Machine Philosophy

A Foundation of Philosophical Methodology

Dissertation
zur
Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Dilectiss Di Sheng Liu

aus

Dalian (China)

2021

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan: Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter: Prof. Dr. Thomas Sattig
Mitberichterstatterin: Prof. Dr. Claudia Maienborn

Tag der mündlichen Prüfung: 19.07.2021

Universitätsbibliothek Tübingen: TOBIAS-lib

Abstract

This essay designs an evolved architecture for doing philosophy, which I call 'machine philosophy'. I identify the core issues of current philosophical practice as an unhealthy mix of boolean argumentation, obsession with ordinary language, and a lack of methodological clarity. Machine philosophy entails that philosophical methodology should be continuous with that of the sciences. Specifically, philosophical theories are descriptive and objective, and the activity of philosophical theorising should be governed by the norms of statistical learning. In this regard, machine philosophy makes two core claims. One: intuitions are fallible evidence in philosophy, which reflect *objective* facts about our *socio-linguistic realities*. Two: philosophical theories are *descriptive* of our socio-linguistic realities in virtue of being *statistically adequate models of our intuitions*. This new architecture does not demand, but enables philosophical theorising to utilise formal, computational, machine learning mechanisms. However, unlike pluralistic proposals, this enabling of distinct mechanisms places a hierarchy on the epistemic quality of each method, measured by their ability to produce *true* descriptions.

For Truth

Table of Contents

Acknowledgements

I would like to thank the following people for their generous contribution in the form of comments, objections, suggestions on machine philosophy: Adam Finnemann; Christian List; Brandon Fitelson; Bennett Holmann; Leon Geerdink; Ed Zalta; Luis Rosa; Merel Semeijn; Gasper Stukelj; Aleksis Vuoksenmaa; Andrei Rodin; Daniel Tiskin. I would like to thank: Alan Hájek for his comments on applying the strategies of my prototype methodological framework on the knowledge problem and for his encouragement; Aleks Knocks for suggesting the generalisation of my prototype methodological framework to beyond the scope of conceptual analysis; Gregory Wheeler for his inputs on the topic of machine epistemology and pragmatism; Hannes Leitgeb for extremely engaging and helpful discussions, and for being a wonderfully supportive mentor during my studies at the MCMP and beyond; Christian Stegelmann for his various inputs and lengthy discussions with me over the course of my three years in Tuebingen, and for being a faithful and supportive friend; Thomas Sattig for believing in me and for his support and comments on my drafts and presentations. I would like to thank my colleagues in philosophy from Munich, Cologne, Moscow, and Tuebingen for their input on my work. I would like to thank my RTG colleagues and supervisor Claudia Maienborn, my friends, and my family for their support. Finally, I would like to thank my partner Liliana for her unwavering support and warm affection, which kept me sane throughout this final phase.

This work is part of the interdisciplinary *RTG 1808: Ambiguity – Production and Perception*, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – 198647426).

Preface

This essay serves two purposes. One, it argues against a kind of philosophical exceptionalism – the idea that philosophical methodology is distinct from the empirical sciences. Two, it proposes a foundation of philosophical methodology, which I call 'machine philosophy'.

This project has been motivated by a combination of two affairs. First, the utter chaos in philosophical practice and treatises on philosophical methodology over the last few decades. In particular, the conflicted role of ordinary language, the arbitrary distinctions between normative vs. descriptive theorising, the mysterious status of intuitions and their pervasiveness in philosophical practice. These issues have plagued philosophical practice. A few of the most notorious candidates are debates on what knowledge is, whether the mind is physical, whether ethics is an objective matter. The glaring problem is that philosophers cannot seem to agree on anything. It is unsurprising, because philosophy lacks an established set of criteria for consensus. This is especially dire given that the empirical sciences have established a working, though imperfect manual on how to do proper science. The lack of progress in philosophy has been measured against the steady progress in the sciences, which is a direct consequence of an established scientific method and in particular, on the convergence of scientific opinions. Although, the idea of a manual on how to do proper philosophy may *prima facie* seem like an oxymoron. This facade stems from the age-old idea that philosophy is unique, in the sense that it's the only field in which everything can be questioned, and so nothing can be truly established. However, this outdated idea is based on a conflation between epistemic uncertainty and reasonable doubt. We know that theories in the empirical sciences are, by the nature of their justification, epistemically uncertain. However, given our understanding of probabilistic reasoning, we no longer draw the faulty inference that an uncertain proposition warrants reasonable doubt. Our understanding of *being reasonable* has also undergone revision in the late twentieth century from an outdated goal of epistemic certainty to a more pragmatic and wholistic one where being reasonable means being *pragmatically reasonable* – or if one dislikes the term 'pragmatic' – being reasonable with respect to a domain or a goal. For example, one need not be a pragmatist to accept the claim that violating logical closure is not irrational under most circumstances, even in mathematics. This is because of the mere fact that some information is more valuable than others. Given our more refined understanding of probability, there is no reason to think that everything can be *reasonably* questioned, in philosophy or not. The lack of agreement in philosophy is not a virtue, but a vice. It is not a consequence of some unique openness of philosophy, but a consequence of a primitive methodological anarchy. I suspect that the lifeblood of this methodological anarchy comes from its comforting illusion of an unrestricted possibility for creative input. However, we do not make money by amusing our neighbours with witty insights. Given how professional philosophy is funded, philosophy is an epistemic, not artistic enquiry. The goal of philosophy is truth. So if we wish to defend the value of philosophy and display philosophical progress, we simply need to evolve out of this methodological anarchy.

The second affair that motivated me to start this project was the progress of the statistical sciences, both in theory and in practice. In particular, the rapid and widespread replacement of logic based artificial intelligence (AI) algorithms by statistics based machine learning models. I see this as the perfect catalyst for rectifying our methodological blackbox. The basic reason is quite simple: the success of statistical reasoning has bypassed and neutralised the issues of Cartesian uncertainty. Although, I should qualify this statement. Machine learning isn't a recent idea. However, before the early 2000s, data was relatively scarce and expensive, and computers weren't quite as efficient for processing large quantities of data. So machine learning models simply didn't perform well enough to compete with AI algorithms, which are based on explicit boolean instructions. Even now, philosophy clings onto boolean reasoning as the default way of argumentation. The entire validity of the Gettier argument in epistemology has been reliant on boolean reasoning. Take it away and the Gettier cases would fall apart. However, in the last decade, big data became possible as data has become cheaper and more abundant, while computing power multiplied. In the light of these two development, machine learning models began to outperform AI. Furthermore, they have solved issues that AI had long struggled to tackle. Examples include: defeating the strongest chess engine StockFish with AlphaZero; providing reliable natural language translations with Google Translate; and lately within the sciences, solving the age old issue of protein folding with Alpha Fold, or the notoriously difficult partial differential equations with numerical methods that classical AI could only dream of. In some areas such as automated driving, machine learning is indispensable. This problem solving capability of machine learning has shown us that contrary to traditional assumptions, low-fidelity trial-and-error learning can be more efficient and accurate than high-fidelity explicit instructional learning. It has shown us that in general, having more data is better than having smarter algorithms, if the goal of our enquiry is to produce a description with good predictive accuracy for a certain dataset. Producing such a description is the activity of theorising in the sciences. Scientists have learned that the evaluation of a scientific theory are pragmatic: most importantly, the truth of a theory is measured only via generality or predictive accuracy. The best way to achieve good predictive accuracy is to do good statistical learning, even without machines. Evidence should be granted statistical significance rather than be outright accepted or denied. i.e. The success of machine learning has shown us that our age-old obsessions with epistemic certainty, boolean reasoning, fitting our data perfectly are distractions from real philosophical work. And it is the goal of this essay to rectify this situation.

