harmful because they show disrespect for the experiences of their opponents. Without respect, healthy relationships between disagreeing parties within the political community cannot be cultivated.

[119] Note that, while Dworkin (2011, p.122) views conceptual congruence as a theoretical virtue, he doesn't himself present it as contributing to the concepts' epistemic value. See Queloz (2024a, pp.4-5) for details.
[120] See Queloz (2024a, pp.14-18) for a detailed discussion.

As the above three examples demonstrate, epistemic and moral considerations often exert opposite pressures on our judgement about what features a concept should possess. The resulting epistemic-moral dilemmas are far from negligible. Considering that humans actively rely on concepts to navigate the world and its normative complexities, it seems impossible to shield them from such dilemmas. Most ordinary language concepts are used across various situations and practices, where their potential effects can be both moral and epistemic. Consequently, normative deliberation about any expression used outside the artificial confines of a single normative domain can encounter epistemic-moral dilemmas. Furthermore, these dilemmas arguably present some of the most interesting problems in conceptual ethics, often concerning aspects of expressions that substantively affect their overall value for users. Ignoring them would thus be a missed opportunity for the project.

7.3. The Overarching Standard of Comparison and the Arbitrariness Problem

It might, however, be doubted whether epistemic-moral dilemmas are resolvable in the first place. Generally speaking, the most common way in which dilemmas between conflicting considerations are resolved is through comparative reasoning, which involves weighing these considerations and reaching a verdict about which matters more. Yet comparability is a three-place relation between the compared options and a standard of comparison. That is, all rational comparisons must be guided by some deliberative standard for determining the relative weights of the compared items (Chang, 1997, pp.4-7; Chang, 2002, p.666; Thomson, 1997, p.276; Orsi, 2015, p.103; Andersson, 2016b, p.380).[121]

[121] I follow here Stocker (1980, p.176; 1997, p.203) and Chang (1997, p.2; 2001, pp.51-54) in understanding *comparability* as a weaker relation than *commensurability*, in the following sense: for two items to be comparable, it's sufficient that there exists a deliberative standard for their *ordinal comparison* — that is, a standard that allows us to rank them based on evaluative differences between the items relative to it. By contrast, for two items to be commensurable, there must also be a deliberative standard for their *cardinal comparison* — a standard that enables not only their ordering but also the precise measurement of *how much* they differ relative to some unit of value. Although some authors use 'comparability' and 'commensurability' interchangeably, taking both to imply the possibility of cardinal comparison (e.g., Raz 1986; Anderson, 1997; Rabinowicz 2021), such an understanding of 'comparability' seems to me too restrictive for two reasons. First, it neglects the plausible possibility that certain standards of comparison may involve too much indeterminacy for precise measurements of evaluative differences between two items to be possible (Parfit, 1986, pp.430-431). Second, the comparability of items on a cardinal scale presupposes what Chang (2002, p.660) calls the *trichotomy thesis* — the thesis that all comparable items can stand to one another only in three comparative relations: better than, worse than, or equally good. In other words, evaluative differences between them are either zero or biased in favour of one item. This, however, neglects an interesting possibility defended by Chang (1997, pp.25-27; 2002): there may be a fourth comparative relation, *parity*, in which two items stand when there is a non-zero evaluative difference between them that isn't biased in favour of either one. While whether parity is a genuine comparative relation remains contested (see, e.g., Carlson, 2010; Rabinowicz, 2008;

To illustrate, consider someone deliberating over whether to buy a leather wallet or a fabric wallet by means of comparative reasoning. It seems plausible that their comparison can proceed rationally only if it's guided by some deliberative standard — for instance, a prudential standard that considers which wallet better serves one's personal interests; a moral standard that considers which option is more ethically acceptable; or an aesthetic standard that considers which wallet looks more attractive or better fits one's personal style. Each standard offers a different way of determining which considerations are relevant and what weight they should carry when deliberating over which wallet to buy. Without such a standard, or some other deliberative standard at play, it's unclear what evaluative relations the person's comparison could track, and thus how their deliberation could count as rational rather than arbitrary.

Now, all the comparisons in the above example take place within a single normative domain. In such cases, there is a clear method for articulating the content of the deliberative standard guiding the comparison: one must ask what the aim of the relevant domain is. For example, suppose one deliberates over which of two possible designs for an expression is epistemically better. The epistemic domain aims to guide agents in organising their cognitive systems so that they accurately track the world and possess various features related to it. Accordingly, one would proceed by identifying the considerations in favour of and against the two alternative designs that pertain to how accurately the expression's users track the world, and then compare their relative contributions to this aim.

Yet, it's far less clear how to articulate a standard of comparison when one wants to conduct a comparison *across* several different normative domains. The possibility of cross-domain comparisons presupposes that the existence of the standard of comparison with the following two characteristics (Sagdahl, 2022, pp.29-33). First, it must be a comprehensive standard that is able to take into account every consideration that is relevant for comparison in at least one of the target domains and assign it a non-arbitrary (even if not fully precise in a cardinal sense) relative weight (cf. Chang, 2004, pp.2-3; Stocker 1990, p.172).[122] Second, it cannot be just any third standard but

Andersson, 2016a), we arguably shouldn't rule out its possibility from the outset simply by how we define 'comparability'.

[122] Note that the belief in such a comprehensive standard doesn't commit one to value monism, i.e., the thesis that there is ultimately just a single super-value to which all other values can be reduced. This is because the comprehensive

the most authoritative standard whose verdicts take normative priority over the verdicts of the qualified standards derived from each domain (cf. Reisner, 2018, p.224; Wedgwood, 2004, p.406; McPherson, 2018, p.265; Copp, 2007, p.294). After all, the very point of making cross-domain comparisons is to step out of individual normative domains and assess which conflicting considerations matter more than others *conclusively*, not just by some standard or other. In the context of the present debate, what we need in order to conduct rational epistemic-moral comparisons is, therefore, *an overarching standard of comparison* that takes both moral and epistemic considerations into account and adjudicates between them in the most authoritative way, which settles which consideration is *all things considered* more important.

The idea that individual normative domains can be compared under an overarching standard of comparison has, however, been met with a lot of scepticism. Some theorists argue that it's an incoherent idea. Copp (2007, ch.9), for instance, offers the following reductio argument against this view: suppose S is the overarching and thus most authoritative standard. Then there must be a standard R that determines S as such. R must itself be the most authoritative standard, or else S's authority would be undermined. But R cannot be identical to S, since no standard can establish its authority merely by meeting its own criteria — any standard could do that. Nor can R be distinct from S, as that would contradict the assumption that S is most authoritative. Hence, there is no coherent way in which S can be the overarching standard of comparison.

Some authors have objected to Copp's argument by suggesting that the authority of the overarching standard can be a brute fact needing no explanation (McLeod, 2001), a conceptual truth requiring no independent justification (Dorsey, 2016), or grounded in something other than another standard (Baker, 2018). However, the last of these authors, Baker (2018, pp.234-236), doubts the coherence of the overarching standard of comparison on independent grounds. He argues that the characterisation of this standard as having special normative force lacks clear content because it relies on metaphors such as 'authoritative', 'overriding', or 'genuinely ought'. These metaphors, he argues, demand clarification, but attempts to provide it either fall into circularity by invoking the original metaphors or invoke psychological terms, which shift the topic to what agents *believe* they ought to do, rather than what they *normatively* ought to do.

standard can be seen as reflecting some higher-level covering value that has lower-level plural values as its parts but goes beyond them (Griffin, 1986, p.90; Chang, 2004, 2015).

It's beyond this chapter's scope to explore whether Copp's and Baker's arguments are surmountable. However, even if these arguments can be defeated, all this shows is that the overarching standard of comparison is a coherent idea. Scepticism about its actual existence remains a reasonable default position. Epistemic and moral domains are governed by very different aims: the former aims to navigate agents in organising their cognitive system so that it accurately tracks the world; the latter aims to navigate agents in making decisions and organising their actions in a way that respects and promotes the well-being, rights, and dignity of others. Both are legitimate but independent sources of normative justification. It's difficult to see how they could be subsumed under a single, authoritative standard that would serve as a common standard through which we could compare them and conclusively establish that one of them is more important than the other (either generally or in particular choice situations).[123]

What adds to this difficulty is the widely recognised asymmetry in the weighing behaviour of epistemic reasons and practical reasons — of which moral reasons are a subspecies (see Feldman, 2000, pp.680-681; Dancy, 2004, p.95; and Berker, 2018, pp.430-433). Epistemic reasons exhibit what Berker (2018, p.430) calls 'prohibitive balancing'. If one has strong, equally balanced epistemic reasons for believing and disbelieving P, and no other epistemic reasons are relevant, epistemic normativity intuitively recommends suspending judgement about P. By contrast, practical reasons exhibit what Berker (2018, p.430) calls 'permissive balancing'. If one has strong, equally balanced practical reasons for performing each of two incompatible actions, and no other practical reasons are relevant, practical normativity intuitively permits choosing either.[124] Any standard attempting to weigh epistemic and practical reasons on a common scale seems committed to offering a unified weighing procedure for both cases, thereby ignoring this asymmetry. In light of this, it seems more plausible to view the epistemic and moral domains as two incomparable normative systems.

Admittedly, there is also the possibility that the overarching standard of comparison has the status of what Chang (2004, p.3) calls a 'nameless value', which is something that 'can do

[123] This kind of pessimism about the existence of an overarching standard of comparison has been expressed by Feldman (2000, pp.691-694), Papineau (2013, p.70) and Berker (2018), with a focus on epistemic-practical conflicts, and by Tiffany (2007) and Sagdahl (2022, chs.6-7), with a focus on prudential-moral conflicts.

[124] As Berker (2018, p.432) points out, asymmetry remains even when considering strong, equally balanced practical reasons for believing and disbelieving something.

normative work even without the benefit of a name' (2004, p.18). This would mean that the overarching standard exists, but we struggle to articulate and grasp it by means of our imperfect conceptual and cognitive resources. But in that case, it's striking that we are unable to grasp it despite being able to grasp multiple other normative standards. A plausible explanation of this contrast is that, if the overarching standard exists, it must have an extremely complex content whose grasp might go beyond our epistemic limits. This still raises doubts about whether appealing to such an intractable standard can be of any use to conceptual ethicists.[125]

Accordingly, the prospects for conceptual ethics in making progress in epistemic-moral conflicts through cross-domain comparisons are questionable, as there is reason to doubt that a standard guiding such comparisons actually exists — or, if it does exist, that it's epistemically accessible. This raises what I call the 'Arbitrariness Problem' for conceptual ethics. The problem runs as follows: if conceptual ethicists can employ neither cross-domain comparative reasoning nor any alternative systematic method to address epistemic-moral conflicts, they are left with the following two options in the face of these conflicts. Firstly, they can suspend any summative judgement about the relative weights of conflicting epistemic and moral considerations. That is, whenever they see an epistemic consideration supports conceptual choice X, while a moral consideration supports conceptual choice Y, they can refrain from weighing them and just state: 'While epistemically speaking, X should be selected, morally speaking, Y should be selected, and that's it.' Or else, they can just spontaneously plump for one consideration, without following any overarching normative standard by which they could justify their choice to those who don't favour it.

Both options are rather unappealing. The first option amounts to leaving deliberation about a concept's design incomplete. This degrades concepts into unfinished products, which seems to be a suboptimal outcome from both moral and epistemic perspectives. The second option isn't significantly better either, as it involves designing a concept through an arbitrary decision whose

[125] Admittedly, even in qualified domains like the epistemic or moral, deliberative standards aren't fully transparent, as shown by our frequent struggles to settle a proper trade-off between specific values within these domains. For instance, it's unclear whether the epistemic standard prioritises the number of true beliefs or their ratio to false ones, or whether the moral standard favours maximising well-being or protecting individual rights. Still, it would be an overstatement to say these standards are as intractable as the overarching standard. In the former case, we have a general sense of the standards and struggle with their application; in the latter, we struggle even to articulate the standard's content.

reliability seems akin to coin-flipping. Since epistemic-moral conflicts often address important questions about a concept's overall functioning, this option risks degrading many concepts into products of questionable quality due to their important features being designed capriciously. It would be disappointing if these were the only two ways to address epistemic-moral conflicts, as the scope of conceptual ethics, as a systematic normative inquiry aimed at ameliorating our concepts, would then be severely limited.

7.4. Three Responses to the Arbitrariness Problem

I will now discuss three possible strategies for countering the Arbitrariness Problem, each aiming to show that the method of cross-domain comparative reasoning, when properly elaborated, can respond to epistemic-moral conflicts in a non-arbitrary manner.[126]

7.4.1. Simion-inspired Response

The first strategy is to argue there are some facts about the constitutive features of concepts from which we can derive a constraint regarding which of the two standards is normatively more authoritative. Hints of this strategy can be found in Simion (2018a). Simion argues that conceptual engineering shouldn't merely be concerned with fixing defective concepts but should also aim to further improve concepts that are already good enough. Most concepts, including the non-defective ones, are suboptimal in the sense that they allow for further improvement. Simion, however, finds it implausible that just any room for improvement justifies further engineering of a concept. That is, she thinks there should be some normative constraint that sets limits on the extent to which it's rationally permissible to improve concepts. To this end, she puts forth the following constraint on legitimate conceptual engineering (2018a, p.923):

[126] While the considered strategies build on representative ideas from the conceptual engineering literature (Simion, 2018a; Podosky, 2018; Thomasson, 2020), the metaethics literature (Chang, 1997, 2009, 2013), and the ethics of belief literature (Reisner, 2008; Howard, 2020; Meylan, 2021), the list isn't exhaustive. Space constraints prevent me from incorporating into my discussion Booth's (2011, 2012) and Steglich-Petersen and Skipper's (2020) arguments for the comparability of practical and epistemic reasons for belief. Let me note, though, that all these arguments rely on highly controversial premises, namely that someone is justified in believing that *p* iff they are blameless in believing that *p* (Booth, 2011); that epistemic reasons can function as guides to belief only if they can be all-things-considered reasons (Booth, 2012); and that epistemic reasons are instrumental reasons whose strength can depend on non-epistemic reasons to pursue an epistemic aim without themselves collapsing to non-epistemic reasons (Steglich-Petersen and Skipper, 2020).

> *The Epistemic Limiting Procedure (ELP)*: A representational device should be ameliorated iff (1) there is all-things-considered reason to do so, and (2) the amelioration doesn't translate into epistemic loss.

ELP gives primacy to the epistemic standard in conceptual ethics because it states that a concept should be designed in a certain way only if there is no epistemic consideration against doing so. Simion motivates this claim as follows:

> 'Concepts, just like beliefs, are representational devices, their function is an epistemic one: to represent the world. In virtue of this function, concepts will be properly functioning when responsive to epistemic reasons, and malfunctional when responsive to practical reasons. Concepts will be good concepts qua concepts when they are epistemically good.' (2018a, p.923).

Simion's point here is that non-epistemic considerations are the *wrong kind of reasons* for engineering a concept. On the common understanding of the wrong kind of reasons, which underpins Simion's argument, these are the reasons for creating a kind that don't anyhow contribute to how well the kind satisfies its constitutive standard of proper functioning.[127]

As I read Simion, she supports her point that non-epistemic considerations in conceptual ethics are the wrong kind of reasons through an argument from analogy (2018a, pp.922-924). She observes that both beliefs and concepts are representational devices. This, according to her, implies that the constitutive function of both beliefs and concepts is to accurately represent the world — that is the constitutive standard of beliefs and concepts by which we evaluate whether they function properly. In the case of beliefs, this means that forming a belief for a non-epistemic reason is forming it for the wrong kind of reason — one that doesn't bear on how well it performs its constitutive function. For example, suppose someone offers you one million dollars to believe, against all available evidence, that you can fly. This might provide an all-things-considered justification for forming the belief, yet it's the wrong kind of reason because it's unrelated to how accurately the belief represents the world. By analogy, the same consequence must then follow for

[127] The category of the wrong kind of reasons is explicated along these lines in Sharadin (2016), Schroeder (2010), and Danielsson and Olson (2007).

concepts: engineering a concept for non-epistemic reasons is engineering it for the wrong kind of reasons, or so the argument goes.

Before examining Simion's argument, we need to apply two substantive weakenings to ELP in order to extract a potential constraint for epistemic-moral comparisons. Firstly, ELP in its current form is clearly too stringent a constraint, even for those who regard the epistemic standard as prior to the moral one. According to the second sub-condition, even if the epistemic standard overall supports designing a concept in a certain way, doing so is impermissible if it involves *any* epistemic loss. Thus, ELP confines the permissible cases of conceptual engineering only to those where there is perfect harmony between what all epistemic considerations recommend.

This, however, seems overly restrictive, considering that tensions between epistemic considerations are common. Here is one example: expressions that pick out heterogeneous categories can be highly informative, as they allow us to explore interesting variations among their members. For example, 'mammal locomotion' refers to a heterogeneous category encompassing walking, flying, swimming, running, burrowing, and more. This makes it useful for studying a wide range of evolutionary adaptations in different mammals. However, as Egré and O'Madagáin (2019, pp.8-9) point out, such expressions are often less projectible. Since the category includes movements arising from very different anatomical systems and ecological pressures, it's difficult to make reliable inductive generalisations about it — for instance, a property observed in one form of locomotion (like energy efficiency in walking) may not apply to others within the same category, such as flying or swimming.

Consequently, the decision to engineer an expression to be more or less heterogeneous often results in both an epistemic gain and an epistemic loss. According to ELP, it's never rationally permissible to make such decisions solely because some epistemic loss will be incurred, regardless of which option is chosen. This strikes me as an unacceptable consequence. While it may be reasonable to refrain from engineering an expression when the epistemic losses outweigh the gains — or perhaps even when gains and losses are equally balanced, given the prohibitive balancing of epistemic reasons discussed above — it seems excessively restrictive to claim that just any epistemic loss renders such engineering impermissible.

But we can easily get around this consequence if we replace ELP with the following revised version proposed by Podosky (2018, p.9):[128]

> *ELP+*: A representational device should be ameliorated iff (1) there is all-things-considered reason to do so, and (2) there isn't all-things-considered epistemic reason to refrain from amelioration.

ELP+ is weaker than ELP because it allows engineering a concept into a certain form even if there is an epistemic consideration against it, insofar as the overall verdict issued by the epistemic standard doesn't recommend against it. This sounds more plausible.

Yet, ELP+ is still unhelpful in the present discussion because its first sub-condition presupposes that there is a fact about how epistemic and moral considerations relevant to engineering a concept compare in their relative weights. Whether this is so is the very point at issue. In this respect, a weakening of ELP+, ELP++, is a better candidate for the constraint governing epistemic-moral comparisons.

> *ELP++*: A representational device should be ameliorated as recommended by an all-things-considered moral reason only if there isn't all-things-considered epistemic reason not to do so.

ELP++ is a consequence of ELP+ that no longer carries the abovementioned problematic presupposition and concerns epistemic-moral comparisons in particular, avoiding additional commitments about whether to ameliorate when other normative standards, such as prudential standards, come into play. Hence, it's better suited for the scope of our investigation.

Nevertheless, it would be ill-conceived to constrain epistemic-moral comparisons by ELP++ because Simion's argument from analogy doesn't work. The problematic step in the argument is the inference from 'beliefs and concepts are representational devices' to 'the constitutive function of beliefs and concepts is to accurately represent the world'. The reason why beliefs have the constitutive function of accurately representing the world isn't merely that they

[128] Podosky (2018) weakens ELP to ELP+ because he thinks ELP+ permits engineering a concept to be representationally inaccurate, as long as this inaccuracy is just a temporary epistemic loss that is eventually eliminated after the concept changes reality.

are representational devices. Even conative attitudes such as desires or intentions are representational devices, as they have representational content, yet this content has a world-to-mind direction of fit; meaning that accurate representation isn't their function. A more plausible explanation is that truth-regarding considerations — considerations relevant to whether or not beliefs are true — play a special role in doxastic deliberation, i.e., the deliberative process constitutive of decision-making about which beliefs to form.[129] Specifically, as Shah (2003) and McHugh (2013) point out, whenever one deliberates about whether to form a belief, one must treat truth-regarding considerations as decisive for the deliberation's outcome. That is, one must immediately recognise that the deliberation is settled by whether the given belief is true. If one doesn't recognise this point, one cannot be said to engage in doxastic deliberation in the first place. Hence, truth-regarding considerations seem inescapable in doxastic deliberation.[130] This observation would be difficult to account for if accurately representing the world were just one of many different features beliefs might have, rather than the constitutive standard of their proper functioning.

By contrast, there is no similar evidence that truth-regarding considerations inescapably regulate conceptual deliberation, i.e., deliberation aimed at deciding which concept to adopt. As shown in previous chapters, when speakers consider which concepts to include in their repertoire, they ask not only what categories are worth representing but also what extra-representational effects those concepts can produce. Carving out the world in a way that enables us to form true beliefs about it is one such effect, but concepts offer much more. They also shape conative, affective, and evaluative attitudes — either by featuring in their content or by evoking them through their usage. They also expand their users' practical agency. Many intentional actions, such as voting or apologising, can be successfully performed only if agents possess the concepts representing them.[131] Furthermore, as my discussion of 'democracy' in Chapter 6 demonstrated, concepts can endow us with new practical perspectives from which we can perceive the

[129] Note that doxastic deliberation isn't just any deliberation about what one should believe, but the deliberation whose aim is to lead to the decision about whether to believe something.

[130] While this suggests that an epistemically rational agent must be responsive to truth-regarding considerations in doxastic deliberation, it doesn't entail the more controversial thesis known as veritism, advocated by Pritchard (2021) and Sylvan (2020), according to which the responsiveness of beliefs to truth provides the most fundamental and exhaustive explanation of what makes them epistemically rational (see Dandelet, 2024, for criticism).

[131] Cf. Anscombe (1957, pp.11-12), who goes so far as to argue that any intentional action is only possible if one possesses a concept describing it. See also Cooper (2004, pp.80-82) for criticism.

environment around us, which deeply affects what possibilities and reasons for actions are salient to us. Some concepts even shape social reality: the very existence of many social kinds depends on the existence of concepts necessary for the formation of attitudes representing them (Searle, 1995; Khalidi, 2015), and social kind concepts often causally influence what they represent (Hacking, 1999; Haslanger, 2015).

All these extra-representational effects show that concepts can be valuable not only as epistemic tools but also as practical tools. When speakers deliberate about whether to adopt a concept, they are often guided less by whether it enables them to form true beliefs and more by its practical import.[132] Accordingly, unlike doxastic deliberation, conceptual deliberation appears possible even when participants don't treat truth-regarding considerations as decisive. Accurate representation of the world is just one optional consideration to which they might be responsive. Simion's analogy between concepts and beliefs is misleading because it overlooks this crucial difference between doxastic and conceptual deliberation. Unless she provides an alternative reason to treat concepts as relevantly similar to beliefs, ELP++ lacks justification as a constraint on epistemic-moral comparisons in conceptual ethics.

While my argument is directed specifically against ELP++, it can be extended to any constraint on epistemic-moral comparisons in conceptual ethics that takes an epistemically weighted constraint on belief-formation as its model. This is an important point, given that some theorists argue practical reasons are overridden by epistemic reasons in determining what one ought to believe only until their weight reaches a relatively high threshold, after which practical reasons can take priority (Reisner, 2008; Howard, 2020). If we formulated an analogous constraint for conceptual amelioration, it would be weaker than ELP++, yet would still prioritise epistemic reasons except in cases where countervailing moral stakes are sufficiently high. Even so, such a constraint would be problematic for the same reasons as ELP++: Simion's analogy between concepts and beliefs is misleading, and giving precedence to epistemic reasons in conceptual amelioration is unjustified.

[132] Cf. Crisp (2025), who goes as far as to argue that truth-regarding considerations in conceptual engineering are derivative from practical reasons.

7.4.2. Thomasson-inspired Response

It's time to address the second possible strategy for countering the Arbitrariness Problem. Similarly to the first strategy, this strategy also posits some constraints on epistemic-moral comparisons in conceptual ethics. However, unlike the first strategy, this strategy doesn't aim to derive such constraints from general facts about the constitutive features of concepts. Instead, the strategy is inspired by Amie Thomasson's idea (2020) that conceptual deliberation is regulated by constraints derived from a specific function that an individual concept serves in a particular linguistic and extra-linguistic context. Thomasson proposes this idea in her defence of what she calls the 'pragmatic approach to conceptual ethics'.[133]

Thomasson presents the pragmatic approach as a deflationary alternative to the heavyweight metaphysical approach to conceptual ethics, according to which there are some deep metaphysical facts about the world and its structure that ground the facts about what concepts we should have (2020, pp.438-439). By contrast, the pragmatic approach interprets conceptual ethics as focusing on the functions that concepts serve for us, as determined by our shared purposes in using them, rather than by any underlying metaphysical truths (2020, pp.440-441).[134]

What Thomasson seeks to show is that even under the pragmatic approach, conceptual ethics doesn't collapse into the pursuit of unconstrained deliberations driven by arbitrary subjective aims of their interlocutors. To this end, she argues that once a concept's function is fixed, we can identify some constraints that conceptual deliberation must be responsive to in order to enable the concept to properly perform this function. Let's call these constraints 'enablement constraints'. According to Thomasson, enablement constraints come in two types. The first type is what she calls 'worldly constraints'. These are the constraints encompassing various empirical facts about the external environment in which the concept is to be used (2020, pp.451-453). The second type is what she calls 'site constraints'. These are the constraints encompassing the facts about other neighbouring concepts and practices that the concept is related to within its surrounding conceptual network (2020, pp.452-454).

[133] Note that although Thomasson proposes this idea to argue that, even under the pragmatic approach, conceptual deliberation isn't arbitrary, she doesn't deploy it to address epistemic-moral conflicts, and the Arbitrariness Problem in particular.

[134] See also Löhr (2025), who argues that the pragmatic approach is better defined not by its functional focus but by its emphasis on shared human purposes as authoritative in conceptual deliberation.

Thomasson (2020, pp.452-454) illustrates how enablement constraints can inform conceptual deliberation using the example of 'death' as examined by Gert et al. (2006). The function of 'death' is to track a property that marks the end of life and thus serves as a suitable criterion for such decisions as when to stop medical care, start funeral preparations, or put survivors' benefits into effect. But what property serves this role is constrained by worldly factors, such as the consideration that an organism can continue its circulatory and respiratory functions entirely with the artificial support of technology; the consideration that this is possible both when the organism remains conscious and when it's in a persistent vegetative state; the consideration that maintaining someone on artificial life-support systems is very expensive; etc. These considerations provide worldly constraints that restrict permissible choices in defining 'death' so that it performs its function properly. Additionally, to enable 'death' to properly perform its decision-guiding function, its definition must be well-coordinated with various neighbouring legal, social, and medical concepts like 'inheritance', 'posthumous rights', 'grief', 'afterlife', 'life', 'funeral', 'resuscitation', 'DNR order', 'medical futility', 'euthanasia', and 'organ donation'. How these concepts operate in our current linguistic and social practices places site constraints on how 'death' should be defined to perform its function properly.

Even setting aside the deflationist motivation behind Thomasson's argument, her point about enablement constraints on conceptual ethics sounds reasonable. All tools possessing a function can perform it properly only if their design is adapted to the circumstances in which they are used, and concepts are no exception. To the extent that concepts are expected to produce some effects in the world where we use them, their design should be informed by relevant empirical facts about this world. Also, since concepts don't operate in isolation but against the backdrop of other concepts and practices in their language, how they fit into their surrounding conceptual networks is also relevant to how well they perform their function.

This naturally raises the question of whether we couldn't also employ enablement constraints in epistemic-moral comparisons. Here is a proposal on how to do so: whenever an epistemic consideration and a moral consideration diverge in what conceptual choice they recommend, we should do three things. First, we should identify what function a concept serves for us. Second, we should examine what enablement constraints the concept must be responsive

to for it to properly perform that function. Third, we should prioritise that of the two conceptual choices which better accommodates these constraints.

What complicates this proposal is that what enablement constraints a concept is subject to depends on what function it serves for its users. There are countless empirical facts about the world, but only those that are relevant to how well a concept performs its function amount to its worldly constraints. Similarly, what conceptual network a concept is embedded in largely depends on its function. For example, imagine a society that neither tries to prevent death nor mourns the deceased, seeing death as a natural, unavoidable phenomenon that humans shouldn't interfere with. While this society would still want 'death' to track a property marking the end of life, they wouldn't want the expression to guide them in decisions about medical care and mourning. Consequently, the function that 'death' would serve for them would differ in some respects from its function for us. This difference would also alter what enablement constraints 'death' would be subject to in their linguistic practices. The aforementioned medical facts wouldn't be relevant as worldly constraints. Also, 'death' wouldn't be associated with concepts like 'grief', 'funeral', 'resuscitation', 'DNR order', or 'euthanasia', which would translate into a difference in its surrounding conceptual network and practices — and thus the site constraints.

Yet, if enablement constraints arise from the function a concept serves, their application is limited to conceptual deliberations in which it's presupposed that a concept should perform its current function. Many conceptual deliberations are unlike that. As explained in Chapters 1 and 2, some concepts — such as oppressive social concepts — serve objectionable functions that also taint the entire conceptual network surrounding them. Conceptual ethics, as Thomasson (2020, pp.448-449, 454-455) envisions it, should critique such functions and networks, and propose their revisions. Here, appealing to enablement constraints tied to a concept's function is unhelpful, since the relevance of these constraints is also disputed.

Our example of Conflict 2 is a clear case of conceptual deliberation involving a dispute over what function a concept *should* serve. Recall its setup: 'democracy' serves a useful function of guiding its users towards organising their public affairs in various distinctive ways. Yet, what helps 'democracy' effectively perform this function is that it's an appraisive concept that encourages ignorance about alternative decision-making systems. The responsiveness to this

worldly constraint enables ‘democracy’ to be an action-guiding concept. However, that doesn’t mean we should be responsive to it when designing the expression. Instead, the constraint prompts us to question whether ‘democracy’ should perform its action-guiding function in the first place, given that doing so involves a considerable epistemic cost. Also, the dilemma cannot be resolved by asking what additional enablement constraints the design of ‘democracy’ must accommodate to serve its action-guiding function. This would be question-begging, as the value of this function is itself a point at issue here.

Furthermore, even when all parties in conceptual deliberation agree on what function a concept should perform, there may still be lingering disputes about how well it should perform its various functional components. This dispute is pertinent when a concept serves to simultaneously produce several effects that impose conflicting constraints on its design. In such cases, a decision must be made about which effects take priority, since all of them cannot be optimally realised. Such dilemmas lurk in the examples in Conflicts 1 and 3.

In Conflict 1, the function of ‘refugee’ is to produce two extra-representational effects: first, to draw attention to those migrants legally entitled to formal privileges like asylum, protection, or non-refoulement; second, to guide its users in allocating both formal and informal privileges (like empathy or sensitive treatment) to migrants, based on moral desert. If ordinary users of the term come to believe that the term’s definition is more inclusive than it in fact is, the term will perform the second functional component better, as it will become more responsive to worldly constraints — the facts about the difficult situation of migrants who flee their countries due to imminent threats other than persecution. However, insofar as this belief is false, its spread obstructs the realisation of the second effect by disregarding site constraints — the facts about the legal definition of ‘refugee’ and related legal terms in the community’s language.

Something structurally similar is going on in our example of Conflict 3. Here, the function of ‘liberty’ is to track a political value satisfying the following two desiderata: first, it should serve as a public ideal that decision-makers can pursue in building a political system based on egalitarian principles. Second, it should be a personal value that all members of a political community can identify with based on their lived political experience. However, the two desiderata conflict because each of them reflects a different aspect of the conceptual network surrounding ‘liberty’.

On the one hand, the expression is sometimes used in a way that suggests that it picks out the political value complementary to the value denoted by 'equality'. On the other hand, it's sometimes also used interchangeably with 'freedom'. These facts are two site constraints that 'liberty' is subject to, even though it cannot be designed in a way that equally respects both.