For that end, machine philosophy makes two core claims. One: intuitions are fallible evidence in philosophy, which reflect *objective* facts about our *socio-linguistic realities*. Two: philosophical theories are *descriptive* of our socio-linguistic realities in virtue of being *statistically adequate models of our intuitions*. These two theses will be defended in Chapters 1 and 2 respectively. They enable and encourage philosophical theorising to adopt the method of statistical learning. In Part II of this essay, I explore two prominent challenges that strike not only machine philosophy, but methodological proposals that align philosophy with the empirical sciences. Chapter 3 deals with the notorious Strawsonian concern on topic change. Chapter 4 shows how normative theorising such as ethics are no different in methodology to descriptive theorising. Discussions in Part II of the essay will also clarify a number of confusions that curse current philosophical practice.

Part I. Groundwork for Machine Philosophy

Chapter 1 On Intuitions as Evidence

'One thing that distinguishes philosophical methodology from the methodology of the sciences is its extensive and avowed reliance on intuition.'[1]

1.1 Introduction

This chapter is about using intuitions as evidence in philosophy. I defend the use of intuitions as defeasible evidence. However, my argument neither rests on nor supports a thus far widely held philosophical exceptionalism: that philosophical methodology is epistemically distinct from the empirical sciences. In particular, I reject the idea that armchair philosophy is epistemically unique or *a priori*. Instead, I argue for the thesis that philosophical intuitions are epistemically analogous to scientific observations. Philosophers begin with intuitions as scientists begin with observations. Both intuitions and observations are empirically derived *and* dependent, and serve to statistically corroborate or contradict theories. They are neither the final word nor to be wholly dismissed. Intuitions constitute the data for philosophers as observations constitute the data for experimental physicists.

For two millennia, deductive support had been considered to be the sole proper source of epistemic justification, which resulted in epistemic scepticism. In the early twentieth century, this attitude was challenged by G.E. Moore when he demonstrated that logically speaking, arguments for epistemic scepticism are no better than the ordinary discourse claim that we do have knowledge of the empirical world.[2] Wittgenstein and the ordinary language philosophers pushed further.[3] They shifted the majority attitude in analytic philosophy towards giving absolute epistemic authority to

[1] Goldman (2007).

[2] Moore (1925)

[3] Wittgenstein (1953). Also Strawson (1950) & (1959), where he famously contrasts a 'descriptivist' view from 'revisionary' methods. I will elaborate on this point in Chapters 2 and 3. Black (1950) and Rorty (1967) advocates for what was then known as 'linguistic analysis' (as a precursor to 'conceptual analysis' a la Jackson (1998)). Malcom (1949) argues for the evidential credibility of ordinary language.

ordinary discourse intuitions. On the other hand, followers of Carnap's scientific philosophy[4] or more generally, those who are sympathetic to revisionist attitudes or formal methods, began to cast doubt on the epistemic credentials of intuitions. As a result, a divide has formed between those who take intuitions for granted, and those who see intuitions as unreliable.[5] I think this is a mistake due to an outdated practice where philosophical arguments are inferentially boolean: that something is either reliable or unreliable, that something is either evidence or non-evidence. Instead, in the spirit of Bayesian epistemology and now machine epistemology, philosophical arguments should abandon its boolean structures and update itself in accord with statistical norms. We should evaluate evidence in a more nuanced manner.[6] Instead of considering possibilities, philosophers should consider probabilities. In particular, we should treat intuitions as data that lends itself to statistical inference.[7]

If successful, my argument has three key upshots. One, my argument improves the clarity of what intuitions are. Two, even though my proposal supports the use of intuitions in philosophy, it calls for a change in how we treat intuitions – as evidence which are empirical, defeasible, theory-laden. I believe that this will serve as a middle ground for all sides of the current intuition debate. Moreover, it should revise our philosophical practice – from either dismissing intuitions or taking them for granted, to a more refined and scientific treatment of intuitions as evidence with varying

[4] a la Carnap (1934); (1950a); (1950b), See Leitgeb (2013) & (2020) for a modern rendition of the Carnapian view, esp. on his scientific philosophy. Cappelen (2018) offers a recent overview and defence of revisionist philosophy.

[5] e.g. Cummins (1998); Bealer (1998)

[6] This has no bearing on whether you think there is one true logic and whether that is classical logic.

[7] There is a distinction between 'statistical' vs 'direct/individual' evidence in the philosophy of law literature. This was first introduced in Schoeman (1987). See also Enoch et. al (2012), Blome-Tillmann (2015), and Günther (2021) for the latest discussion on this topic. In this literature, 'direct' evidence refers to evidence resulting from 'direct experience' rather than 'statistical inference'. This has some etymological relation with the distinction between 'statistical' and 'anecdotal' evidence in the sciences. Though in the latter usage, 'statistical' refers to the broader category of 'empirical evidence', rather than evidence exclusively derived from statistical *inference*. The focus in scientific enterprise is more so on the replicability or robustness of a piece of evidence rather than on how that piece of evidence is arrived at (directly or not). Neither of these distinctions are meant to say that 'direct' or 'anecdotal' evidence do not lend themselves to statistical reasoning. In so far as intuitions are concerned, I'm arguing that it's analogous to observational evidence, and hence 'statistical evidence' in the latter usage (in the sciences). The point is that intuition isn't boolean or deductive.

degrees of statistical strength. Third, my thesis will allow for a more refined version of armchair philosophy to be compatible with philosophical empiricism, philosophical naturalism, or experimental philosophy. In fact, my thesis encourages armchair philosophy to work together with empirical and computational methods.

To draw the conclusion that intuitions are epistemically analogous to observations, we need to compare them side by side on three key features: their epistemic source (including sources of error); their roles in theories as both topic and evidence; and how they are necessarily theory-laden. The body of this paper is divided into five sections. In 1.2, I specify the key epistemic parameters for the kind of intuition that is used in philosophical arguments. I argue that intuitions are marked by their lack of non-trivial epistemic grounding apart from a trivial kind of testimony (that intuitions are public, not private). In a similar vein, observations are also void of non-trivial epistemic grounding apart from the similarly trivial testimony (that observations are public). In particular, both intuitions and observations are domain dependent. In 1.3, I argue that intuitions are derived from meaning and concept acquisition, and the intuitions are *of* the entities expressed by the meaning or concept in question. Specifically, intuitions measure our community dependent socio-linguistic realities. In this regard, intuitions are empirical in the same sense as observations are empirical. Observations are ultimately derived from perception (often with instrumental aid), and the observations are *of* the entities shown by the measurement in question, which measure our domain dependent physical realities. In 1.4, I argue that intuitions play a constitutive role for theories in the same way that observations do. We begin the study of what water is via identifying water as is observed by us. Similarly, we begin the study of what truth is via identifying truth as is conceived by us. In 1.5, I argue that intuitions are necessarily theory-laden. In particular, the popular notion of a ‘folk intuition’ is a fantasy that isn’t even desirable. I show that observations are necessarily theory-laden in the same epistemic manner. In 1.6, I conclude that intuitions in philosophy serve the same role as observations in the sciences. They are defeasible evidence in theorising, subject to statistical evaluation. I provide a formal description of intuition as vectors of domain specific parameters.

1.2 What are Intuitions?

1.2.1 The Epistemic Parameters of Intuitions

In the literature on philosophical methodology, the term ‘intuition’ corresponds most closely with the following lexical entries:

> ‘The *immediate* apprehension of an object by the mind *without the intervention of any reasoning process;* a particular act of such apprehension… *Immediate* apprehension *by the intellect alone*; a particular act of such apprehension… In a more general sense: *Direct* or *immediate* insight; an instance of this.’[8]

In the context of philosophical methodology, the ‘immediate’ in the lexical entry has no temporal significance. It has more to do with the lack of ‘any reasoning’, or more accurately, the lack of epistemic grounding for the intuited proposition. The ‘object’ in the context of philosophical practice are propositions. So by ‘intuition’, we mean something along the lines of: ‘the apprehension of a *proposition* by the mind *without further epistemic grounding*, or a particular act of such apprehension.’ The task of this section is to elaborate on this basic description, in the hope of making it precise in a way that's relevant for our argument. Let us begin with some examples derived from philosophical literature:

i. It is irrational to violate elementary logical laws.[9]

[8] Entries 5a, 5b, and 6 of “intuition”, OED, retrieved 5 Dec 2019, emphasis mine.