In both examples, the expression's function comprises two divergent components, each reflecting different aspects of the worldly and linguistic circumstances in which the concept is embedded. These aspects are enablement constraints, as they affect how well the expression performs its function. Yet they exert opposing pressures on how the expression should be designed. Moreover, the tension between them mirrors the tension between epistemic and moral considerations at stake. As a result, it's unclear whether the epistemically or morally supported conceptual choice better accommodates the enablement constraints in these examples. Determining this would require comparing their relative weights — but these weights would presumably be assigned based on the very epistemic and moral considerations whose comparability is at issue. Thus, appealing to enablement constraints merely pushes us one step backward in resolving these conflicts.

7.4.3. Chang-inspired Response

There is one more strategy for countering the Arbitrariness Problem that is worth considering because it builds upon two influential ideas that Ruth Chang (1997, 2009, 2013) introduced to the general debate about cross-domain comparisons. This response comes in two steps. In the first step, the response attempts to lend plausibility to the possibility of epistemic-moral comparison by pointing out that some easy cases of epistemic-moral conflicts intuitively seem resolvable. In the second step, the response elaborates on how cross-domain comparative reasoning can also resolve some difficult cases of epistemic-moral conflicts by appealing to the existence of will-based reasons. Let me go through these steps in order.

The first step of the response is a version of the argument from nominal–notable comparisons that Chang (1997, p.32) develops to defend the comparability of prudential and moral considerations, and that has recently been adopted by Meylan (2021, pp.208-211) to argue for the

comparability of practical and epistemic considerations in belief-formation.[135] While the argument hasn't yet been adapted to conceptual ethics, it isn't difficult to see what such an adaptation might look like. In the context of the present discussion, the argument involves showing that there are easy cases of epistemic-moral conflicts in conceptual ethics where it intuitively seems clear which option is, all things considered, preferable. These are the cases in which the compared considerations are such that one consideration is only of marginal importance relative to its own standard, while the second consideration is of notable importance relative to its own standard. In such cases, we are strongly inclined to conclude that the option supported by the second consideration is, all things considered, preferable over the option supported by the first consideration.

To illustrate, imagine a modification of our example of Conflict 1 in which the country's citizens are generally hostile towards all newcomers, regardless of whether they believe these newcomers qualify as refugees. The only difference in how they treat those they apply 'refugee' to is that, out of formality, they give them a smile and a cookie upon first seeing them. Thus, even if we deceive the citizens about the definition of 'refugee', the only moral payoff will be that more newcomers will receive nicer, superficial treatment from citizens upon their first encounter with them. Otherwise, the citizens will treat these newcomers just as unfriendly and discriminatorily as they would if they didn't consider them refugees. Let's call the modified scenario 'Cookie'. In Cookie, there is a strong intuition that the moral benefit of more newcomers initially receiving a smile and a cookie isn't, all things considered, worth the epistemic cost of deceiving the citizens about who qualifies as a refugee in their society. If this is true, there presumably must be some overarching standard of epistemic-moral comparison that makes this verdict true — or so the argument goes.

The first step of the response offered some hope that we might resolve the easy cases of epistemic-moral conflicts by cross-domain comparative reasoning after all. But what about cases such as Conflict 1, Conflict 2, and Conflict 3, where there is a tension between a notable epistemic consideration and a notable moral consideration? Even if there really is an overarching standard of epistemic-moral comparison, how can we explain that, unlike in the easy cases where it's clear

[135] Many other philosophers have used this argument to defend the comparability between moral and prudential considerations. Dorsey (2016, p.119), Crisp (2006, p.132), and Parfit (2011, p.132) are a few representative examples.

what this standard recommends, no obvious resolution of these cases follows from it? And how can the given standard guide us in resolving these cases? These are the questions that the response addresses in its second step. In doing so, the response takes inspiration from Chang's proposal (2009, 2013) that the exercise of will can play an important role in rational deliberation because it can create new reasons when all the other reasons bearing on a choice situation underdetermine which option to choose.[136] She uses the following example to motivate this claim:

> 'Suppose you are faced with a choice between a career as a philosopher and one as a trapeze artist. You have investigated each career from every angle, vividly imagined yourself writing philosophy articles and swinging under the big top, carefully considered and re-considered the reasons for and against each career, thought long and hard about how the reasons for and against each relate, sought advice from people whose judgement you respect, and so on. Suppose that, as a result of careful and thorough investigation, you come to believe that, all things considered, the reasons for and against each career have run out.' (2009, pp.249-250).

Chang argues that although the reasons given to you in this situation underdetermine whether you should become a philosopher or a trapeze artist, you can change this through an act of will by deciding that one of the two careers is, in some respect, of greater personal significance to who you are. The fact that you have made this decision will then be a new will-based reason to pursue the given career that will make a normative difference to the situation by tipping the overall normative balance in its favour. You will then have all-things-considered justification for pursuing this career, which will be grounded in how the exercise of your will has reshaped your rational identity in the course of deliberation.

It might be suggested that will-based reasons can also be a key to resolving conflicts between notable epistemic and moral considerations. The difficulty in resolving them might perhaps be explained by the hypothesis that, in these conflicts, all the epistemic and moral reasons

[136] Several proposals adjacent to Chang's can be found in the literature. For example, Holton (2009, ch.3) argues that the will can help us choose what to do when we don't know how to make a comparative judgement about alternative options. Raz (1999, ch.3), meanwhile, suggests *our desires* play a central role in practical deliberation whenever we choose between several rationally eligible but incomparable alternatives, while maintaining that they can do so without themselves creating new reasons (see Chang 2012, pp.121-123, for discussion).

for preferring one option over the other run out, just as they did in Chang's example. That is, the overarching standard for epistemic-moral comparison underdetermines which of the two considerations is, all things considered, more important. However, conceptual ethicists can perhaps change this directly during their deliberation by deciding that one of the two options is of greater significance to the collective identity of the linguistic community that will use the negotiated concept. Just like in Chang's example, this decision might reshape the community's rational identity by creating a new will-based reason in favour of this option. Once the decision is made, the overall balance of reasons gets tipped, and the overarching standard recommends that, all things considered, this option is the preferable one for the community to take.

I find both steps of this response unconvincing. Let's start backward by assuming the response succeeds in its first step. Even then, the second step remains problematic. The paradigmatic choice situations Chang uses to support will-based reasons involve deliberation over personal life choices affecting only an individual agent. In contrast, the expected outcome of conceptual deliberation is a product that serves the entire linguistic community of its users. Therefore, these deliberations involve multiple agents, all of whom arguably have a say in what their concept should be like. Yet, we cannot simply presuppose that there is a wide agreement among agents about which option in epistemic-moral conflicts matters most for their collective identity. More often than not, a linguistic community may be deeply divided on this substantive point: its members may split into two similarly sizeable groups, one wanting the concept to perform better epistemically; the other prioritising its moral performance.

The appeal to will-based reasons here is far from straightforward. Presumably, we must appeal to will-based reasons deriving from the collective decision of the linguistic community as a group agent. Several authors argue that a community qualifies as a group agent by forming collective attitudes that aren't mere aggregations of individual attitudes — such as by simple majority — but relate to them in more complex ways.[137] For instance, the community may have a hierarchical organisation allowing some members greater influence in forming collective attitudes. If so, the community could be a group agent capable of deciding whether epistemic or moral

[137] See especially List and Pettit (2006; 2011, ch.3), Pettit and Schweikard (2006), and List (2014). These authors, in fact, argue for a stronger claim: to meet certain rationality requirements necessary for qualifying as a group agent, a community must possess a suitable organisational structure that renders its collective attitudes irreducible to a mere aggregation of individual attitudes.

aspects are more important for its group identity, even if its members widely disagree on the question. It can do so in virtue of a few members of the community having the institutional power to decide in favour of one option on behalf of the whole community. Still, it's unclear whether such a decision has the *normative* power to create a new will-based reason that's binding for all other members of the community.

Plausibly enough, such a decision could create a new will-based reason for those members of the community to whom it can be justified on the grounds that it accurately reflects their own view of the concept's role in group identity. However, for members who feel alienated from that conception, the decision would have normative force only if the decision-making mechanism through which it was generated could either be independently justified to them or was one to which they had previously consented. Otherwise, the decision would seem more like an attempt by some subset of members to impose a specific conceptual choice on dissenting members, rather than a resolution to the underlying epistemic-moral conflict. As such, it would materialise the concerns expressed about the risk of conceptual engineering degenerating into a coercive project that illegitimately exerts control over others' conceptual resources (Queloz & Bieber, 2022; Shields, 2021).

Accordingly, for collective decisions to have normative power in epistemic-moral conflicts faced by linguistic communities deeply divided on the significance of a concept for their group identity, these communities would need to establish a decision-making mechanism that enjoys sufficient uptake among their members to count as legitimate. This is a substantive task. Considering that our linguistic communities have thus far invested little effort in developing reasonable collective arrangements for regulating normative conceptual disagreements and justifying them to their members, I doubt such mechanisms are already in place. Therefore, the second step of the Chang-inspired response currently seems unfeasible as a general method for resolving epistemic-moral conflicts.

But couldn't we say that the second step of the Chang-inspired response can still address cases in which a linguistic community widely agrees on the significance of a concept for its group identity, while in other cases, it at least serves as the source of motivation for future action aimed at resolving epistemic-moral conflicts through comparative reasoning? For this to be the case, the

first step of the response must be successful — it must establish that epistemic and moral considerations are genuinely comparable.

Unfortunately, even this step can be called into question. As Sagdahl (2022, ch.5) points out, we can account for our intuitions about nominal-notable comparisons even without appealing to the overarching standard of comparison. Sagdahl makes this point in relation to prudential-moral conflicts, but I will now show that it also extends to epistemic-moral conflicts. Consider again the easy case Cookie. Recall that we are strongly disposed to judge that the option supported by the notable epistemic consideration, all things considered, outweighs the option supported by the nominal moral consideration. Still, this alone doesn't testify to cross-domain comparability because the term 'all things considered' can be plausibly interpreted in this judgement not as meaning 'when assessed by weighing two standards on a balance', but as meaning 'when independently assessed by each of the two standards'. Let's call the former interpretation the 'comparative interpretation' and the latter interpretation the 'convergence interpretation'. While the former interpretation invites us to weigh epistemic and moral considerations on a balance, the latter interpretation invites us to consider whether, despite the conflict between an individual epistemic consideration and an individual moral consideration, the epistemic standard and the moral standard might issue the same overall verdict about what option to take.

Why do I find it plausible that 'all things considered' in our judgement about Cookie should be understood through the convergent interpretation? This is because it isn't difficult to argue that both the epistemic and moral standards recommend the option supported by the notable epistemic consideration. It goes without saying that the epistemic standard recommends it. But, plausibly enough, even the moral standard does. It would be surprising if morality instead recommended spreading lies about substantive topics such as who qualifies as a refugee merely to ensure newcomers receive a smile and a cookie. After all, there seems to be a non-negligible moral value in living a truthful life, which isn't worth sacrificing for trivial moral gains such as this one.[138]

[138] See especially Murdoch (1970, pp.30-41, pp.64-65; 1992, p.406), who argues that there is a tight connection between the morality of an agent and their capacity to pay attention to reality (see also Mason, 2023, for discussion). Additionally, some authors in the ethics of belief literature have expressed a related idea — namely, that compliance with epistemic standards is often practically desirable (e.g., Heil, 1992; Maguire & Woods, 2020, p.235).

Moreover, a compelling case for the convergence between moral and epistemic standards can be made even in the reversed cases, where there is a conflict between a notable moral consideration and a nominal epistemic consideration. As an example, imagine a community that performs exceptionally well with regard to animal welfare. Animals living alongside this community receive maximally sensitive and fair treatment. However, suppose we know the community treats animals so well only because the term 'animal' historically acquired spiritual connotations in their social practices, promoting the irrational belief that animals are the messengers of God. This belief has the status of a benign falsehood, for it doesn't anyhow hinder the community's inquiry about the world. That is, all the other beliefs of the community's members about the world are just as accurate as they would be if they didn't have the given spiritual belief. Hence, the epistemic loss this belief brings about is minor, while the moral benefit it has for the world is substantive.

The following question arises: should we try to convince the community to strip the term 'animal' of spiritual associations, thereby encouraging them to abandon their irrational spiritual belief about animals? Morality clearly recommends against doing so, given the significant moral benefit for animals this belief brings about. And, upon reflection, there is good reason to think the epistemic standard doesn't recommend this course of action either after weighing up all the relevant epistemic considerations.

For sure, abandoning the spiritual belief will present a small epistemic gain for the community. However, we must also consider potential downstream epistemic effects of the abandonment. Without the spiritual belief, the community's treatment of animals will likely deteriorate, aligning with the poor treatment of animals seen globally. There is a risk that this will lead to significant epistemic losses for the community, such as self-deception and ignorance. We commonly see that when humans cause substantive moral harm to others, they struggle to admit it to themselves. Instead, they tend to close their eyes to the harmfulness of their actions or downplay and rationalise it. The risk that the community under consideration will behave similarly should be taken seriously, given the widespread human tendency to overlook and explain away their mistreatment of animals. Considering all this, it seems plausible that these potential epistemic

losses override the epistemic gain of eliminating one benign falsehood.[139] This illustrates how both epistemic and moral standards can independently recommend the option supported by a notable moral consideration over the one supported by a nominal epistemic consideration.

As we can see, our judgements about which option is, all things considered, preferable in the nominal-notable cases of epistemic-moral conflicts can be plausibly interpreted as being motivated by the convergent interpretation. This considerably weakens the force of the argument from nominal-notable comparisons, since, under the convergent interpretation, assessing how an expression should be designed all things considered doesn't require comparing the two domains. Instead, it merely involves conducting two separate intra-domain comparisons: one of the relative weights of all relevant epistemic considerations, and another of the relative weights of all relevant moral considerations. No overarching standard of cross-domain comparison is presupposed here.

7.5. The Integrative Approach to Epistemic-Moral Conflicts

We have explored three strategies for addressing the Arbitrariness Problem, each attempting to show how it might be possible to non-arbitrarily compare epistemic and moral considerations in conceptual deliberation. While these strategies turned out to be unsuccessful, I now want to argue that the Arbitrariness Problem is still surmountable by other means. That is, in the face of epistemic-moral conflicts, conceptual ethicists aren't forced to either arbitrarily choose one of the options under consideration or leave their deliberation about conceptual design incomplete. However, as I see it, the key to surmounting the problem lies not in finding a non-arbitrary method for cross-domain comparison, but in adopting the so-called *integrative approach* to epistemic-moral conflicts.

As seen in the easy cases of epistemic-moral conflict, epistemic and moral standards can converge in their overall verdicts on how an expression should be designed, despite tensions between individual considerations. Admittedly, there are also hard cases of epistemic-moral

[139] My opponent might tweak the scenario so that the agents are psychologically sophisticated enough to commit moral harm without self-deception or ignorance. But once the agents become so different from us, it's unclear whether we should still trust our intuitions about them. Since most humans are prone to such distortions when committing substantive moral harms, our judgement about what the community should, all things considered, do may be skewed by projecting our own tendencies onto them. Moreover, given that conceptual ethics concerns which representational devices we, as actual human agents, ought to use, it's unclear how relevant such hypothetical, dissimilar agents really are.

conflict, such as the three examples discussed in §7.2, whereby the epistemic and moral standards are likely to diverge in their overall verdicts. Yet, the guiding idea behind the integrative approach is that such divergence often isn't fixed and immutable but results from contingent psychological and social factors within human control. Accordingly, the aim of the integrative approach is to explore the possibility of creating favourable conditions for the convergence between epistemic and moral standards on how an expression should be designed by appropriately changing how we engage with language. To the extent that this possibility proves realistic, the approach guides us towards integrating epistemic and moral standards by implementing these changes and choosing such a design of the expression that is recommended by both standards once the changes are in place.

The integrative approach has so far been advocated by several philosophers in relation to moral-prudential conflicts. Scheffler (1992, chs.7-8; 2008) extensively argues that the relationship between morality and self-interest is one of potential convergence, as the extent to which the two conflict depends on an agent's psychological makeup as well as external societal conditions. On his view, human practices and social institutions 'help to determine the prevalence, both of the motivational patterns that lead people to try to shape their projects in such a way as to satisfy moral requirements, and of the factors that have the potential to frustrate such attempts' (1992, p.131). Specifically, he thinks societies can integrate morality and prudence by becoming well-ordered and just — fostering the development of moral sensibilities and a sense of responsibility for the community — since in such societies, moral behaviour becomes more attractive and easier to live up to than in unjust and lawless ones (1992, pp.138-145). Similarly, Nagel (1986, p.206) argues that 'the most important task of political thought and action is to arrange the world so that everyone can live a good life without doing wrong, injuring others, benefiting unfairly from their misfortune, and so forth'. More recently, Wallace (2006, pp.133-135) and Sagdahl (2022, pp.221-223) have independently argued that integration offers a promising way to incorporate both morality and prudence into practical deliberation, presenting it as an alternative to weighing the two values against each other.

In what follows, I draw methodological inspiration from these authors, aiming to show that, although originally developed for moral-prudential conflicts, the integrative approach holds

significant — yet underexplored — potential for addressing challenging cases of epistemic-moral conflict in conceptual ethics, such as Conflicts 1-3.

7.5.1. What Kind of Tools Are Linguistic Expressions?

In each example of Conflicts 1-3, we encounter a situation where a notable epistemic consideration about an expression's design stands against a notable moral consideration. Attending to these conflicts through the integrative approach involves asking under what conditions (if any) we could achieve that one of these considerations is overridden within its own domain, i.e., that the standard governing that domain overall favours the same conceptual choice as the standard governing the other domain. I want to make two general points about what kind of tools linguistic expressions are, factoring in which will allow us to see that the prospects of bringing about such convergence in the overall verdicts issued by the epistemic and moral standards are brighter than they might initially seem.

My first point is underpinned by reasoning similar to the one Thomasson (2020, pp.452-454) uses to support her point about site constraints. Linguistic expressions are *holistic tools* that cannot effectively operate in isolation but only in coordination with a broader conceptual network of neighbouring expressions they are enmeshed in. Therefore, it would be illusory to think a single expression can produce any substantive epistemic or moral effect on its users even if the surrounding conceptual network doesn't cooperate.

To illustrate this point, consider Haslanger's (2000, p.42) famous proposal to redefine gender expressions so that 'woman' picks out individuals systematically subordinated based on observed or imagined female bodily features, while 'man' picks out individuals systematically privileged based on observed or imagined male bodily features. This example is helpful because Haslanger argues that we benefit both epistemically and morally from these definitions. By explicitly defining gender expressions in terms of subordination and privilege, they increase awareness of gender oppression and better serve critical feminist inquiry. These epistemic benefits also yield moral benefits by supporting feminist efforts to combat gender oppression by empowering those facing it (Haslanger, 2000, pp.46-48).[140]

[140] See also Barnes (2017, pp.2421-2422), who interprets Haslanger's proposal as aiming at these epistemic and moral effects.

It seems difficult to deny, though, that the extent to which the revisions of 'man' and 'woman' can successfully deliver these epistemic and moral benefits is constrained by the conceptual network in which they are embedded. If surrounding expressions promote ignorance of gender-based oppression, these revisions alone are unlikely to significantly increase awareness or drive substantial moral progress. Consider communities with deep misconceptions about related terms. They misconstrue 'female' and 'male' as reflecting biological differences that justify privileging men, and misunderstand 'feminism' and 'oppression', thinking the former seeks special privileges for women and the latter applies to any harm, regardless of power or systemic patterns. Even if such communities adopt Haslanger's redefinitions of gender expressions, these are unlikely to have a significant illuminating or empowering effect unless they also revise their understanding of surrounding expressions.

This example illustrates that, whatever notable epistemic and moral differences expressions make to the lives of their users, they do so only to the extent to which their surrounding expressions enable it (cf. Marchiori, 2025, pp.13-15). An expression may initially seem impactful under a given design, but its surrounding network may, on closer inspection, hinder its impact. If the impact is negative, this may weaken the considerations against the design; if it's positive, a key question is whether large-scale revision of the network is feasible. If not, our expectations for the design should be moderated accordingly. When integrating epistemic and moral domains, we must pay attention to these considerations, as they can significantly shift the overall balance of reasons within each of them.

Next, the second general point I want to make about expressions is this: when we design expressions, we plan them as representational devices with *stable applicability*. By this, I mean that we don't design expressions as local and temporary tools whose applicability is confined to a specific user setting, but as tools with cross-temporal and counterfactual applicability. This point becomes clearer when we consider what our practice of defining expressions looks like. Suppose, for example, we stipulatively define 'person' as follows:

P: 'Person' means 'a being that is rational and self-aware'.

If this stipulation successfully fixes the meaning of 'person', P is typically expected to specify what 'person' applies to — not just in the actual world and present time, but also in a range of

possible worlds and times. If P fell short by leaving the extension of 'person' indeterminate beyond the actual world and the present time, that alone would suggest it's an unsatisfactory definition. There are two indications that this expectation underlies our defining practices.

First, we routinely use expressions in counterfactual statements, in statements about the past and future, and in contexts beyond the immediate user setting. We do so confidently and unreflectively, without first checking whether these situations fall within the expression's application scope. This indicates that we, by default, expect their scope of application to be cross-world and cross-temporal. Second, we commonly test definitions like P for adequacy using the method of cases.[141] For instance, we ask whether there are hypothetical cases where an entity we aren't disposed to call 'a person' is rational and self-aware, or vice versa. These cases serve as counterexamples to P being the definition of 'person' precisely because adequate definitions are expected to govern their application across counterfactual circumstances such cases help us imagine.

This point calls for explanation: why do we define expressions with the expectation that their scope of application extends beyond the actual world and the present time? There are arguably both epistemic and moral benefits to stable applicability. If expressions were defined solely to represent users' immediate environment, leaving their extensions in other worlds and times indeterminate, we couldn't use them in counterfactual or cross-temporal reasoning. This would clearly impoverish us epistemically, as much of our understanding of the behaviour of objects and causal relationships between them depends on engaging in counterfactual, predictive, and retrospective thoughts. Without cross-modal and cross-temporal applicability, our ability to entertain and communicate such thoughts would be severely hindered.

Further, the loss of the above ability would likely also limit us in moral deliberation. Our ability to make good moral decisions becomes more reliable to the extent that we deliberate in accordance with reason. Such reason-governed deliberation is guided by the question of which among several alternative courses of action is supported by the strongest balance of reasons. This future-oriented question requires both counterfactual and predictive reasoning. In answering it, an

[141] See, notably, Weinberg et al. (2001), Deutsch (2015), and Machery (2017) for discussions of how exactly the method of cases works.

agent imagines possible future courses of action, predicts their likely consequences, and chooses the course they most strongly endorse. Moreover, in predictive reasoning, the ability to entertain and communicate thoughts about past consequences of similar actions is useful for extrapolation. Thus, our ability to reason counterfactually and cross-temporally is also important in our moral pursuits.

As we can see, having at our disposal expressions that are stably applicable and thus useful for counterfactual and cross-temporal reasoning is both morally and epistemically advantageous. Additionally, stable applicability also enhances the potential for using expressions across various circumstances. If expressions are defined so that actual and present speakers can use them to represent cross-temporal and counterfactual situations, this also significantly boosts their representational utility for the speakers located in these situations, who can then use the expressions to describe their immediate environment. To illustrate, a consequence of the extension of 'water', 'addiction' or 'money' being determinate in a range of counterfactual, past and future situations is that, even when we are located in these situations, we can make use of these expressions to represent our immediate surroundings. Of course, this doesn't imply that their extensions cannot be empty in these situations. There might be worlds in which there is no water, addiction or money. But even the speakers situated in such worlds can help themselves to these expressions when describing what is absent in their immediate environment.

Accordingly, our defining practices reveal that we design expressions as stably applicable tools when it comes to their representational utility. Of course, whether stable applicability positively contributes to an expression's moral or epistemic value is a further question. The answer ultimately depends on what extra-representational effects the expression produces in different situations as a result of being applicable in them. If an expression is defined in a way that consistently produces problematic extra-representational effects, such as the formation of false beliefs or oppression, it may be better to limit its applicability. Further, even if an expression's applicability currently yields valuable extra-representational effects, it may not be able to reproduce them in many other situations or may only do so without retaining their positive value.[142]

[142] An example of the former possibly was discussed earlier: even if Haslanger's ameliorative definitions of gender terms are empowering in some societies, they may fail in others where surrounding expressions are shaped by misconceptions. As for a potential example of the latter possibility, it might be argued that while the expression 'economic growth' once played a positive action-guiding role in post-war recovery, it has become harmful in modern

In such cases, the mere fact that an expression has representational utility for speakers in those situations presumably doesn't make it any better.

Suppose, however, that we design a concept that is stably applicable in three ways: (i) it applies across diverse cross-world and cross-temporal situations; (ii) it consistently produces extra-representational effects there; and (iii) these effects consistently have positive value. This stable applicability — call it 'well-rounded applicability' — is clearly a positive feature that enhances the concept's moral or epistemic value by boosting its extra-representational performance. Whatever epistemic or moral goods this concept delivers, its stable applicability amplifies them by making these goods accessible across more situations. The key takeaway is that designing concepts as stably applicable tools is beneficial, provided that stability applies not only to their representational potential but also to the positive value of their extra-representational effects.

7.5.2. Exploring the Potential of Epistemic-Moral Integration

With the above two points under our belts, let's now revisit the examples of Conflicts 1-3 using the integrative approach. That is, let's examine the prospect of integrating epistemic and moral standards in such a way that they recommend the same conceptual choice in these examples, while factoring in considerations about expressions' surrounding conceptual network — hereafter referred to as *holistic considerations* — and considerations about the stability of expressions' positive extra-representational effects — hereafter referred to as *stability considerations*. I intend to show that, in light of these considerations, such integration is entirely possible — provided that we cultivate the right kinds of practices and dispositions among expression users.

Consider first the example of Conflict 1. A key question here is whether there's an alternative way to motivate citizens to treat newcomers as welcomingly as they treat refugees, without deceiving them about the correct application of 'refugee'. One such alternative is the *discursive strategy*, which has two components: first, cultivating a healthy culture of rational deliberation among citizens about metalinguistic questions, such as what expressions should mean to serve as the most valuable representational tools; and second, drawing upon this culture to

capitalist societies by promoting unsustainable resource use and a culture of productivism, no longer needed to improve living standards (cf. Saito, 2024).

initiate a public metalinguistic debate about what 'refugee' should mean.[143] In this debate, citizens can be presented with rational arguments showing that the situation of those fleeing political persecution is relevantly similar to those fleeing for other reasons, such as civil war, invasion, or climate disaster. These arguments can then be used to convince citizens that there is no good reason to include the former group under 'refugee' while excluding the latter — especially if this leads to different treatment of the two groups.

What's remarkable about the discursive strategy is that, if successful, its outcomes appear not only epistemically but also morally preferable to those of the deceptive strategy. This is because it promotes deeper revisions in citizens' conceptual and belief network, beyond simply changing what they believe 'refugee' refers to. As a result, the moral gains it produces are more stable. Specifically, the strategy seeks to make citizens more normatively reflective about what 'refugee' *should* include by encouraging them to recognise relevant similarities between concepts like 'persecution', 'civil war', 'climate disaster', and 'invasion'. Recognising these similarities is likely to lead to more sensitive treatment of newcomers. This effect, however, will be more *stable* than if it were achieved through deception. After all, citizens will improve their treatment of newcomers in response to rationally forming a normative belief about what deserves to be called 'refugee'.[144] The risk of them abandoning this belief after a short period is far lower than the risk of abandoning a false belief acquired through deception: while abandoning the former requires being exposed to a compelling counterargument, the latter would be abandoned as soon as the deception is uncovered.[145]

Hence, in light of holistic and stability considerations, there is strong motivation to apply the discursive strategy in this example. By fostering a healthy discursive culture that enables citizens to engage in public debate over what 'refugee' should mean, we restructure the example's choice situation so that lying to them about the meaning of 'refugee' is no longer morally preferable,

[143] See Sterken (2020) and Cantalamessa (2021) for discussion of how such metalinguistic debates can be initiated by temporary disruptions in communicative patterns that presuppose a term's extant meaning.
[144] Recognising the similarities between these expressions may help citizens adopt a more morally apt perspective towards newcomers, even if they remain unsure whether to call them 'refugees' (cf. Sliwa, 2024, p.130).
[145] Additionally, Kitsik (2023a, 2023b) would argue that the deceptive strategy is objectionable on different grounds: it violates citizens' epistemic autonomy by shaping their beliefs without giving them the opportunity for rational deliberation.

all things considered. In doing so, we create a measure of fit between the verdicts of moral and epistemic standards.

Next, consider the example of Conflict 2. When assessing this example through a moral lens, it's likewise crucial to consider under what conditions the action-guiding effects associated with 'democracy' can be stably produced. This is an open question, as the example's setup only specifies that these effects are easiest to produce when 'democracy' is an appraisive expression. However, I suspect that designing 'democracy' as an appraisive expression may not be the most effective way to stably produce these effects over time.

To explain, recall the epistemic reason against 'democracy' being appraisive in this case: its appraisiveness promotes ignorance of alternative decision-making systems. At first glance, this ignorance might seem morally unproblematic so long as democracy is the morally best decision-making system. However, once stability considerations are factored in, this view seems short-sighted. If 'democracy' prevents users from critically evaluating it or considering alternatives, it promotes blind allegiance. Yet, action-guiding force based on blind allegiance is typically less stable over time than that based on critical judgement following reflective comparison. This shouldn't come as a surprise. If 'democracy' elicits strong positive associations that promote allegiance, but its proponents don't understand what justifies these associations, there's no guarantee they won't, after a while, change their minds and shift their allegiance to one of its alternatives. If that happens, 'democracy' will be replaced by a different expression in its current action-guiding role.

This shows that while 'democracy' can easily gain immense action-guiding force as an appraisive concept, it can just as easily lose it. This should give us pause and prompt reflection on whether we couldn't develop educational initiatives — like those in Chapter 6 — designed to teach people to engage more critically with 'democracy' in public life. If successful, such initiatives could still allow speakers to be guided by the reasons 'democracy' offers for organising public affairs, while also encouraging them to question why these reasons should outweigh those offered by alternative political concepts. Although this strategy diminishes the status of 'democracy' as an expression tracking the ultimate political ideal, and thus weakens its action-guiding force, it gives the expression a more grounded role in its conceptual network — one from which it can guide

action more moderately, yet more *stably*. Even from a moral perspective, this strategy is preferable to the uncritical allegiance to 'democracy' that currently prevails. It thus offers a promising way to integrate epistemic and moral standards.

Last but not least, consider the example of Conflict 3. The integrative approach seems applicable even here. In fact, I take this example to be a good illustration of a situation in which holistic considerations reveal that a substantive negative effect associated with an expression can be easily prevented, provided that the concept is properly situated within its conceptual network.

In more concrete terms, while the tension between 'liberty' and 'equality' can embody an epistemic flaw in speakers' conceptual network, whether it does depends on whether it's opaque or transparent to them. As noted in §7.1, the presence of this tension doesn't necessarily mean speakers hold incoherent or disunified beliefs about the two concepts. It simply means 'liberty' and 'equality' often cannot be correctly applied to the same object, as the object may be unable to jointly manifest the values each concept represents. Incoherence arises only if speakers fail to recognise this constraint, mistakenly assuming societies and political systems can always manifest both values simultaneously. However, if speakers internalise both concepts within their conceptual network in a way that makes the tension salient and intelligible, this risk can be avoided.