[9] This concerns the normativity of logic for human reasoning. This proposition was the basis for the conclusion in Wason (1968); Wason & Evans (1983) that people are by default irrational. Nowadays this proposition in its unqualified form is no longer taken for granted, since the idea of rationality itself has been refined, and we take rationality to be broader than *epistemic* or *logical* rationality, or adherence to classical inference rules.

ii. Someone who has not yet seen something *red* does not know *what it is like* to see red.[10]

iii. Torturing a sentient being for fun is wrong.[11]

iv. I cannot know that *p* if I would have believed *p* even if it were false.[12]

v. A proposition cannot be both true and not true.[13]

These propositions are often treated as being *intuitive* in philosophical arguments.[14] For those who take intuitions seriously, such propositions have been used as *ungrounded premises* in arguments. It's clear why attitudes divide on the practice of using intuitions as evidence. On the one hand, such propositions seem undeniably true, and there seems to be no better evidence for the corresponding topics. On the other hand, such propositions are used without epistemic support independent from the fact that there are people who hold such propositions to be true. This dichotomy stands on a blackbox with respect to what intuitions are. Current literature on the nature of intuitions have no one focused domain. We have accounts of intuitions in terms of ontological, psychological, as well as epistemic parameters from different philosophers. One might hold that intuitions are doxastic attitudes such as beliefs, dispositions, judgements: 'Our "intuitions" are simply opinions; our philosophical theories are the same. Some are commonsensical, some are sophisticated; some are particular, some general; some are more firmly held, some less. But they are all opinions.'[15] One might hold that intuitions are intellectual

[10] This was the intuition behind the conclusion in Jackson (1986)'s knowledge argument that the conscious experience involves non-physical properties.

[11] This is used frequently in ethics as an example of a clear ethical wrongdoing.

[12] This is one of the intuitions used in post-Gettier epistemology, as a counter-example to X in X.

[13] This proposition fails to hold in a paraconsistent logic, where a proposition can be both true and not true without an absurd consequence of entailing anything. However, it is debatable whether such logics actually violate this intuition.

[14] Being intuitive is of course agent dependent. A proposition that is intuitive to one need not be intuitive to another. A proposition that is at first intuitive could later be supported or rejected by arguments. That being said, when philosophers talk about intuitions, we refer to those that are more or less shared by the community in question, and those for which we have no non-trivial arguments.

[15] Lewis (1983). See also van Inwagen (1997); Sosa (1998); Williamson (2007) for similar accounts.

seemings, where a person has the intuition that *p* iff it seems so that *p*.[16] One might hold that intuitions are conceptual or linguistic competence: ‘What are called “intuitions” in philosophy are just applications of our ordinary capacities for judgement.’[17] Or, a combination thereof. As far as methodological considerations go, only epistemic parameters matter, since the practice of using intuitions as evidence is a specifically *epistemic* issue.

Again, accounts such as the aforementioned examples sit in distinct domains. We can consistently assert that intuitions in the context of philosophical arguments are beliefs, intellectual seemings, *and* linguistic competence. The claim that intuitions are linguistic competence is a claim about the *epistemic source* of intuitions rather than an outright *intuitions = linguistic competence*. After all, literally speaking, competence cannot *be* intuitions, it can only *enable* or *induce* intuitions. Intellectual seemings on the other hand can also be beliefs. Some might argue that intuitions need not be beliefs (in the cognitive or psychological sense), since one can intuit that *p* without believing that *p*. However, this would be missing the point of the debate; it’s a conflation of the epistemic domain with the psychological one. When we use a piece of intuition as a premise for an argument, we treat the intuition as (if it’s) a true proposition. The intricacies of how we use the term ‘intuition’ in ordinary discourse, *as far as psychology is concerned*, are irrelevant. The intuition must *act* as an object of belief to *epistemically function* as a premise. Whether we believe what we intuit is beside the point. So in the context of assessing intuitions as evidence, it makes no difference whether we say ‘we intuit that *p*’ or ‘we believe that *p*’, as long as we stay on topic. One might object that our intuitions are epistemically more primitive than the basic beliefs we use as premises. For example, one might insist that our intuitions on what’s morally permissible have lead us to believe that iii. ‘torturing a sentient being is wrong’. However, iii. itself isn’t a piece of intuition. My reply is that as long as there is a groundless premise, it is a piece of intuition for all the relevant purposes

[16] Bealer (1998); Boghossian (2009)

[17] Williamson (2004). See also Ludwig (2007); Cappelen (2012).

that intuitions serve in philosophical arguments. As Weinberg et. al would point out: ’Epistemic intuition is simply a spontaneous judgement about the epistemic properties of some specific case – a judgement for which the person… may be able to offer no plausible justification.’[18] The fact that our intuitions in some cases are not explicitly presented has nothing to do with whether they function as basic premises in philosophical arguments. I can present a perfectly valid mathematical proof without explicitly writing down the axioms of set theory. However, the assumption on which the proof’s validity depends is that all parties involved in verifying the proof share beliefs in those basic axioms. If there’s really some mysterious intuition that cannot be explicitly stated, then that intuition might as well not exist. If you cannot tell me why you think that torturing a sentient being is wrong, then I take that you *intuit* that torturing a sentient being is wrong. It is redundant and confusing to postulate some mysterious set of intuitions to further ground the said proposition. Remember, the very conflict on whether we should use intuitions is because they are ungrounded.

To see the evidential credibility of intuitions, what we need is to find the epistemic source of intuitions. It matters not whether we think of intuitions as judgements or as dispositions to believe. That cannot change the *epistemic* source for the underlying propositions. For example, if I intuit that there is a cup on the table because it seems to me that there is a cup on the table, then the epistemic source of my intuition is the combination of my sensory input of there being some entities in some spacial-temporal configuration, my judgement that this configuration constitutes the state of affair that there is a cup on a table, and background presuppositions such as that I am not hallucinating, and that my vision works as expected etc. If I were to use the proposition that ‘there is a cup on the table’ in an argument, it matters not whether I judged it to be so or disposed to believe it to be so. Note that the epistemic source does not constitute epistemic grounding. Because I cannot non-circularly justify the proposition that ‘there is a cup on the table’ with the proposition that ‘my senses and my linguistic competence compel me to say that there is a cup on the

[18] Weinberg et. al. (2001) made a similar point.

table.' The issue of epistemic source will be elaborated upon in the next section. For now, the point is to argue that ontological or psychological parameters are irrelevant for the purpose of judging whether intuitions constitute legitimate evidence. Sure, some might argue against the use of intuitions as evidence on the ground that they are private – that the epistemic source of an intuition is a person's individual psychology. However, this is no different from an argument against (especially direct, but also indirect) observations on the ground that observations are essentially in our mind. The fact that observations are mental states has nothing to do with where the *content* of our observations come from. The fact that intuitions are mental states has nothing to do with where the content of our intuitions come from. Of course, I have nothing to say to a diehard idealist, for the diehard idealist also doubts scientific practice. On the other hand, a scientific realist might argue that while observations in science are grounded in some objective state of affairs, intuitions in philosophy generally have no such grounding. Knowledge just is fundamentally a different entity from water. The latter is a real substance *in the world*, whereas we cannot, so to speak, physically point at knowledge. So, the realist asks: how can intuitions also have objective truth makers? The short answer is that mathematical propositions are objectively true, if anything is objectively true. The fact that numbers do not physically exist has nothing to do with the objectivity of propositions about numbers. The long answer is that our intuitions of various propositions *are* grounded in an objective reality, in the sense that our socio-linguistic practice is real. This is not linguistic trickery. I mean this very seriously, in the sense that our communities have acquired the *correct* use of terms such as 'know', 'good', 'beautiful' from our common categorisation of such states. The epistemic state that we would ordinarily call 'knowledge' is as real and objective as the event of the French revolution, or as the tea in our cup. The fact that there are disagreements between communities on what should be called 'knowledge' is as much a disagreement as that on which entity should be given the privilege of being a 'planet', a 'donut', a bottle of 'cognac' etc.