But doesn't it then follow that, as long as the tension is transparent to speakers, the epistemic standard merely *permits* — but doesn't *recommend* — retaining it, unlike the moral standard? If so, full epistemic-moral integration still wouldn't be achieved. However, there is reason to think the epistemic standard *does* recommend retaining the tension — provided it's recognised. To explain, the recognition of this tension can be seen not only as a means of fixing or preventing an epistemic flaw, but also as a source of positive epistemic enrichment for speakers. It deepens speakers' understanding of the relationship between the two values, sharpens awareness of their limits, and helps anticipate political challenges arising from their conflict. Such insight fosters a more realistic grasp of political life than if the tension were obscured. Moreover, since the tension between liberty and equality appears to be a persistent and perhaps unavoidable feature of human political interaction, rather than a passing anomaly, the insights it yields carry enduring epistemic value.

Therefore, the proper integrative strategy in this case is to foreground the tension in speakers' minds, increasing collective awareness of it. Whereas the integrative strategy in the previous two examples involved aligning the moral standard with the epistemic, here the reverse occurs: by raising awareness of the tension between 'liberty' and 'equality', we restructure the choice situation so the epistemic standard aligns with the moral one — overall recommending that the tension be retained.

7.6. Conclusion

I conclude that, although the prospects of resolving epistemic-moral conflicts through cross-domain comparison seem bleak, the integrative approach holds underexplored potential worth further investigation. We have seen that assessing linguistic expressions as holistic tools with stable applicability reveals promising ways to restructure choice situations so that epistemic and moral standards can be brought into harmony.

Each example offered practical guidance on the patterns of engagement with language that users should collectively cultivate to realise this potential. The first conflict highlighted the importance of fostering a healthy discursive culture that encourages metalinguistic debate and responsiveness to rational argument. The second stressed the need for speakers to learn to be guided by expressions they critically engage with, rather than blindly follow. The third showed that we shouldn't eliminate tensions in our conceptual repertoires at any cost, but instead attend to the insights they reveal and raise awareness of them (cf. Queloz, 2024a; 2025, chs.5-6). Cultivating such engagement aims to build conceptual repertoires that don't force — or at least minimise — trade-offs between epistemic and moral values, but instead allow both to be upheld together.

Two final remarks. First, the feasibility of the integrative approach depends on whether speakers can restructure their linguistic engagement in the aforementioned ways. This, in turn, depends on whether they have sufficient collective control over their linguistic practices. While speakers are unlikely to have immediate collective control over these practices, it nonetheless seems reasonable that they possess sufficient long-range collective control: if many members of a linguistic community strive to change these practices through a series of possibly interrupted

actions over a significant period of time, they are likely to succeed.[146] This level of control seems sufficient to make epistemic-moral integration in conceptual ethics a feasible — though challenging and long-term — project.

Second, much has been written about the role moral considerations can play in epistemic justification. Some argue that value judgements, including moral ones, may play an indirect evidential role in scientific inquiry (e.g., Longino, 1990; Anderson, 2004; Douglas, 2009), while others defend moral encroachment — the view that moral factors can directly affect whether a belief is epistemically justified (e.g., Basu, 2019; Basu & Schroeder, 2019; Bolinger, 2018; Fritz, 2017). None of the integration cases I have discussed here presupposes such dependence; in each, moral and epistemic standards converge for independent reasons. While this makes my argument less theoretically burdened by contested claims, I suspect that if moral considerations do contribute to epistemic justification, the case for epistemic-moral integration in conceptual ethics is even stronger. Exploring this possibility is a promising direction for future research.

[146] See Alston (1988, pp.274-278) and Koch (2021b, pp.337-339) for discussions of the distinction between immediate and long-range control in belief formation and semantic change, respectively.

Conclusion

This brings us to the end of this dissertation. To recap, the first part argued that, in assessing referential expressions, we must attend to the functions they serve for a user group, as these functions reveal which representational and extra-representational effects are central to the group's motivation for using them. This then allows us to understand which interests, evaluative outlooks, and desires underlie the group's linguistic practices — and thereby, why the expressions matter to its members. As we have seen, such understanding is valuable for the responsible pursuit of cross-linguistic conceptual ethics and for the analysis and evaluation of thick ethical concepts. The second part urged caution against discarding expressions whose reference is mismatched, including cases where the mismatch arises from semantic indeterminacy. It suggested we first consider how such expressions might enrich our conceptual repertoire and how they might be ameliorated. The third part argued that when conceptual assessment is pulled between competing epistemic and moral considerations, we shouldn't be confined to comparative reasoning but should instead consider how the tension between these considerations might be mitigated through the restructuring of our linguistic practices. All in all, I hope that the dissertation serves as a compass in the intricate process of assessing referential expressions, offering useful practical guidance on how conceptual ethics ought to be pursued or, at the very least, enabling us to view this process from a range of novel perspectives.

References

Andler, M. S. (2017). Gender Identity and Exclusion: A Reply to Jenkins. *Ethics*, 127(4), 883-895.

Al Jazeera. (2017, July 16). *Thousands Rally Against Court Reforms in Poland*. Al Jazeera. Retrieved March 26, 2025, from https://www.aljazeera.com/news/2017/7/16/thousands-rally-against-court-reforms-in-poland

Alston, W. P. (1988). The Deontological Conception of Epistemic Justification. *Philosophical Perspectives*, 2, 257-299.

Alexopoulos, G. (2017). *Illness and Inhumanity in Stalin's Gulag*. Yale University Press.

Allen, S. R. (2021). Kinds Behaving Badly: Intentional Action and Interactive Kinds. *Synthese*, 198(12), 2927-2956.

Anderson, E. (1997). Practical Reason and Incommensurable Goods. In R. Chang (Ed.), *Incommensurability, Incomparability, and Practical Reason* (pp.90-109). Harvard University Press.

Anderson, E. (2004). Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce. *Hypatia*, 19(1), 1-24.

Anderson, P. C., & Peña-Chocarro, L. (2014). Diversity in Harvesting Techniques. In A. van Gijn, J. Whittaker, & P. C. Anderson (Eds.), *Explaining and Exploring Diversity in Agricultural Technology* (pp.85-132). Oxbow Books.

Andersson, H. (2016a). Parity and Comparability: A Concern Regarding Chang's Chaining Argument. *Ethical Theory and Moral Practice*, 19(1), 245-253.

Andersson, H. (2016b). Vagueness and Goodness Simpliciter. *Ratio*, 29(4), 378-394.

Andow, J. (2020). Fully Experimental Conceptual Engineering. *Inquiry*, 1-27.

Anscombe, G. E. M. (1957). *Intention*. Harvard University Press.

Appiah, K. A. (1996). Race, Culture, Identity: Misunderstood Connections. *The Tanner Lectures on Human Values*, 17, 51-136.

Aristotle. (c. 350 BCE/1960). *Posterior Analytics, Topica* (H. Tredennick & E. S. Forster, Trans.). Harvard University Press.

Austin, M. (2018). *Humility and Human Flourishing: A Study in Analytic Moral Theology*. Oxford University Press.

Baker, D. (2018). Skepticism About Ought Simpliciter. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol.13, pp.230-252). Oxford University Press.

Baker, L. R. (2007). *The Metaphysics of Everyday Life*. Cambridge University Press.

Bagehot, W. (1867/1963). *The English Constitution*. Cornell University Press.

Barber, B. (1984). *Strong Democracy: Participatory Politics for a New Age*. University of California Press.

Barnes, E. (2014). XV-Going Beyond the Fundamental: Feminism in Contemporary Metaphysics. *Proceedings of the Aristotelian Society*, 114(3_pt_3), 335-351.

Barnes, E. (2017). Realism and Social Structure. *Philosophical Studies*, 174(10), 2417-2433.

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

Basu, R., & Schroeder, M. (2019). Doxastic Wronging. In B. Kim & M. McGrath (Eds.), *Pragmatic Encroachment in Epistemology* (pp.181-205). Routledge.

Beaver, D., & Stanley, J. (2023). *The Politics of Language*. Princeton University Press.

Bedke, M. (2017). Cognitivism and Non-Cognitivism. In T. McPherson & D. Plunkett (Eds.), *The Routledge Handbook of Metaethics* (pp.292-307). Routledge.

Bell, R. (2025). 'Just the Facts': Thick Concepts and Hermeneutical Misfit. *The Philosophical Quarterly*, 75(2), 373-395.

Belleri, D. (2021a). Downplaying the Change of Subject Objection to Conceptual Engineering. *Inquiry*, 1-24.

Belleri, D. (2021b). On Pluralism and Conceptual Engineering: Introduction and Overview. *Inquiry*, 1-19.

Belleri, D. (2025). 'You're Changing the Subject': An Unfair Objection to Conceptual Engineering? *The Philosophical Quarterly*, 75(3), 858-877.

Bergström, L. (2002). Putnam on the Fact-Value Dichotomy. *Croatian Journal of Philosophy*, 2(1), 1-13.

Berker, S. (2018). A Combinatorial Argument Against Practical Reasons for Belief. *Analytic Philosophy*, 59(4), 427-470.

Bertelsmann Stiftung. (2024). *Transformation Index BTI 2024*. Retrieved March 26, 2025, from https://www.bti-project.org/en/reports/global-findings

Bettcher, T. M. (2007). Evil Deceivers and Make-Believers: On Transphobic Violence and the Politics of Illusion. *Hypatia*, 22(3), 43-65.

Bigelow, J., & Pargetter, R. (1987). Functions. *The Journal of Philosophy*, 84(4), 181-196.

Blackburn, S. (1992). Through Thick and Thin. *Proceedings of the Aristotelian Society, Supplementary Volume*, 66, 284-299.

Blackburn, S. (1998). *Ruling Passions: A Theory of Practical Reasoning*. Oxford University Press.

Blackburn, S. (2013a). Disentangling Disentangling. In S. Kirchin (Ed.), *Thick Concepts* (pp.121-135). Oxford University Press.

Blackburn, S. (2013b). Pragmatism in Philosophy: The Hidden Alternative. *Philosophic Exchange*, 41(1), 2-13.

Blackburn, S. (2017). Pragmatism: All or Some or All and Some? In C. Misak & H. Price (Eds.), *The Practical Turn: Pragmatism in Britain in the Long Twentieth Century* (pp.61-74). Oxford University Press.

Blum, L. (2010). Racialized Groups: The Sociohistorical Consensus. *The Monist*, 93(2), 298-320.

Bolinger, R. J. (2018). The Rational Impermissibility of Accepting (Some) Racial Generalizations. *Synthese*, 1-17.

BonJour, L. (1980). Externalist Theories of Empirical Knowledge. *Midwest Studies in Philosophy*, 5(1), 53-74.

Booth, A. R. (2011). Epistemic Ought is a Commensurable Ought. *European Journal of Philosophy*, 22(4), 529-539.

Booth, A. R. (2012). All Things Considered Duties to Believe. *Synthese*, 187(2), 509-517.

Boyd, R. (1988). How to Be a Moral Realist. In G. Sayre-McCord (Ed.), *Essays on Moral Realism* (pp.181-228). Cornell University Press.

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

Brandom, R. (1994). *Making It Explicit*. Harvard University Press.

Brey, P. (2024, May 2). *A Theory of Moral Conceptual Change* [Lecture recording]. CEN Lecture Series. https://www.youtube.com/watch?v=-wDhOYuxhCI&t=435s

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

Brown, É. (2019). Fake News and Conceptual Ethics. *Journal of Ethics & Social Philosophy*, 16(2), 144-154.

Brun, G. (2016). Explication as a Method of Conceptual Re-Engineering. *Erkenntnis*, 81(6), 1211-1241.

Brun, G. (2022). Re-Engineering Contested Concepts. A Reflective-Equilibrium Approach.

Synthese, 200(2), 168.

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

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

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

Burgess, A., & Plunkett, D. (2020). On the Relation Between Conceptual Engineering and Conceptual Ethics. *Ratio*, 33(4), 281-294.

Burton, S. L. (1992). 'Thick' Concepts Revised. *Analysis*, 52(1), 28-32.

Cambridge University Press. (n.d.). Democracy. In *Cambridge English Dictionary*. Retrieved March 26, 2025, from https://dictionary.cambridge.org/dictionary/english/democracy

Camp, E. (2019). Perspectives and Frames in Pursuit of Ultimate Understanding. In S. R. Grimm (Ed.), *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology* (pp.17-46). Oxford University Press.

Cantalamessa, E. A. (2021). Disability Studies, Conceptual Engineering, and Conceptual Activism. *Inquiry*, 64(1-2), 46-75.

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

Cappelen, H. (2023). *The Concept of Democracy: An Essay on Conceptual Amelioration and Abandonment*. Oxford University Press.

Cappelen, H., & Dever, J. (2001). Believing in Words. *Synthese*, 127(3), 279-301.

Cappelen, H., & Plunkett, D. (2020). Introduction—A Guided Tour of Conceptual Engineering and Conceptual Ethics. In A. Burgess, H. Cappelen, & D. Plunkett (Eds.), *Conceptual Engineering and Conceptual Ethics* (pp.1-34). Oxford University Press.

Carlson, E. (2010). Parity Demystified. *Theoria*, 76(2), 119-128.

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

Carrara, M., & Vermaas, P. E. (2009). The Fine-Grained Metaphysics of Artifactual and Biological Functional Kinds. *Synthese*, 169(1), 125-143.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.

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

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

Chang, R. (1997). *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press.

Chang, R. (2001). Against Constitutive Incommensurability or Buying and Selling Friends. *Philosophical Issues*, 11, 33-60.

Chang, R. (2002). The Possibility of Parity. *Ethics*, 112(4), 659-688.

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

Chang, R. (2009). Voluntarist Reasons and the Sources of Normativity. In D. Sobel & S. Wall (Eds.), *Reasons for Action* (pp.243-271). Cambridge University Press.

Chang, R. (2013). Incommensurability (and Incomparability). In H. LaFollette (Ed.), *The International Encyclopedia of Ethics* (pp.2591-2604). Blackwell.

Chang, R. (2015). Value Pluralism. In J. D. Wright (Ed.), *International Encyclopedia of the Social and Behavioural Sciences* (pp.21-26). Elsevier.

Chappell, S. G. (2013). There Are No Thin Concepts. In S. Kirchin (Ed.), *Thick Concepts* (pp.182-196). Oxford University Press.

Chatelain, R. (2020, November 18). *Poll: 77% of Trump Backers Believe Biden's Win Due to Fraud*. NY1. Retrieved March 26, 2025, from https://www.cnn.com/2021/02/04/politics/2020-election-donald-trump-voter-fraud/index.html

Chomsky, N. (2002). *On Nature and Language*. Cambridge University Press.

Christiano, T. (2008). Democracy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2008 edition). Stanford University. Retrieved March 26, 2025, from https://plato.stanford.edu/archives/fall2008/entries/democracy/

Cillizza, C. (2021, February 4). *Three Quarters of Republicans Believe a Lie About the 2020 Election*. CNN Politics. Retrieved March 26, 2025, from https://www.cnn.com/2021/02/04/politics/2020-election-donald-trump-voter-fraud/index.html

Coady, D. (2024). Stop Talking About Echo Chambers and Filter Bubbles. *Educational Theory*, 74(1), 92-107.

Cohen, S. (1984). Justification and Truth. *Philosophical Studies*, 46(3), 279-295.

Copp, D. (2007). *Morality in a Natural World: Selected Essays in Metaethics*. Cambridge University Press.

Coppedge, M., Gerring, J., Altman, D., Bernhard, M., Fish, S., Hicken, A., Kroenig, M., Lindberg, S. I., McMann, K., Paxton, P., & Semetko, H. A. (2011). Conceptualizing and Measuring Democracy: A New Approach. *Perspectives on Politics*, 9(2), 247-267.

Cooper, R. (2004). Why Hacking is Wrong About Human Kinds. *The British Journal for the Philosophy of Science*, 55(1), 73-85

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

Crane, T. (2017). The Unity of Unconsciousness. *Proceedings of the Aristotelian Society*, 117(1), 1-21.

Crane, T., & Thompson, J. R. (2023). Implicit Cognition and Unconscious Mentality. In J. R. Thompson (Ed.), *The Routledge Handbook of Philosophy and Implicit Cognition* (pp.56-68). Routledge.

Craig, E. (1990). *Knowledge and the State of Nature*. Oxford University Press.