1.2.2 *What Intuitions are Not?*

Before specifying the epistemic parameters for 'intuition', I make three preliminary clarifications, in the spirit of doing *feature selection*.[19] These are on the exclusion of analytic propositions; on pseudo-intuitions; and on the ambiguity between intuiting vs. the intuited.

We should exclude analytic propositions from our discussion. Analytic propositions such as 'vixens are female foxes' or v.: 'a proposition cannot be both true and not true' are intuitive if by intuitive we simply mean 'appears true'. However, they do have clear epistemic grounding – they are true in virtue of the meaning of their constituent terms.[20] The source of their truth is clearly verifiable, *independent of the fact that they are intuitive*. To be clear, it isn't an issue to *say* that analytic propositions are 'intuitive'. However, analytic propositions are clearly legitimate evidence, *because they are analytically true*. On the other hand, it is debatable whether the kind of intuitions under scrutiny are legitimate evidence, and that is because we cannot identify their truth-makers. So the target of dispute should exclude analytic propositions.

To follow up on the exclusion of analytic propositions, I further clarify the notion of intuitions by excluding pseudo-intuitions. A pseudo-intuition is a proposition that *prima facie* appears like any other intuition, but has an explicit epistemic grounding independent of someone having the intuition. For example, *i* from 1.2 is a pseudo-intuition. The holding of *i* requires assenting to the proposition that *a person is rational only if the person acts in accord with holding all and only true beliefs*. Harman's rejection of Wason's conclusion was grounded on a rejection of this

[19] Feature selection is a statistical technique for reducing the number of parameters for a given dataset. In this case, I'm removing some problematic parameters that could easily corrupt our account of what intuitions are. More on feature selection will show up in the next chapter. For now, I'm just doing a tongue-in-cheek demonstration of what machine philosophy is all about.

[20] I take the ordinary language meaning of 'not', 'and', 'proposition' to be sufficient for determining the truth of v. Debates on whether non-classical systems use 'not', 'and' with different meanings are irrelevant here. Discussions become pointless if we begin to conflate ordinary languages with formal languages.

presumption.[21] What Wason did was to make the mistake of equivocating rationality with a narrow form of *epistemic* rationality. Although, oftentimes it is practically *ir*rational to hold certain true beliefs, given our cognitive limitations such as computational capacity or memory. Examples of pseudo-intuitions are everywhere. Mathematical intuitions on whether a proof is valid are pseudo-intuitions, since the validity of a proof is to be grounded in first-order logic. The pre-scientific intuition that the earth is flat is a pseudo-intuition, since it is based on the false presupposition that what we experience extends to larger scales. Clearly, pseudo-intuitions are not the topic of our discussion. They are supported by independent propositions. So the question of whether they constitute legitimate evidence is simply a question of whether they are conclusions of sound arguments. In contrast, the proposition that 'justification is defeasible' *is* intuition proper. The only reason we believe this proposition is the fact that in ordinary life, we often grant epistemic justification with uncertainty. In other words, justification is defeasible because we often take justification to be defeasible.

When we talk about using intuitions as evidence, we could mean one of two things. First, we could refer to the propositional content of an intuition that *p*. Second, we could refer to the fact that *someone has the intuition that p*. This ambiguity is substantial just in case we wish to endorse using intuitions as evidence. There are two reasons for this.

One, if we want *p* to function as a premise, we must assent to the claim that *p* is true *of something.* So the reasons for *p* and for 'having the intuition that *p*' can differ. Two, while philosophical theorising largely operate on the assumption of treating *p* as the premise, rather than 'having the intuition that *p*', it often seems that we argue for the use of *p* by referring to 'having the intuition that *p*'. Some philosophers defend this practice on the grounds that *p* is an *intuition*.[22] While this line of argument can

[21] Harman (1986)

[22] See Goldman (1999a), (2007) for arguments along these lines. Kripke (1980) provides a famous example of explicitly arguing that *intuiting* that *p* is the strongest evidence for believing that *p*. i.e. The intuition that *p* is justified by *having the intuition that p*.

risk overtly psychologising intuitions,[23] it seems that a defence of using intuitions as evidence must address this phenomenon. The problem is, the two individually reasonable consequents come into conflict. Treating *p* as a premise entails that the source for *p* and for *having the intuition that p* could be distinct. But the fact that *p* is an intuition appears to entail that 'having the intuition that *p*' is the one and only epistemic grounding for *p*.

In the next sub-section, I argue that this apparent conflict, and the underlying ambiguity could both be neatly resolved by having a clear account of 1. the target of intuitions (what *p* could be true of) as community-dependent socio-linguistic realities, and 2. the source of having a proper intuition and the source of the intuitions themselves as both socio-linguistic. In this regard, the epistemic relation between community-dependent socio-linguist realities and proper intuitions or having proper intuitions is the same kind of epistemic relation between physical reality and our proper observations or having proper observations. Consequently, the role of agency for assessing the epistemic credentials of an intuition is akin to that for assessing the reliability of an observation. In both cases *intuiting/observing that p* can affect the quality of the data, but neither are *the* grounds for that *p* (unless, again, you are an idealist). To keep things clear, when I mention 'the intuition that *p,*' I henceforth refer to the propositional content of *p*, in the context of *p* being a premise. If I need to refer to 'intuiting that *p*', I would say so.

1.2.3 *Intuitions are Community Dependent*

Suppose I have an intuition that *p*. Then my intuition that *p* is dependent on my correct understanding of the constituent terms of *p*, on certain values that I uphold, on certain beliefs or dispositions that I may have etc. For example, a person who intuits that 'torturing a sentient being is wrong' would understand the proposition in a specific way. 'Torturing a sentient being' would refer to an act that causes pain to a being who is able to experience pain, and if we elaborate further, anyone who

[23] Pust (2001)

understands the phrase correctly could also *imagine* or *conceive* of such a scenario. ‘Wrong’ would be understood in the context of the intuition as ‘moral wrongness’ or simply ‘immoral’ or ‘unethical’. A person who understands what it means to be unethical would associate unethical acts with acts that one *ought not to do*.[24] However, merely understanding the terms would still be insufficient for having the intuition. In particular, it would be insufficient for categorising the act of torturing a sentient being as moral wrongdoing. For example, one might have a value system wherein not all sentient beings are to be given equal moral considerations. Then there could be scenarios where one simply does not intuit that ‘torturing a sentient being is wrong’.[25] Both the understanding of the terms and the value system are jointly necessary for having the intuition that ‘torturing a sentient being is wrong’. These are community dependent factors.[26] In this particular example, the community in question could be within a certain anglophone community. Perhaps if we allow for linguistic cognates, we can say that the community is one that upholds certain values, beliefs, categories etc. With this community dependence in mind, let’s make a new attempt at characterising intuitions.

At the core, the question of how intuitions confer epistemic warrant is a question of how seemingly true propositions that have no non-trivial justification confer epistemic warrant. This understanding is a refinement over the ordinary language conception of intuition as: ‘the apprehension of a proposition by the mind *without further epistemic grounding*’. The refinement is to exclude analytic propositions and pseudo-intuitions, which can both be understood as ‘intuitions’ in ordinary discourse.

[24] Of course the converse isn’t true, but we need not go into a corpus study of what ‘immoral’ or ‘unethical’ means in discourse. The point is that anyone who understands the proposition has a certain conception that makes itself obviously plausible for the person assenting to the proposition, which in effect makes the proposition an intuition.

[25] Williamson (2007) raised a similar but even stronger point about how people can fail to assent to analytic propositions in virtue of holding on to certain theoretical beliefs. e.g. A person who correctly understands all the terms of the proposition ‘all vixens are female foxes’ can still rationally refuse to assent to the proposition. This is because the person might, say, believe that 1. vixens don’t exist and that 2. we cannot quantify over empty sets.

[26] It matters not whether language can be private. Even if there were to be a private language, the bearer of that language could mark her own community, or be a member of another community in virtue of sharing certain values or categories.

Here is an attempt at an epistemic characterisation of intuitions in the sense of being a seemingly true proposition with no non-trivial justification:

A proposition *p* is a piece of intuition iff

Ia. a community *C* would use *p* as a premise for an argument; *and*

Ib. there are no justification for *p*, except for the fact that the common discourse/ practice of *C* makes accepting *p* reasonable.

Here a *community* is to be understood as a group of people who share a particular linguistic or conceptual practice. The delineation of a community is dynamic. Philosophers delineate a community. Anglophones delineate a community. Mathematicians delineate a community. The community that is relevant for a piece of intuition *p* is the community in which an argument using *p* would be presented and understood. For example, philosophers have little trouble agreeing on the basic Gettier intuitions: that the agent in the Gettier case has a justified true belief of some proposition *p* without knowing that *p*. However, once you take it outside of the philosophy classroom, you would find puzzled faces among academics and non-academics alike.[27]

Condition Ia restricts the domain of intuitions to those that a particular community would use in arguments. This condition eliminates idiosyncratic intuitions, but allows for communities to disagree over a particular piece of intuition. This is what we want. The intuitions that philosophers take seriously in arguments are the ones shared by a certain community. This community is usually a subfield of philosophers. However, there is clear evidence that such intuitions could be rejected by outsiders.[28] Sometimes, a community could even be internally divided on a particular piece of intuition.

[27] Weinberg et. al (2001) shows how beyond the philosophy classroom, folks would systematically disagree with some core Gettier intuitions. There has since been ongoing debates about the significance of the experimental results in Weinberg et al (2001). e.g. Nagel (2012); Seyedsayamdost (2015). Stich (2013) replies. It's not my intention to dwell on this debate. The bottom line is, Gettier intuitions aren't accepted across different communities. At the least, in my various conversations with mathematicians on the Gettier issue, they simply do not buy into the Gettier story, and this failure to accept the Gettier intuitions isn't idiosyncratic. In any case, I will demonstrate how this isn't an issue for using intuition as data.

[28] Ibid.

The first clause of condition Ib further restricts propositions used as premises to those that are ungrounded. The second clause identifies the minimal condition on which a community's intuitions depend. Perhaps trivially so, the very fact that different communities could assign different evidential values to a piece of intuition in the context of an argument is indicative that their intuitions are tied to their epistemic discourse/practice. What I mean by epistemic discourse/practice can be understood roughly as a Carnapian linguistic framework.[29] Physicists operate within a certain linguistic framework, in which certain propositions are taken to be true, and certain forms of reasoning permissible. Investors operate within another linguistic framework, in which different forms of reasoning are taken to be permissible, and a different set of propositions true. To illustrate this idea with an example, think about the basic Gettier intuition. If you are a traditional epistemologist, you would likely take the Gettier intuition that the agent has a justified belief for granted. On the other hand, a mathematician or a non-academic might be puzzled. The reason is that when you operate within traditional epistemology, you have background assumptions about what counts as being justified, what counts as being knowledge etc. that other communities might not share. For example, our implicit criteria for being justified is more lenient than that of a mathematician, while our implicit criteria for knowledge is stricter than that of a non-academic. We can in general say that the discourse and practice of a community is what distinguishes its set of intuitions from those of outsiders.

This characterisation would of course raise the question of whether such 'implicit criteria' just are presuppositions (albeit implicit) in the sense of sentence *i* from the examples above. If so, then all intuitions would be pseudo-intuitions, since they would be grounded on other propositions. My reply is this: such implicit criteria are *not* propositional attitudes. They are the product of a community's linguistic framework in the following sense. The *practice* of an epistemologist – working within

[29] a la Carnap (1928, 1934). Although, my framework needn't be *linguistic* in any salient sense, it could be conceptual or some other kind if the reader wishes. I shall take it to be linguistic throughout this text for simplicity of illustration.

a certain tradition of solving certain problems – has conditioned the epistemologist to *simply see/judge/believe/agree* that the agent in the Gettier case has a justified true belief. There is no specific proposition *p* that is derived from the epistemologist's practice, such that *p* would ground the claim that the agent in the Gettier case has a justified true belief. If you were to ask an epistemologist why she believes that the agent is justified, she might tell you that 'under ordinary circumstances, we would take the belief to be justified.' This is at best a circular reasoning for the intuition itself. It is just another way to say 'our intuitions agree'. If you were to ask her about why the agent has no knowledge, she might tell you that 'the agent doesn't know, because the belief could have been false, or that the belief was only true by luck etc.'. In this case, the epistemologist is invoking epistemic intuitions that arose from a tradition of philosophers writing about knowledge. i.e. Although the proposition that 'the agent has no knowledge' isn't a piece of intuition, the proposition that 'I cannot know that *p* if I would have believed *p* even if it were false' is. You might like to ask then why epistemologists espouse so-called 'folk intuitions' for justification. This is itself due to a philosophical tradition. It is due to the wide acceptance of ordinary language philosophy, and various ways of dealing with the Cartesian sceptic by appeal to folk discourse in 20th century analytic philosophy. Since then, philosophers have been trying to study terms/concepts/things *as they are conceived by the folk*. If philosophers in the 19th century were to read Gettier's paper, they would more readily dismiss Gettier's argument simply because they think that the agent's belief wouldn't be justified. As for why epistemologists adhere to folk intuitions re justification but not re knowledge, the answer is this: Gettier's story was a response to Plato's thesis that a true belief alone is not knowledge. In ordinary practice, we often equate *knowledge* with *true belief*. The very notion of knowledge in the Gettier stories must be sensitive to Plato's conception of knowledge, or else it would change the topic. It is not the literal extension of 'knowledge' in ordinary discourse. The claim that it *is* the study of the folk conception of knowledge can be defended quite easily by claiming that the topic in question is a particular sub-category of what folks would call 'knowledge', rather than a modification of the folk conception.

When we theorise about some X in some domain D, we can provide either a complete or a partial characterisation of X in D. The theory that water is H_2O is a complete characterisation of water in chemistry, while the theory that $F = ma$ is a partial characterisation of force in physics, since the characterisation is a particular description of the amount of force that acts on a resting mass when the mass is accelerated. There are other situations where force requires a different physical characterisation, such as the gravitational force between two objects. In philosophy, we like to talk about pre-theoretic extensions as products of folk intuitions. Given our characterisation of what intuitions are, it is easy to see how intuitions contribute to our pre-theoretic extensions. There are two ways in which this happen. One, our intuitions in the form of explicit propositions such as i–v. provide the criteria for delineating the extension of some pre-theoretic 'X'. Two, our linguistic practice prompts us to simply identify some entities as X and others as not-X, without providing us with an explicit reason for doing so.