Crewe, B., Ievins, A., Larmour, S., Laursen, J., Mjåland, K., & Schliehe, A. (2023). Nordic Penal Exceptionalism: A Comparative, Empirical Analysis. *The British Journal of Criminology*, 63(2), 424-443.

Crisp, R. (2006). *Reasons and the Good*. Oxford University Press.

Crisp, R. (2025). Towards an Ethics of Conceptual Engineering. *Inquiry*, 68(2), pp.755-768.

Cull, M. J. (2019). Against Abolition. *Feminist Philosophy Quarterly*, 5(4), 1-16.

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

Cunningham, H. (2021). *Children and Childhood in Western Society Since 1500* (3rd ed.). Routledge.

Dahl, R. A. (1971). *Polyarchy: Participation and Opposition*. Yale University Press.

Dahl, R. A. (2006). *A Preface to Democratic Theory* (Expanded ed.). University of Chicago Press.

Dancy, J. (1995). In Defence of Thick Concepts. *Midwest Studies in Philosophy*, 20(1), 263-279.

Dancy, J. (2004). Enticing Reasons. In R. J. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (pp.91-118). Oxford University Press.

Dandelet, S. (2024). Epistemic Rationality and the Value of Truth. *The Philosophical Review*, 133(4), 329-365.

Danielsson, S., & Olson, J. (2007). Brentano and the Buck-Passers. *Mind*, 116(463), 511-522.

Dasgupta, S. (2014). The Possibility of Physicalism. *The Journal of Philosophy*, 111(9/10), 557-592.

Dembroff, R. (2020). Beyond Binary: Genderqueer as Critical Gender Kind. *Philosophers' Imprint*, 20(9), 1-31.

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

Deutsch, M. (2015). *The Myth of the Intuitive: Experimental Philosophy and Philosophical Method*. MIT Press.

Deutsch, M. (2020). Speaker's Reference, Stipulation, and a Dilemma for Conceptual Engineers. *Philosophical Studies*, 177(12), 3935-3957.

Devitt, M. (1981). *Designation*. Columbia University Press.

Devitt, M. (2011). Experimental Semantics. *Philosophy and Phenomenological Research*, 82(2), 418-435.

Dipert, R. R. (1993). *Artifacts, Art Works, and Agency*. Temple University Press.

Donnellan, K. S. (1970). Proper Names and Identifying Descriptions. *Synthese*, 21(3-4), 335-358.

Dorsey, D. (2016). *The Limits of Moral Authority*. Oxford University Press.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Dreier, J. (2004). Meta-Ethics and the Problem of Creeping Minimalism. *Philosophical Perspectives*, 18(1), 23-44.

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

Dunbar, R. I. M. (1998). *Grooming, Gossip, and the Evolution of Language*. Harvard University Press.

Dworkin, R. (2006). *Is Democracy Possible Here? Principles for a New Political Debate*. Princeton University Press.

Dworkin, R. (2011). *Justice for Hedgehogs*. Harvard University Press.

Economist Intelligence Unit. (2024). *Democracy Index 2024*. Retrieved March 26, 2025, from https://www.eiu.com/n/campaigns/democracy-index-2024/

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

Elstein, D. Y., & Hurka, T. (2009). From Thick to Thin: Two Moral Reduction Plans. *Canadian Journal of Philosophy*, 39(4), 515-536.

Eklund, M. (2017). *Choosing Normative Concepts*. Oxford University Press.

Eklund, M. (2021). Conceptual Engineering in Philosophy. In J. Khoo & R. Sterken (Eds.), *Routledge Handbook of Social and Political Philosophy of Language* (pp.15-30). Routledge.

Eklund, M. (2024). Conceptual Engineering and Conceptual Innovation. *Inquiry*, 1-24.

Epstein, B. (2015). *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. Oxford University Press.

Epstein, B. (2019). Anchoring Versus Grounding: Reply to Schaffer. *Philosophy and Phenomenological Research*, 99(3), 768-789.

Estlund, D. (2008). *Democratic Authority: A Philosophical Framework*. Princeton University Press.

Euronews. (2024, November 18). *Slovakians Rally Against Populism on Anniversary of Fall of Communist System*. Euronews. Retrieved March 26, 2025, from

https://www.euronews.com/my-europe/2024/11/18/slovakians-rally-against-populism-on-anniversary-of-fall-of-communist-system

Evans, G. (1973). The Causal Theory of Names. *Proceedings of the Aristotelian Society, Supplementary Volume*, 47(1), 187-225.

Evnine, S. J. (2016). *Making Objects and Events: A Hylomorphic Theory of Artifacts, Actions, and Organisms*. Oxford University Press.

Evnine, S. J. (2022). The Historicity of Artifacts: Use and Counter-Use. *Metaphysics*, 5(1), 1-13.

Fabre, C. (2022). III—Doxastic Wrongs, Non-Spurious Generalizations and Particularized Beliefs. *Proceedings of the Aristotelian Society*, 122(1), 47-69.

Feldman, R. (2000). The Ethics of Belief. *Philosophy and Phenomenological Research*, 60(3), 667-695.

Fine, A. (1975). How to Compare Theories: Reference and Change. *Noûs*, 9(1), 17-32.

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

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

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

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

Frege, G. (1892/1980). On Concept and Object. In M. Black & P. Geach (Eds. & Trans.), *Translations from the Philosophical Writings of Gottlob Frege* (pp.42-55). Basil Blackwell.

Frege, G. (1893/1980). On Sense and Reference. In M. Black & P. Geach (Eds. & Trans.), *Translations from the Philosophical Writings of Gottlob Frege* (pp.56-78). Basil Blackwell.

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

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

Gelfert, A. (2018). Fake News: A Definition. *Informal Logic*, 38(1), 84-117.

Gert, B., Culver, C. M., & Danner, C. K. (2006). *Death in Bioethics: A Systematic Approach* (2nd ed.). Oxford University Press.

Gheaus, A. (2023). Feminism Without "Gender Identity". *Politics, Philosophy & Economics*, 22(1), 31-54.

Gibbard, A. (1992). Thick Concepts and Warrant for Feelings. *Proceedings of the Aristotelian Society*, *Supplementary Volume*, 66, 267-283.

Gibbard, A. (2003). *Thinking How to Live*. Harvard University Press.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Psychology Press.

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

Godfrey-Smith, P. (1994). *A Modern History Theory of Functions*. Noûs, 28(3), 344-362.

Goetze, T. S. (2021). Conceptual Responsibility. *Inquiry*, 64(1-2), 20-45.

Goldman, A. I. (1976). Discrimination and Perceptual Knowledge. *The Journal of Philosophy*, 73(20), 771-791.

Golkar, S. (2024, February 28). *Iran's Democratic Façade: Elections and the Regime's Manipulation. Washington Institute for Near East Policy*. Retrieved March 26, 2025, from https://www.washingtoninstitute.org/policy-analysis/elections-expose-irans-fading-democratic-pretensions

Graeber, D. (2011). *Debt: The First Five Thousand Years*. Melville House.

Grimm, S. R. (2010). *The Goal of Explanation. Studies in History and Philosophy of Science Part A*, 41(4), 337-344.

Griffiths, P. E. (1993). Functional Analysis and Proper Functions. *The British Journal for the Philosophy of Science*, 44(3), 409-422.

Griffin, J. (1986). *Well-Being: Its Meaning, Measurement and Moral Importance*. Clarendon Press.

Habermas, J. (1998). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (W. Rehg, Trans.). MIT Press.

Habgood-Coote, J. (2019a). Stop Talking About Fake News! *Inquiry*, 62(9-10), 1033-1065.

Habgood-Coote, J. (2019b). What's the Point of Knowing How? *European Journal of Philosophy*, 27(3), 693-708.

Habgood-Coote, J. (2022). Fake News, Conceptual Engineering, and Linguistic Resistance: Reply to Pepp, Michaelson and Sterken, and Brown. *Inquiry*, 65(4), 488-516

Hamilton, A., Madison, J., & Jay, J. (1787-1788/1961). *The Federalist* (J. E. Cooke, Ed.). Wesleyan University Press.

Hacking, I. (1999). *The Social Construction of What?* Harvard University Press.

Harcourt, E., & Thomas, A. (2013). Thick Concepts, Analysis, and Reductionism. In S. Kirchin (Ed.), *Thick Concepts* (pp.20-43). Oxford University Press.

Hardimon, M. O. (2017). *Rethinking Race: The Case for Deflationary Realism*. Harvard University Press.

Harding, S. R., Flannelly, K. J., Weaver, A. J., & Costa, K. G. (2005). The Influence of Religion on Death Anxiety and Death Acceptance. *Mental Health, Religion and Culture*, 8(3), 253-261.

Hare, R. M. (1952). *The Language of Morals*. Clarendon Press.

Hare, R. M. (1963). *Freedom and Reason*. Oxford University Press.

HarperCollins Publishers. (n.d.). Democracy. In *Collins English Dictionary*. Retrieved March 26, 2025, from https://www.collinsdictionary.com/dictionary/english/democracy

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

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

Haslanger, S. (2005). What Are We Talking About? The Semantics and Politics of Social Kinds. *Hypatia*, 20(4), 10-26.

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

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique*. Oxford University Press.

Haslanger, S. (2015). Social Structure, Narrative and Explanation. *Canadian Journal of Philosophy*, 45(1), 1-15.

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

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

Hawthorne, J., & Lepore, E. (2011). On Words. *The Journal of Philosophy*, 108(9), 447-485.

Heil, J. (1992). Believing Reasonably. *Noûs*, 26(1), 47-61.

Held, D. (2006). *Models of Democracy* (3rd ed.). Polity Press.

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

Hilpinen, R. (1993). Authors and Artifacts. *Proceedings of the Aristotelian Society*, 93(1), 155-178.

Hills, A. (2016). Understanding Why. *Noûs*, 50(4), 661-688.

Hirschman, D., & Manser, A. R. (1973). Function and Explanation. *Proceedings of the Aristotelian Society*, *Supplementary Volume*, 47(1), 19-38.

Hardimon, M. O. (2017). *Rethinking Race: The Case for Deflationary Realism*. Harvard University Press.

Hochman, A. (2017). Replacing Race: Interactive Constructionism About Racialized Groups. *Ergo*, 4(22), 61-92.

Holm, S. (2017). The Problem of Phantom Functions. *Erkenntnis*, 82(2), 233-241.

Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford University Press.

Horgan, T., & Timmons, M. (1991). New Wave Moral Realism Meets Moral Twin Earth. *Journal of Philosophical Research,* 16, 447-465.

Horgan, T., & Timmons, M. (1992). Troubles for New Wave Moral Semantics: The "Open Question Argument" Revived. *Philosophical Papers*, 21(3), 153-175.

Houkes, W., & Vermaas, P. (2004). Actions Versus Functions: A Plea for an Alternative Metaphysics of Artifacts. *The Monist*, 87(1), 52-71.

Houkes, W., & Vermaas, P.E. (2010). *Technical Functions: On the Use and Design of Artefacts* (Vol. 1). Springer.

Howard, C. (2020). Weighing Epistemic and Practical Reasons for Belief. *Philosophical Studies*, 177(8), 2227-2243.

Irmak, N. (2019). An Ontology of Words. *Erkenntnis*, 84(5), 1139-1158.

Isaac, M. G. (2020). How to Conceptually Engineer Conceptual Engineering? *Inquiry*, 63(9-10), 1033-1056.

Isaac, M. G. (2021). Which Concept of Concept for Conceptual Engineering? Erkenntnis, 1-25.

Isaac, M. G., Koch, S., & Nefdt, R. (2022). Conceptual Engineering: A Road Map to Practice. *Philosophy Compass*, 17(10), e12879.

Isaac, M. G. (2024). Post-Truth Conceptual Engineering. *Inquiry*, 67(1), 199-214.

Isaac, M. G. (2025). The Hallmark Problem for Conceptual Engineering. *Metaphilosophy*, 1-18.

Izumi, Y., Kasaki, M., Zhou, Y., & Oda, S. (2018). Definite Descriptions and the Alleged East–West Variation in Judgments About Reference. *Philosophical Studies*, 175(5), 1183-1205.

Jackson, F. (1998). Reference and Description Revisited. *Philosophical Perspectives*, 12, 201-218.

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

Jenkins, K. (2018). Toward an Account of Gender Identity. *Ergo*, 5(27), 713-744.

Jeffers, C. (2019). Cultural Constructionism. In J. Glasgow, S. Haslanger, C. Jeffers, & Q. Spencer (Eds.), *What Is Race?: Four Philosophical Views* (pp.38-72). Oxford University Press.

Johnson, D. (2015). *God is Watching You: How the Fear of God Makes Us Human*. Oxford University Press.

Jonas, E., & Fischer, P. (2006). Terror Management and Religion: Evidence That Intrinsic Religiousness Mitigates Worldview Defense Following Mortality Salience. *Journal of Personality and Social Psychology*, 91(3), 553-567.

Jonas, H. (1969). Philosophical Reflections on Experimenting with Human Subjects. *Daedalus*, 98(2), 219-247.

Jones, K. (2014). Intersectionality and Ameliorative Analyses of Race and Gender. *Philosophical Studies*, 171(1), 99-107.

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

Jorem, S. (2022). The Good, the Bad and the Insignificant—Assessing Concept Functions for Conceptual Engineering. *Synthese*, 200(2), 106.

Jorem, S., & Löhr, G. (2024). Inferentialist Conceptual Engineering. *Inquiry*, 67(3), 932-953.

Juvshik, T. (2023). On the Social Nature of Artifacts. *Theoria*, 89(6), 910-932.

Kaplan, D. (1990). Words. *Proceedings of the Aristotelian Society*, *Supplementary Volume*, 64(1), 93-119.

Kaplan, D. (2011). Words on Words. *The Journal of Philosophy*, 108(9), 504-529.

Kapusta, S. J. (2016). Misgendering and Its Moral Contestability. *Hypatia*, 31(3), 502-519.

Karmanau, Y. (2025, March 25). *Lukashenko Claims Belarus is More Democratic Than Critics Suggest*. AP News. Retrieved March 26, 2025, from https://apnews.com/article/7b5d85b8400d678a19608f3054e63350

Kazankov, D., & Yi, E. (2024). Linguistic Imposters. *The Philosophical Quarterly*, 74(4), 1182-1206.

Kelp, C. (2014). Two for the Knowledge Goal of Inquiry. *American Philosophical Quarterly*, 51(3), 227-232.

Khalidi, M. A. (2015). Three Kinds of Social Kinds. *Philosophy and Phenomenological Research*, 90(1), 96-112.

Khalifa, K. (2013). The Role of Explanation in Understanding. *The British Journal for the Philosophy of Science*, 64(1), 161-187.

Kirchin, S. (2010). The Shapelessness Hypothesis. *Philosophers' Imprint*, 10(4), 1-28.

Kirchin, S. (2017). *Thick Evaluation*. Oxford University Press.

Kitcher, P. (1993). Function and Design. *Midwest Studies in Philosophy*, 18(1), 379-397.

Kitcher, P. (2007). Does 'Race' Have a Future? *Philosophy and Public Affairs*, 35(4), 293-317.

Kitsik, E. (2023a). Epistemic Environmentalism and Autonomy: The Case of Conceptual Engineering. *Canadian Journal of Philosophy*, 53(6), 487-501.

Kitsik, E. (2023b). Epistemic Paternalism via Conceptual Engineering. *Journal of the American Philosophical Association*, 9(4), 616-635.

Knight, J., & Johnson, J. (1997). What Sort of Equality Does Deliberative Democracy Require? In J. Bohman & W. Rehg (Eds.), *Deliberative Democracy: Essays on Reason and Politics* (pp.279-320). MIT Press.

Knoll, V. (2020). Verbal Disputes and Topic Continuity. *Inquiry*, 1-22.

Koch, S. (2021a). Engineering What? On Concepts in Conceptual Engineering. *Synthese*, 199(1), 1955-1975.

Koch, S. (2021b). The Externalist Challenge to Conceptual Engineering. *Synthese*, 198(1), 327-348.