1.3 What Do Intuitions Measure?

Observations measure physical reality. It's not so obvious what intuitions measure. For those who use intuitions extensively in philosophical arguments, intuitions measure a kind of objective, or at least intersubjective reality; though it's unclear if not physical, what kind of reality this might be, and how we can establish it as a legitimate subject matter. One of the most popular candidates is to say that intuitions are about public concepts. This is the default position by those who are engaged in conceptual analysis, or more generally, those who think that the subject matter of philosophy is concepts. For these philosophers, intuitions measure our collective conceptual reality.

Williamson has advanced a position wherein he supports the *methodology* of traditional armchair philosophy, but argues against their ontology and claims that philosophy isn't about concepts or language, but *things*.[30] For example, when

[30] Williamson (2007)

philosophers study *knowledge*, they're not merely studying the *concept of knowledge*, but *of knowledge* itself. This position defends a continuity between the empirical sciences and philosophy in virtue of their subject matter. However, the deadly issue with this approach is that it forgets about the domain dependency of theories. When a chemist studies *water*, she aims to *chemically* describe water. And *for this purpose*, she would have to examine *the chemical composition of the thing* that we call 'water' in English. However, when a philosopher studies *knowledge*, she aims to examine the *epistemological features* of knowledge, rather than neurological, sociological, psychological features. And *for the purposes of epistemology*, there is absolutely no difference between *the thing* we call 'knowledge' in the epistemic realm (whatever that may even be), and the term 'knowledge'.[31] Therefore, I take the exact opposite approach – I argue for a methodology that's continuous with the sciences, but a subject matter that's distinct. In particular, I think philosophy should study our community dependent linguistic/conceptual realities, but with statistical methods rather than treating philosophical enquiry as a priori. Therefore, agreeing with the linguistic/conceptual analysts solely on this point, I maintain that intuitions measure community-dependent social-linguistic realities. I argue that the sources of our intuitions are the sources of our linguistic competence, and intuitions are thereby empirical. We can then establish what it means to say that our intuitions measure community-dependent socio-linguistic realities.

1.3.1 Intuitions are Empirical

In section 1.2.3, I argued that intuitions depend crucially on our adherence to a particular linguistic framework – a community's ways of expression and reasoning. I now argue that the epistemic sources of our intuitions are the epistemic sources of our linguistic competence. This linguistic competence is our competence to participate in a certain community's discourse, or our fluency in a particular community's ways of

[31] Or Cappelen's apparently safe term 'representational devices', which I think is unnecessary and rather problematic, as 'representational device' is far more ambiguous than simply 'terms' or even 'concepts'. See Cappelen (2018).

expression and reasoning. The sources for these competences are empirical, not a priori. We familiarise ourselves with a community's ways of expression and reasoning via a gradual process of learning its practices and repeating it ourselves. We improve our mastery by obtaining experience, not by reading a manual on the foundations of a subject. A student who merely masters the theory of piano performance would not develop the skills to play the piano well. The student must *practice* in order to play well. A student who simply masters the knowledge of how a language works, even with a total knowledge of vocabulary and grammar, would not be able to express herself at will or notice grammatical errors without conscious effort. Fluency can only be achieved thorough repeatedly *using* the language. A student who masters every textbook on the foundations of mathematics would not be proficient at identifying a provable or refutable theorem and then designing a mathematical proof. The student must *exercise* proving mathematical propositions. Our intuitions are ultimately due to how we have acquired our languages and adopted social norms. Here both 'languages' and 'social norms' are to be understood in a broad sense: languages could include specialised languages; social norms could be within specialised domains, such as the norms of doing mathematical research or playing chess. In other words, our intuitions are obtained via a machine-learning-like low-fi procedure, where our 'machine learning algorithms' train on our socio-linguistic experience.[32] Instead of arguing for this top-down, I illustrate why (non-analytic) intuitions are substantially empirical with the mathematical intuition of there being infinitely many natural numbers.

Very little of what we now call 'mathematical intuitions' are intuitions proper. Most of them are pseudo-intuitions. The aforementioned intuitions on the validity of proofs are pseudo-intuitions, since their validity are grounded on first-order logic. Even the axioms of ZFC set theory are not intuitions proper. Some of the axioms are analytic, while others have been formulated to meet certain desiderata. For example, the axiom of extensionality is analytic, since it is in the very conception of a set that

[32] The term 'low-fi' vs. 'high-fi' is from Sterelny (2012). Low-fi learning is learning via trial and error, and high-fi learning is learning via explicit descriptions of rules. The way that we acquire intuition is very much in the spirit of Skyrms (2010), in which he discusses trial and error learning for meaning acquisition.

two sets are the same just in case they have the same elements. On the other hand, the axiom of infinity is not analytic, since the conception of a set itself does not dictate whether there exists infinite sets. Rather, the axiom of infinity was formulated to meet the desideratum of allowing for the existence of infinite sets, which is needed to formally ground intuitively true propositions such as there are infinitely many natural numbers. Now why do mathematicians hold such a desideratum? Well, because the desideratum is intuitive. First, the desideratum cannot be non-trivially grounded. It is logically equivalent to the intuition that there are infinitely many natural numbers. There is no justification for this intuition other than the fact that most mathematicians hold the intuition.[33] Second, the intuition is not analytic, since it is not in the very idea of natural numbers that there are infinitely many of them.[34] For example, some societies such as those of the Australian aborigines have no way of distinctly

[33] Gödel and Russell thought that the axioms of mathematics are true *because* they serve to prove intuitively correct mathematical propositions. So although *within mathematics*, the proposition that there are infinitely many natural numbers is justified *by* the axioms (including the axiom of infinity), the epistemic grounding should really be the other way around – the axiom of infinity was construed to satisfy the intuition that there are infinitely many natural numbers. This example shows that mathematical proofs are epistemic justification for mathematical propositions in a two-directional manner. i.e. The success of a proof could be evidence for certain axioms (such as the axiom of infinity), or for the proven proposition (whether we can square a circle). Which direction the epistemic justification goes depends on whether the axioms or the conclusion hold more epistemic priority. This priority could be, for example, whether the conclusion is intuitively true. If an axiom disagrees with an intuitively true mathematical proposition, then mathematicians could reject the axiom for that reason. If a mathematical foundation fails to decide on widely held mathematical propositions, then we reject the system, not the propositions. In general, one should caution against seeing mathematical proofs as epistemic justification for the mathematical propositions that they prove. In most of the cases, mathematicians undergo the task of finding proofs for propositions they already believe to be true, via some other reasoning within mathematics (e.g. that the square root of 2 is irrational, the Goldbach conjecture).

[34] Set-theoretic definitions of natural numbers are *theories*, not definitions in the sense of a dictionary definition, which are topic fixing (more on this in the next section). They are akin to the chemical definition of water as H_2O, which are after all, empirically discovered, by fixing the reference *based on* the (ordinary language) definition of water. Furthermore, the modern set-theoretic definitions of numbers are constructed *post* the acceptance that there are infinitely many natural numbers. So the latter clause cannot be analytically grounded on the set-theoretic definitions. Rather, the definitions (theories) are grounded on these intuitions.

expressing numbers larger than ten.[35] For someone outside of the mathematical community, there might as well be finitely many natural numbers, since after all, the universe has finitely many things.[36] So how did mathematicians develop the intuition, which we nowadays take for granted, that there are infinitely many natural numbers?

In the late 17th century, the need for computing quantities relating to continuous change forced mathematicians to use the idea of infinity, which enabled calculus.[37] Since then, mathematicians have gradually grown accustomed to the notion of infinity. However, the widespread intuition that there are infinitely many natural numbers is a recent development. It required the notion of infinite sets, which was introduced and studied by Cantor. Before then, there was no proper definition of 'infinity' in mathematics.[38] It would have therefore been imprudent for mathematicians to talk about infinitely many natural numbers when the ascription was meaningless. For instance, when Leibniz developed calculus, he considered infinity to be an ideal entity, different in nature from the countable numbers.[39] This conception contradicts the modern notion of infinity. We now know that the cardinality of natural numbers is *countably* infinite. Furthermore, we also know that infinity comes in different sizes with the cardinality of natural numbers being the smallest infinity. So the development of the intuition that there are infinitely many natural numbers was gradual, and the process was empirical. It was not based on a priori reasoning[40], but a

[35] This does not entail that they have no way of counting more than ten items. https://aiatsis.gov.au/sites/default/files/e_access/serial/m0005975_v_a.pdf In fact, in each *natural* language there is an upper limit on the expressibility of a number. If one recalls the anecdote about Bertrand Russell's peculiar IQ test performance, we see that the exprhssibilty of n-polygons is an extension of this limitation. In modern times, we have learned to create new expressions on the fly for large numbers, such as a 'googol'. This is not to suggest that the limit of language is the limit of thought. There are expressions such as 'Rayo's number', though it's no longer in the domain of natural language, since its definition depends on set-theory.

[36] The English definition of a 'number' is 'a word or symbol that represents an amount or a quantity'.

[37] This in turn enabled solutions to previously unsolved problems, such as Zeno's paradox. More concretely, the introduction of infinitesimals dissolved the apparent conflict in Zeno's puzzle.

[38] Allen (2003)

[39] Jesseph (1998)

[40] Currently the default view is that intuitions are a priori. It seems that we *simply* have them. Of course, this is absurd once we realise that intuitions can differ between communities, and there are concrete reasons why they differ.

process of acquainting oneself with the *practice* and the *language* of the mathematical community as it develops. In this particular case, the intuition that there are infinitely many natural numbers became mainstream as mathematicians became more accustomed to the practice of talking about infinities, using them in proofs, continually refining the very notion. This process is akin to how the musical community in the western world have grown accustomed to atonality over the course of the late 19th to early 20th century due to composers such as Scriabin and Schoenberg; or how the anglophone community have grown accustomed to the idea that a 'painting' could be non-referential over the course of the 20th century due to painters such as Kandinsky.[41] The result is that, the western musical community has acquired the intuition that a tonal centre is unnecessary for music;[42] and the anglophone community nowadays takes for granted the intuition that a painting need not be referential.

Here I must clear an imminent misunderstanding. One might point out that Cantor introduced and drew conclusions about infinite sets on a priori grounds. Then how could the resulting intuition be empirical? Here is my reply: Because the grounds on which Cantor introduced infinite sets are not the source of our intuition that *there are* infinite sets. The axiom of infinity is an *axiom* within mathematics.[43] Therefore it cannot be grounded within mathematics, with all the methods of mathematical proof. This includes any internal grounding such as consistency.[44] Of course, there certainly are external grounds for the axiom, such as considerations of theoretical simplicity, power, fruitfulness. However, these external grounds are not the reasons why modern

[41] Of course, this change in the notion of painting was not exclusive to the anglophone community. Here I simply use the anglophone community as an example since the term 'painting' is English.

[42] As expected of a piece of intuition, this has been disputed by a minority of musicologists on theoretical grounds. For example, Westergaard (1968) argues that listeners will hear tonality in everything, though they could be mistaken.

[43] Meaning that other axioms within mathematics are insufficient to prove the axiom of infinity. This of course doesn't meant that the axiom of infinity cannot be proven using mathematical propositions. However, this would

[44] If consistency were a reason, we could prove the axiom using *reductio ad absurdum.* Although we do indeed prove the position that 'there is no largest prime number' by appealing to consistency, this by itself cannot entail the axiom of infinity.

mathematicians have the *intuition* that there *are* infinite sets. Instead, our intuition comes from the fact that a significant portion of what we know as established mathematics depends on the existence of infinite sets. This dependence in turn happened over time, due to the gradually increasing acceptance and impact of set theory. Similarly, Scriabin transcended tonality on explicit grounds. For example, he introduced musical devices such as the incomplete spans or the 'mystical chord' on mystical, aesthetic, and philosophical grounds.[45] However, these grounds cannot be and is not the source of the modern intuition that music need not have a tonal centre. Apart from the fact that many of Scriabin's beliefs are not shared by the wider community, the intuition required the acceptance by the musical community that Scriabin's works are indeed atonal, and that they are indeed music. This acceptance is the source of our intuitions, and it is empirical.

Now a second objection might follow: If Cantor had explicit grounds on which he introduced infinite sets, and if Scriabin had explicit grounds on which he transcended tonality, then why aren't our intuitions pseudo-intuitions? i.e. Aren't our intuitions simply grounded on those of Cantor and Scriabin, even if those are not the reasons for which we have the intuitions? Here is my response: The groundings for Cantor or for Scriabin are external to the linguistic framework of the community in which they are intuitive. Therefore, they cannot *epistemically* ground the propositions that we take to be true *within* those frameworks. For example, the axiom of infinity is not and cannot be epistemically grounded on considerations of theoretical simplicity or fruitfulness. To use simplicity as an example: We do not think that a theorem is true *because* it's simple. We certainly *prefer* simpler theories, but simplicity is not an epistemic, but a pragmatic consideration. In general, all external considerations can be thought of as pragmatic groundings.[46] This is why the axiom of infinity is an *axiom*. It must be *stipulated* within mathematics. There is no proof for it. Similarly, Scriabin's reasons

[45] Baker (1980)

[46] For a diehard pragmatist, it could seem that pragmatic groundings are legitimate reasons for believing a proposition. i.e. pragmatic groundings *are* epistemic groundings. In that case, there are no intuitions, and the diehard pragmatist can happily go about their business of using what most people would call intuitions in their arguments, and stop here with the current paper.

for transcending tonality are not epistemic groundings why a piece of music need not have a tonal centre. The reason is simple: by the time Scriabin wanted to transcend tonality, the very definition of music required tonality. So Scriabin's project was not musical, but meta-musical – he contributed to the revision of what 'music' denotes. However, the revision only succeeded because the *musical community* approved of such a revision. This decision of revision was neither a priori nor musical, but pragmatic, since the revision could not have been grounded on established musical theory, but on external considerations of innovation, aesthetics etc.

The distinctions between internal vs. external grounding and that of the justification vs. the source of our intuitions can be subject to confusion. So before moving on, I shall provide a structural summary of the difference between the three items: the reasons for *conceiving/introducing* a proposition *p*, the reasons for *intuiting* that *p*, and the reasons *for* that *p*. Another way to describe these three items are the external grounding/justification for *p*, the empirical sources that give us the intuition that *p*, and the internal grounding/justification for *p*. If one prefers the locution of 'the context of discovery vs. the context of justification', one more way to delineate them are the context of *discovery*, the context of *belief*, and the context of *justification*. If *p* is *not* a pseudo-intuition, then there is no justification for *p*. Now to use a concrete example to illustrate the structural dependencies between the three, let us consider the physical proposition that 'an aircraft can stay aloft.' Cayley developed the idea of a flying machine that became the basis for modern planes in the late 18th century by introducing four vector forces: thrust, lift, drag, and gravity, based on experiments. However, the physics of an aircraft's lift is still unsolved. i.e. No one knows how an aircraft stays afloat, only that it appears to do so. Now, our intuition that an aircraft can stay afloat is obviously neither because we have read on Cayley's ideas nor because we know of the various hypotheses on lift. Rather, it is because we have grown accustomed to flying airplanes!