Koch, S. (2023). Why Conceptual Engineers Should Not Worry About Topics. *Erkenntnis*, 88(5), 2123-2143.

Koch, S., & Lupyan, G. (2024). What Is Conceptual Engineering Good For? The Argument From Nameability. *Philosophy and Phenomenological Research*, 1-18.

Koch, S. (2025). Should We Stop Talking About "Democracy"? Conceptual Abandonment and the Perils of Political Discourse. *Asian Journal of Philosophy*, 4(1), 28.

Koch, S., Löhr, G., & Pinder, M. (2023). Recent Work in the Theory of Conceptual Engineering. *Analysis*, 83(3), 589-603.

Koch, S., & Ohlhorst, J. (2024). Heavy-Duty Conceptual Engineering. *Noûs*, pp.1-19.

Kocurek, A. W. (2022). What Topic Continuity Problem? *Inquiry*, 1-21.

Köhler, S., & Veluwenkamp, H. (2024). Conceptual Engineering: For What Matters. *Mind*, 133(530), 400-427.

Köhler, S., & Veluwenkamp, H. (2025). Proper Foundations for Conceptual Ethics. *Synthese*, 206(2), 1-20.

Koslicki, K. (2012). Essence, Necessity, and Explanation. In T. E. Tahko (Ed.), *Contemporary Aristotelian Metaphysics* (pp.187-206). Cambridge University Press.

Koslicki, K. (2018). *Form, Matter, Substance*. Oxford University Press.

Koslicki, K. (2023). Artifacts and the Limits of Agentive Authority. In M. Garcia-Godinez (Ed.), *Thomasson on Ontology* (pp.209-241). Palgrave Macmillan.

Koslow, A. (2022). Meaning Change and Changing Meaning. *Synthese*, 200(2), 94.

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

Kroes, P. (2012). *Technical Artefacts: Creations of Mind and Matter: A Philosophy of Engineering Design* (Philosophy of Engineering and Technology, Vol. 6). Springer.

Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press.

Kyle, B. G. (2020). The Expansion View of Thick Concepts. *Noûs*, 54(4), 914-944.

Lam, B. (2010). Are Cantonese Speakers Really Descriptivists? Revisiting Cross-Cultural Semantics. *Cognition*, 115(2), 320-332.

Lanius, D. (2021). What Is the Value of Vagueness? *Theoria*, 87(3), 752-780.

Landes, E. (forthcoming). How Language Teaches and Misleads: "Coronavirus" and "Social Distancing as Case Studies. In: M.G. Isaac, K. Scharp. and S. Koch (eds.), *New Perspectives on Conceptual Engineering*. Synthese Library.

Landes, E. (2025). Conceptual Engineering Should Be Empirical. *Erkenntnis*, pp.1-21.

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

Laurence, S., & Margolis, E. (1999). *Concepts and Cognitive Science*. In E. Margolis & S.

Laurence (Eds.), Concepts: Core Readings (pp.3-81). MIT Press.

Lepoutre, M. (2024). Mobilizing Falsehoods. *Philosophy & Public Affairs*, 52(2), 106-146..

Levine, R. J. (1988). *Ethics and Regulation of Clinical Research* (2nd ed.). Yale University Press.

Levy, N. (2017). The Bad News About Fake News. *Social Epistemology Review and Reply Collective*, 6(3), 20-36.

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

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

Lijphart, A. (1999). *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*. Yale University Press.

List, C. (2014). Three Kinds of Collective Attitudes. *Erkenntnis*, 79(Suppl 9), 1601-1622.

List, C., & Pettit, P. (2006). Group Agency and Supervenience. *Southern Journal of Philosophy*, 44(S1), 85-105.

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford University Press.

Locke, J. (1689/1975). *An Essay Concerning Human Understanding* (P. H. Nidditch, Ed.). Oxford University Press.

Löhr, G. (2020). Concepts and Categorization: Do Philosophers and Psychologists Theorize About Different Things? *Synthese*, 197(5), 2171-2191.

Löhr, G. (2021). Commitment Engineering: Conceptual Engineering Without Representations. *Synthese*, 199(5), 13035-13052.

Löhr, G. (2025). What Is Pragmatist Conceptual Engineering? *Inquiry*, 1-31.

Löhr, G., & Veluwenkamp, H. (2025). Rethinking Philosophical Methodology: Conceptual Engineering Meets Value Sensitive Design. *Metaphilosophy*, 1-16.

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

López-Bultó, O., Piqué, R., Antolín, F., Barceló, J. A., Palomo, A., & Clemente, I. (2020). Digging Sticks and Agriculture Development at the Ancient Neolithic Site of la Draga (Banyoles, Spain). *Journal of Archaeological Science: Reports*, 30, 102-193.

Ludwig, K. (2007). The Epistemology of Thought Experiments: First-Person Versus Third-Person Approaches. *Midwest Studies in Philosophy*, 31(1), 128-159.

Lupyan, G., & Zettersten, M. (2021). Does Vocabulary Help Structure the Mind? In M. D. Sera & M. Koenig (Eds.), *Minnesota Symposia on Child Psychology* (Vol.40, pp.160-199). Wiley.

Machery, E. (2009). *Doing Without Concepts*. Oxford University Press.

Machery, E., Mallon, R., Nichols, S., & Stich, S. (2004). Semantics, Cross-Cultural Style. *Cognition*, 92(1), 1-12.

Machery, E. (2017). *Philosophy Within Its Proper Bounds*. Oxford University Press.

Machery, E. (2021). A New Challenge to Conceptual Engineering. *Inquiry*, 1-24.

Macarthur, D., & Price, H. (2007). Pragmatism, Quasi-Realism, and the Global Challenge. In C. Misak (Ed.), *New Pragmatists* (pp.91-121). Oxford University Press.

Maguire, B., & Woods, J. (2020). The Game of Belief. *The Philosophical Review*, 129(2), 211-249

Mallon, R. (2006). 'Race': Normative, Not Metaphysical or Semantic. *Ethics*, 116(3), 525-551.

Mallon, R. (2007). Human Categories Beyond Non-Essentialism. *Journal of Political Philosophy*, 15(2), 146-168.

Mallon, R., Machery, E., Nichols, S., & Stich, S. (2009). Against Arguments From Reference. *Philosophy and Phenomenological Research*, 79(2), 332-356.

Marchiori, S. (2025). Conceptual Affordances:(How) Should They Inform Conceptual Engineering?. *Synthese*, 206(2), 1-23.

Marques, T. (Forthcoming). Representing or Shaping Reality? What 'Class' Can Teach About 'Woman'. In M. G. Isaac, S. Koch, & K. Scharp (Eds.), *New Perspectives on Conceptual Engineering*. Synthese Library.

Marques, T. (2020). Amelioration vs Perversion. In T. Marques & Å. Wikforss (Eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variation* (pp.260-284). Oxford University Press.

Mason, C. (2023). Reconceiving Murdochian Realism. *Ergo*, 10, 649-672.

Mathiesen, T. (2012). Scandinavian Exceptionalism in Penal Matters: Reality or Wishful Thinking? In T. Ugelvik & J. Dullum (Eds.), Penal Exceptionalism? Nordic Prison Policy and Practice (pp.13-38). Routledge.

Matsui, T. (2024). Local Conceptual Engineering in a Linguistic Subgroup and the Implementation Problem. In P. Stalmaszczyk (Ed.), *Conceptual Engineering: Methodological and Metaphilosophical Issues* (pp.117-133). Brill.

McDowell, J. (1981). Non-Cognitivism and Rule-Following. In *Mind, Value, and Reality* (pp.198-219). Harvard University Press.

McHugh, C. (2013). Normativism and Doxastic Deliberation. *Analytic Philosophy*, 54(4), 447-465.

McKenna, R. (2018). No Epistemic Trouble for Engineering 'Woman': Response to Simion. *Logos & Episteme*, 9(3), 335-342.

McLeod, O. (2001). Just Plain Ought. *Journal of Ethics*, 5(4), 269-291.

McPherson, T. (2018). Authoritatively Normative Concepts. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol.13, pp.253-275). Oxford University Press.

McPherson, T. (2025). Meaninglessness and the Ethics of Lexical Abandonment. *Asian Journal of Philosophy*, 4(1), 19.

McPherson, T., & Plunkett, D. (2021). The Vindicatory Circularity Challenge to the Conceptual Ethics of Normativity. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol.16, pp.207-232). Oxford University Press.

McPherson, T., & Plunkett, D. (2024). Topic Continuity in Conceptual Engineering and Beyond. *Inquiry*, 67(9), 2847-2873.

Medina, J. (2013). *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. Oxford University Press.

Meylan, A. (2021). Doxastic Divergence and the Problem of Comparability. P*hilosophy and Phenomenological Research*, 103(1), 199-216.

Michaelson, E. (2022). The Vagaries of Reference. *Ergo*, 9(52).

Mill, J. S. (1865/1958). *Considerations on Representative Government* (3rd ed., C. V. Shields, Ed.). Bobbs-Merrill.

Miller, J. T. (2021). A Bundle Theory of Words. *Synthese*, 198(6), 5731-5748.

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

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Millikan, R. G. (1989). In Defense of Proper Functions. *Philosophy of Science*, 56(2), 288-302.

Millikan, R. G. (1999). Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function. *The Journal of Philosophy*, 96(4), 191-206.

Mills, C. W. (1998). *The Racial Contract*. Cornell University Press.

Mills, C. W. (2007). White Ignorance. In S. Sullivan & N. Tuana (Eds.), *Race and Epistemologies of Ignorance* (pp.26-31). State University of New York Press.

Mitchell, S. D. (1993). Dispositions or Etiologies? A Comment on Bigelow and Pargetter. *The Journal of Philosophy*, 90(5), 249-259.

Moore, G. E. (1942). A Reply to My Critics. In P. A. Schilpp (Ed.), *The Philosophy of G.E. Moore* (pp.535-677). Open Court.

Murdoch, I. (1970). *The Sovereignty of Good*. Routledge.

Murdoch, I. (1992). *Metaphysics as a Guide to Morals*. Chatto & Windus.

Nagel, E. (1977). Teleology Revisited. *The Journal of Philosophy*, 74(5), 261-301.

Nagel, T. (1986). *The View From Nowhere*. Oxford University Press.

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

Nado, J. (2021b). Classification Procedures as the Targets of Conceptual Engineering. *Philosophy and Phenomenological Research*, 106(1), 136-156.

Nado, J. (2021c). Conceptual Engineering via Experimental Philosophy. *Inquiry*, 64(1-2), 76-96.

Nado, J. (2023). Taking Control: Conceptual Engineering Without (Much) Metasemantics. *Inquiry*, 66(10), 1974-2000.

Neander, K. (1991). The Teleological Notion of 'Function'. *Australasian Journal of Philosophy*, 69(4), 454-468.

Nefdt, R. M. (2024). Concepts and Conceptual Engineering: Answering Cappelen's Challenge. *Inquiry*, 67(1), 400-428.

Nelson, J. A. (1982). Schwartz on Reference. *The Southern Journal of Philosophy*, 20(3), 359-365.

Nimtz, C. (2024a). Engineering Concepts by Engineering Social Norms: Solving the Implementation Challenge. *Inquiry*, 67(6), 1716-1743.

Nimtz, C. (2024b). The Power of Social Norms: Why Conceptual Engineers Should Care About Implementation. *Synthese*, 203(6), 215.

Oderberg, D. S. (2011). Essence and Properties. *Erkenntnis*, 75(1), 85-111.

Ohlhorst, J. (2023). Engineering Virtue: Constructionist Virtue Ethics. *Inquiry*, 1-20.

Orsi, F. (2015). *Value Theory*. Bloomsbury.

Oxford University Press. (n.d.). Democracy. In *Oxford English Dictionary*. Retrieved March 26, 2025, from https://www.oed.com/dictionary/democracy_n

Pagano, E. (2024). Social Construction, Social Kinds and Exportation. *Analysis*, 84(1), 83-93.

Papineau, D. (2013). There Are No Norms of Belief. In T. Chan (Ed.), *The Aim of Belief* (pp.64-79). Oxford University Press.

Parfit, D. (1986). *Reasons and Persons*. Clarendon Press.

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

Parsons, G. (2019). Phantom Functions and the Evolutionary Theory of Artefact Proper Function. *Grazer Philosophische Studien*, 96(1), 154-170.

Passinsky, A. (2021). Should Bitcoin Be Classified as Money? *Journal of Social Ontology*, 6(2), 281-292.

Passinsky, A. (2025). Social Kind Essentialism. *Philosophical Studies*, 182(3), 1023-1046.

Pateman, C. (1970). *Participation and Democratic Theory*. Cambridge University Press.

Peacocke C. (1992). *A Study of Concepts*. MIT Press.

Pepp, J., Michaelson, E., & Sterken, R. (2022). Why We Should Keep Talking About Fake News. *Inquiry*, 65(4), 471-487.

Pettit, P. (2018). *The Birth of Ethics: Reconstructing the Role and Nature of Morality*. Oxford University Press.

Pettit, P., & Schweikard, D. (2006). Joint Actions and Group Agents. *Philosophy of the Social Sciences*, 36(1), 18-39.

Pinker, S., & Bloom, P. (1990). Natural Language and Natural Selection. *Behavioral and Brain Sciences*, 13(4), 707-784.

Pinder, M. (2017). Does Experimental Philosophy Have a Role to Play in Carnapian Explication? *Ratio*, 30(4), 443-461.

Pinder, M. (2020). Conceptual Engineering, Speaker-Meaning and Philosophy. *Inquiry*, 63(9-10), 925-939.

Pinder, M. (2021). Conceptual Engineering, Metasemantic Externalism and Speaker-Meaning. *Mind*, 130(517), 141-163.

Plantinga, A. (1993). *Warrant: The Current Debate*. Oxford University Press.

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

Plunkett, D. (2016). Negotiating the Meaning of "Law": The Metalinguistic Dimension of the Dispute Over Legal Positivism. *Legal Theory*, 22(3-4), 205-275.

Plunkett, D., & Sundell, T. (2013). Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint*, 13(23), 1-37.

Plunkett, D., & Sundell, T. (2021a). Metalinguistic Negotiation and Matters of Language: A Response to Cappelen. *Inquiry*, 1-25.

Plunkett, D., & Sundell, T. (2021b). Metalinguistic Negotiation and Speaker Error. *Inquiry*, 64(1-2), 142-167.

Pohlhaus, G. (2012). Relational Knowing and Epistemic Injustice: Toward a Theory of Willful Hermeneutical Ignorance. *Hypatia*, 27(4), 715-735.

Pollock, J. L. (1984). Reliability and Justified Belief. *Canadian Journal of Philosophy*, 14(1), 103-114.

Pollock, J. (2021). Content Internalism and Conceptual Engineering. *Synthese*, 198(12), 11587-11605.

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

Podosky, P.M.C. (2022). Don't Count Truth Out Just Yet: A Response to Isaac. *Inquiry*, pp.1-11.

Pratt, J. (2008a). Scandinavian Exceptionalism in an Era of Penal Excess: Part I: The Nature and Roots of Scandinavian Exceptionalism. *The British Journal of Criminology*, 48(2), 119-137.

Pratt, J. (2008b). Scandinavian Exceptionalism in an Era of Penal Excess: Part II: Does Scandinavian Exceptionalism Have a Future? *The British Journal of Criminology*, 48(3), 275-292.

Pratt, J., & Eriksson, A. (2014). *Contrasts in Punishment: An Explanation of Anglophone Excess and Nordic Exceptionalism*. Routledge.

Preston, B. (1998). Why Is a Wing Like a Spoon? *The Journal of Philosophy*, 95(5), 215-254.

Preston, B. (2013). *A Philosophy of Material Culture*. Routledge.

Price, H. (2011). *Naturalism Without Mirrors*. Oxford University Press.

Price, H., Blackburn, S., Brandom, R., Horwich, P., & Williams, M. (Eds.). (2013). *Expressivism, Pragmatism and Representationalism*. Cambridge University Press.

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

Pritchard, D. (2021). Intellectual Virtues and the Epistemic Value of Truth. *Synthese*, 198, 5515-5528.

Putnam, H. (1975). The Meaning of 'Meaning'. In K. Gunderson (Ed.), *Language, Mind, and Knowledge* (pp.131-193). University of Minnesota Press.

Putnam, H. (2002). *The Collapse of the Fact/Value Dichotomy and Other Essays.* Harvard University Press.

Queloz, M. (2019). The Points of Concepts: Their Types, Tensions, and Connections. *Canadian Journal of Philosophy*, 49(8), 1122-1145.

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

Queloz, M. (2022). Function-Based Conceptual Engineering and the Authority Problem. *Mind*, 131(524), 1247-1278.

Queloz, M. (2024a). The Dworkin-Williams Debate: Liberty, Conceptual Integrity, and Tragic Conflict in Politics. *Philosophy and Phenomenological Research*, 109(1), 3-29.

Queloz, M. (2024b). Virtues, Rights, or Consequences? Mapping the Way for Conceptual Ethics. *Studia Philosophica: The Swiss Journal of Philosophy*, 83(1), 9-22.

Queloz, M. (2025). *The Ethics of Conceptualization: Tailoring Thought and Language to Need*. Oxford University Press.

Queloz, M., & Bieber, F. (2022). Conceptual Engineering and the Politics of Implementation. *Pacific Philosophical Quarterly*, 103(3), 670-691.

Railton, P. (1986). Moral Realism. *The Philosophical Review*, 95(2), 163-207.

Rabinowicz, W. (2008). Value Relations. *Theoria*, 74(1), 18-49.

Rabinowitz, W. (2021). Incommensurability Meets Risk. In H. Andersson & A. Herlitz (Eds.), *Value Incommensurability: Ethics, Risk, and Decision-Making*. Routledge.

Rawls, J. (1999). *A Theory of Justice* (Rev. ed.). Harvard University Press.

Raz, J. (1986). *The Morality of Freedom*. Clarendon Press.

Raz, J. (1999). The Central Conflict: Morality and Self-Interest. In *Engaging Reason* (pp.303-332). Oxford University Press.

Reisner, A. (2008). Weighing Pragmatic and Evidential Reasons for Belief. *Philosophical Studies*, 138, 17-27.

Reisner, A. (2018). Two Theses About the Distinctness of Practical and Theoretical Normativity. In C. McHugh, J. Way, & D. Whiting (Eds.), *Normativity: Epistemic and Practical*. Oxford Academic.

Rey, G. (1983). Concepts and Stereotypes. *Cognition*, 15(3), 237-262.

Richardson, K. (2024). Value Magnetism: Why Conceptual Engineering Requires Objective Values. *Global Philosophy*, 34(1), 21.

Riggs, J. (2019). Conceptual Engineers Shouldn't Worry About Semantic Externalism. *Inquiry*, 62(9-10), 1-22.

Riggs, J. (2021). Deflating the Functional Turn in Conceptual Engineering. *Synthese*, 199(3), 11555-11586.

Rini, R. (2017). Fake News and Partisan Epistemology. *Kennedy Institute of Ethics Journal*, 27(2), 43-64.

Roberts, D. (2011). Shapelessness and the Thick. *Ethics*, 121(3), 489-520.

Roberts, D. (2013a). It's Evaluation, Only Thicker. In S. Kirchin (Ed.), *Thick Concepts* (pp.78-98). Oxford University Press.

Roberts, D. (2013b). Thick Concepts. *Philosophy Compass*, 8(8), 677-688.

Roberts, D. (2017). Depending on the Thick. *Proceedings of the Aristotelian Society*, *Supplementary Volume*, 91(1), 197-220.

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

Russell, B. (1903). *Principles of Mathematics*. Allen and Unwin.

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

Rushing, S. (2013). What Is Confucian Humility? In S. Angle & M. Slote (Eds.), *Virtue Ethics and Confucianism* (pp.173-181). Routledge.

Sagdahl, M. S. (2022). *Normative Pluralism: Resolving Conflicts Between Moral and Prudential Reasons*. Oxford University Press.

Sainsbury, R. M., & Tye, M. (2012). *Seven Puzzles of Thought and How to Solve Them: An Originalist Theory of Concepts*. Oxford University Press.

Saito, K. (2024). *Slow Down: The Degrowth Manifesto*. Astra Publishing House.

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

Salmon, W. C. (1997). *Causality and Explanation*. Oxford University Press.

Sauer, P. (2024, March 18). Putin Election Result Hailed by Kremlin Despite Condemnation of Lack of Democracy. *The Guardian*. Retrieved March 26, 2025, from https://www.theguardian.com/world/2024/mar/18/putin-election-result-russia-ukraine-war

Saul, J. (2006). Philosophical Analysis and Social Kinds: Gender and Race. *Proceedings of the Aristotelian Society*, 106(1), 119-143.

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

Schaffer, J. (2019). Anchoring as Grounding: On Epstein's the Ant Trap. *Philosophy and Phenomenological Research*, 99(3), 749-767.

Scharp, K. (2020). Philosophy as the Study of Defective Concepts. In A. Burgess, H. Cappelen, & D. Plunkett (Eds.), *Conceptual Engineering and Conceptual Ethics* (pp.396-416). Oxford University Press.

Scheffler, S. (1987). Morality Through Thick and Thin: A Critical Notice of Ethics and the Limits of Philosophy. *The Philosophical Review*, 96(3), 411-434.

Scheffler, S. (1992). *Human Morality*. Oxford University Press.

Scheffler, S. (2008). Potential Congruence. In P. Bloomfield (Ed.), *Morality and Self-Interest* (pp.303-332). Oxford University Press.

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

Schroeder, M. (2018). When Beliefs Wrong. *Philosophical Topics*, 46(1), 115-128. Schroeter, L., &

Schroeter, F. (2015). Rationalizing Self-Interpretation. In C. Daly (Ed.), *The Palgrave Handbook of Philosophical Methods* (pp.419-447). Palgrave Macmillan.

Schumpeter, J. A. (1943/2003). *Capitalism, Socialism and Democracy*. Routledge.

Searle, J. R. (1958). Proper Names. *Mind*, 67(266), 166-173.

Searle, J. R. (1995). *The Construction of Social Reality*. Simon and Schuster.

Sękowski, K., & Landes, E. (2024). Conceptual Engineering Is Old News. *The Philosophical Quarterly*, pqae087.

Sellars, W. (1963/2007). Philosophy and the Scientific Image of Man. In K. Scharp & R. B. Brandom (Eds.), *In the Space of Reasons* (pp.369-408). Harvard University Press.

Shah, N. (2003). How Truth Governs Belief. *The Philosophical Review*, 112(4), 447-482.

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

Sherratt, A. G. (1981). Plough and Pastoralism: Aspects of the Secondary Products Revolution. In I. Hodder, G. Isaac, & N. Hammond (Eds.), *Pattern of the Past: Studies in Honour of David Clarke* (pp.261-305). Cambridge University Press.

Shields, M. (2021). Conceptual Domination. *Synthese*, 199(5), 15043-15067.

Shields, M. (2023). Conceptual Engineering, Conceptual Domination, and the Case of Conspiracy Theories. *Social Epistemology*, 37(4), 464-480.

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

Siderits, M. (2020). Against Reducing Arthāpatti. In M. Keating (Ed.), *Controversial Reasoning in Indian Philosophy: Major Texts and Arguments on Arthâpatti* (pp.289-310). Bloomsbury Academic.

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

Simion, M. (2018b). Epistemic Trouble for Engineering 'Woman'. *Logos & Episteme*, 9(1), 91-98.

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

Sliwa, P. (2024). Making Sense of Things: Moral Inquiry as Hermeneutical Inquiry. *Philosophy and Phenomenological Research*, 109(1), 117-137.

Smith, P. S. (2012). A Critical Look at Scandinavian Exceptionalism. In T. Ugelvik & J. Dullum (Eds.), *Penal Exceptionalism? Nordic Prison Practice and Policy* (pp.38-57). Routledge.

Smyth, N. (n.d.). *The New Philosopher-Kings: Conceptual Engineering and Social Authority* [Unpublished manuscript]. PhilPapers. Retrieved May 26, 2025, from https://philpapers.org/rec/SMYTNP

Smyth, N. (2023). Purity and Practical Reason: On Pragmatic Genealogy. *Ergo*, 10(37), 1057-1081.

Sojka, R. E., Bjorneberg, D. L., & Entry, J. A. (2005). Irrigation: Historical Perspective. In R. Lal (Ed.), *Encyclopedia of Soil Science* (pp.745-749). CRC Press.

Sosis, R., & Bressler, E. R. (2003). Cooperation and Commune Longevity: A Test of the Costly Signaling Theory of Religion. *Cross-Cultural Research*, 37(4), 211-239.

Spike, J. (2025, March 25). *Protesters Block Traffic in Hungary Opposing Ban on LGBTQ Pride Events*. AP News. Retrieved March 26, 2025, from https://apnews.com/article/protesters-block-traffic-hungary-opposing-ban-lgbtq-pride-events-b461894acf829683fe1603e55588e98c

Spinoza, B. (1677/1997). *On the Improvement of the Understanding*. Dover Publications.

Stearns, P. N. (2017). *Childhood in World History* (3rd ed.). Routledge.

Steglich-Petersen, A., & Skipper, M. (2020). An Instrumentalist Account of How to Weigh Epistemic and Practical Reasons for Belief. *Mind*, 129(516), 1071-1094.

Sterken, R. (2020). Linguistic Interventions and Transformative Communicative Disruption. In H. Cappelen, D. Plunkett, & A. Burgess (Eds.), *Conceptual Engineering and Conceptual Ethics* (pp. 417–434). Oxford University Press.

Stevenson, C. L. (1944). *Ethics and Language*. Yale University Press.

Stocker, M. (1997). Abstract and Concrete Value: Plurality, Conflict and Maximization. In R. Chang (Ed.), *Incommensurability, Incomparability, and Practical Reason* (pp.196-214). Harvard University Press.

Stojnić, U. (2022). Just Words: Intentions, Tolerance and Lexical Selection. *Philosophy and Phenomenological Research*, 105(1), 3-17.

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

Strevens, M. (2019). *Thinking off Your Feet: How Empirical Psychology Vindicates Armchair Philosophy*. Harvard University Press.

Sundell, T. (2020). Changing the Subject. *Canadian Journal of Philosophy*, 50(5), 580-593.

Sylvan, K. (2020). An Epistemic Non-Consequentialism. *The Philosophical Review*, 129(1), 1-51.

Sytsma, J., Livengood, J., Sato, R., & Oguchi, M. (2015). Reference in the Land of the Rising Sun: A Cross-Cultural Study on the Reference of Proper Names. *Review of Philosophy and Psychology,* 6(2), 212-230.

Tappolet, C. (2004). Through Thick and Thin: Good and Its Determinables. *Dialectica*, 58(2), 207-221.

Taylor, P. C. (2000). Appiah's Uncompleted Argument: WEB Du Bois and the Reality of Race. *Social Theory and Practice*, 26(1), 103-128.

Thomasson, A. L. (2003). Realism and Human Kinds. *Philosophy and Phenomenological Research*, 67(3), 580-609.

Thomasson, A. L. (2007). Artifacts and Human Concepts. In E. Margolis & S. Laurence (Eds.), *Creations of the Mind: Theories of Artifacts and Their Representation* (pp.52-73). Oxford University Press.

Thomasson, A. L. (2014). Public Artifacts, Intentions, and Norms. In M. Franssen, P. Kroes, T. A. C. Reydon, & P. E. Vermaas (Eds.), *Artefact Kinds: Ontology and the Human-Made World* (pp.45-62). Springer.

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

Thomasson, A. (2021). Conceptual Engineering: When Do We Need It? How Can We Do It? *Inquiry*, 64(7), 727-752.

Thomasson, A. L. (2025). *Rethinking Metaphysics*. Oxford University Press.

Thomson, J. J. (1997). The Right and the Good. *The Journal of Philosophy*, 94(6), 273-298.

Tiffany, E. (2007). Deflationary Normative Pluralism. *Canadian Journal of Philosophy*, 37(5), 231-262.

Tilton, E. C. R. (2024). "That's Above My Paygrade": Woke Excuses for Ignorance. *Philosophers' Imprint*, 24(1), 1-19.

Torregrossa, C. (2024). Experimental Aesthetics and Conceptual Engineering. *Erkenntnis*, 89(3), 1027-1041.

Ugelvik, T. (2012). The Dark Side of a Culture of Equality: Reimagining Communities in a Norwegian Remand Prison. In T. Ugelvik & J. Dullum (Eds.), *Penal Exceptionalism? Nordic Prison Practice and Policy* (pp.121-138). Routledge.

Ugelvik, T. (2013). Book Review: John Pratt and Anna Eriksson, 'Contrasts in Punishment: An Explanation of Anglophone Excess and Nordic Exceptionalism'. *Theoretical Criminology*, 17(4), 580-582.

United Nations. (1951). *Convention Relating to the Status of Refugees*. Retrieved July 21, 2025, from https://www.refworld.org/legal/agreements/unga/1951/en/39821

V-Dem Institute. (2024). *Democracy Reports*. Retrieved May 5, 2024, from https://v-dem.net/publications/democracy-reports/

Vaidya, A. (2020). Arthâpatti: An Anglo-Indo-Analytic Attempt at Cross-Cultural Conceptual Engineering. In M. Keating (Ed.), *Controversial Reasoning in Indian Philosophy: Major Texts and Arguments on Arthâpatti* (pp.311-332). Bloomsbury Academic.

Väyrynen, P. (2013). *The Lewd, the Rude and the Nasty*. Oxford University Press.

Vermaas, P. E., & Houkes, W. (2003). Ascribing Functions to Technical Artefacts: A Challenge to Etiological Accounts of Functions. *The British Journal for the Philosophy of Science*, 54(2), 261-289.

Wallace, R. J. (2006). *Normativity and the Will*. Oxford University Press.

Wedgwood, R. (2004). The Metaethicists' Mistake. *Philosophical Perspectives*, 18, 405-426.

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and Epistemic Intuitions. *Philosophical Topics*, 29(1), 429-460.

Weltman, D. (2024). What Do We Want? To Eliminate Gender! When Do We Want It? Later! *Pacific Philosophical Quarterly*, 105(3), 510-540.

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

Willemsen, P., & Reuter, K. (2020). Separability and the Effect of Valence: An Empirical Study of Thick Concepts. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp.794-800). Cognitive Science Society.

Williams, B. (1985/2006). *Ethics and the Limits of Philosophy*. Routledge.

Williams, B. (1995). Truth in Ethics. *Ratio*, 8(3), 227-242.

Williams, B. (2005). From Freedom to Liberty: The Construction of a Political Value. In G. Hawthorne (Ed.), *In the Beginning Was the Deed: Realism and Moralism in Political Argument* (pp.75-96). Princeton University Press.

Williams, M. (2011). Pragmatism, Minimalism, Expressivism. *International Journal of Philosophical Studies*, 18(3), 317-330.

Williamson, T. (2002). *Knowledge and Its Limits*. Oxford University Press.

Wilson, W. (1885/1956). *Congressional Government*. Johns Hopkins University Press.

Wiredu, K. (1995). *Conceptual Decolonization in African Philosophy: Four Essays* (O. Oladipo, Ed.). Hope Publications.

Wiredu, K. (1998). Toward Decolonizing African Philosophy and Religion. *African Studies Quarterly*, 1(4), 17-46.

Wodak, D. (2022). Of Witches and White Folks. *Philosophy and Phenomenological Research*, 104(3), 587-605.

Wright, L. (1973). Functions. *The Philosophical Review*, 82(2), 139-168.

Zack, N. (2002). *Philosophy of Science and Race*. Routledge.

Zalta, E.N. (2001). Fregean Senses, Modes of Presentation, and Concepts. *Philosophical Perspectives*, 15(1), 335-359.

Zettersten, M., & Lupyan, G. (2020). Finding Categories Through Words: More Nameable Features Improve Category Learning. *Cognition*, 196, 539-547.

Zuber, A. (2025). Forward-Looking Concept Functions and the Function/Accident Distinction. *Erkenntnis*, 1-27.