1.3.2 Intuitions Measure Community-Dependent Socio-Linguistic Realities

A key feature of observation is that it can be *false* or *invalid*. An observations is invalid just in case it fails to correctly measure reality. For example, if my thermometer reads 37°C while in fact the entity measured has a temperature of 40°C, then my observation via the thermometer has failed to measure physical reality, and the observation is thereby invalid. By establishing a target of measurement for intuitions, we likewise can say that an intuition is valid or invalid in virtue of correctly measuring its target reality.[47]

One might doubt the parallel between observations and intuitions in the following way. While there is only one physical reality, there are multiple community-dependent socio-linguistic realities. If there are a multiplicity of realities for our intuitions, then how can philosophers decide which realities are the 'right' ones to study? In particular, it seems to tread dangerously on the thread of reducing philosophy to lexicography. After all, if my intuition of knowledge is only meant to correctly measure a particular community's epistemic conception of knowledge, then how can my theory of knowledge have any general significance? Moreover, it seems that my intuition that 'there are infinitely many natural numbers' doesn't merely captures a mathematician-dependent *socio-linguistic* reality, but a *mathematical* reality that's community-*independent*.

To answer these worries, we need to distinguish a *domain* from a *community*. A domain is an area of study, such as mathematics or physics. We can think of a domain as a linguistic framework with its rules of reasoning, values, presuppositions etc. A community is a group of people who share a common set of socio-linguistic facts. For every domain there is a corresponding community who pursue epistemic activities within the domain. Trivially, a domain is established by the socio-linguistic rules of its community. The domain of mathematics is established by the community of mathematicians, with what they think are valid rules of inferences, good practices,

[47] Whether that reality is objective can complicate things a little, and will be discussed below in section 1.5. For now, it's enough to clarify that there is a parallel between observation and intuition with regard to measuring their respective reality.

sound presuppositions etc.[48] However, not every community has a corresponding domain. We can have a community of *anglophone mathematicians* without the possibility of the community establishing a valid domain, since mathematics happens to be a natural-language independent game. i.e. English plays no epistemic significance in mathematics.

Hitherto we have suggested that domains and communities are interchangeable, because they are *if* we restrict ourselves to only communities that correspond to a domain. However, a problem arises once we begin to consider communities such as 'anglophone mathematicians' or 'fountain pen lovers'. Consider the last worry about the intuition on there being infinitely many natural numbers. Going by what I claim, we'd have to say that this intuition measures a mathematician-dependent socio-linguistic reality. The complaint is that while a socio-linguistic reality is community dependent (i.e. it can change over time, it isn't objective etc.), a mathematical reality shouldn't be. However, recall that 'socio-linguistic' just is the set of specialised practices and languages of its corresponding domain (when such as domain exists). In particular, a socio-linguistic reality within the community of mathematicians demarcates exactly the mathematical language and rules of reasoning within mathematics. In other words, the language and rules of mathematics establishes the game or domain of mathematics, so to say that our intuition is about the mathematician-dependent socio-linguistic reality *just is* to say that our intuition is about mathematics. A similar point would hold for ethics, epistemology etc. To say that our epistemological intuition is about the socio-linguistic reality of epistemologists is to say that the intuition is about epistemology. Because trivially, by the above formulae, epistemology is a domain that is governed by the socio-linguistic rules of epistemologists. So to avoid problems, we simply do not consider domain-less communities. This is more than easy to do since we're only here to talk about intuitions that play some evidential role in a domain.

[48] There have been plenty of examples when mathematicians have (attempted) to revise the very domain of mathematics, from its language (introducing the foundations of mathematics with set theory) to its practices (Bourbaki's insistence on algebraic or verbal proofs, rejecting visual proofs invalid).

One might jump then to the conclusion that I am endorsing a kind of anti-realism, since it sounds as if I am saying that the realities we describe *just is* our constructions thereof, or even 'language games'. However, this would be a terrible misunderstanding. I am making no metaphysical commitments here. The point is purely epistemic and pragmatic. The way to think about it is to think about mathematical *practice*. The *method* with which a mathematician studies mathematical objects is like analysing a much more complicated version of chess, studying the 'best moves' etc. There is no doubt that the studies are objective, but ultimately, the manners in which the mathematicians study mathematical objects is via manipulating the mathematical language with certain rules. This has nothing to do with whether there exists mathematical objects. Perhaps a better example is to think of philosophy, where the entities under study usually do exist in some physical form (e.g knowledge is also a psychological state). In no way do I deny the existence of knowledge by saying that the intuition of epistemologists are about the socio-linguistic reality of epistemology. What the socio-linguistic qualifier does is to establish a possible way for us to assess our intuitions akin to how physicists could assess observations (even if they don't have an *actual* way to do that). i.e My intuition that *p* would only be legitimate or valid just in case certain presuppositions that *p* depends on are true. For example, if I misunderstood the constituent term 'knowledge' in the proposition that 'the agent in the Gettier case has no knowledge of the target *p*', then I could end up with an intuition that fails to correctly capture the socio-linguistic reality of epistemology. In the trivial case, my misunderstanding of 'knowledge' could provide me with a different extension from that of the epistemological community. Then my resulting intuitions wouldn't be legitimate evidence in virtue of violating the linguistic rules of the community.

I hope that the reader is thus far convinced that intuitions are empirical in the same way that observations are. In the next section, I argue that intuitions are epistemically akin to observational data in the sense that they *constitute* the topics on which we theorise. i.e. My intuitions on what knowledge is constitute the data for the study of knowledge.

1.4 Intuitions as Topic and Tacit Theories in Science

Consider the question: 'What is knowledge?' *Prima facie*, the question asks for a description on this thing that is *ordinarily called 'knowledge',* or *considered to be knowledge*. When epistemologists provide such a description, they are providing a theory of knowledge in epistemology. In other words, a theory of knowledge is a theory of the thing that folks would *intuitively* consider to be knowledge. To use some examples from other domains: When chemists studied the chemical composition of water, they were studying entities that is *ordinarily called 'water'*. When mathematicians studied the geometrical properties of circle, they were studying the shape that is ordinarily called 'circle'. In general – other than stipulated entities – the theorising of some *X* must begin with identifying the stuff that we would intuitively call '*X'*. i.e. Intuitions play a topic-fixing, or constitutive role for theorising. In this section, I defend the following two conditions:

Ta. A theory of X in a domain D $T_D(X)$ is a D-specific description of the class of entities that the community of D would intuitively classify as X, and

Tb. If the community of D has the intuition that F(X), then $T_D(X)$ is a D-specific description of the class of entities in which at least F is satisfied.[49]

Condition Ta is straightforward, it claims that if I theorise about some X, then I am theorising about the stuff that I intuitively take to be X. It takes care of the intuitions that directly *identifies* X. For example, most epistemologists share the intuition that in a basic Gettier case, the agent in question has no propositional knowledge. Condition Tb takes care of the intuitions that *predicates* X. For example, most epistemologists also share the intuition that a piece of knowledge is a piece of true belief. So when epistemologists theorise about knowledge, they only consider entities that are also true beliefs. Most intuitions used in philosophical arguments are in the form of Tb – we usually have intuitions that some entity X in question entails some condition F. I illustrate Tb again with one of our initial examples: the intuition

[49] Note that this is *not* saying that all instantiations of $T_D(X)$ would satisfy F. Perhaps $T_D(X)$ revises the pre-theoretic X and in the process violates F for certain values of $T_D(X)$.

that torturing a sentient being for fun is wrong is an intuition that 'torturing a sentient being for fun' is a sufficient condition for *wrongdoing*. When ethicists theories about ethical actions, they must treat the action of *torturing sentient beings for fun* as a piece of data for unethical actions.

To defend the pair of conditions Ta and Tb, I argue that they neatly resolve the paradox of analysis. A version of it in epistemology has been presented as the problem of the criterion:[50]

'To know what *knowing* is, we must have a *procedure* for distinguishing genuine cases of knowing from non-knowing. But to know whether our procedure is a good procedure, we have to know whether it really *succeeds* in distinguishing genuine cases of knowing from those that are not. And we cannot know whether it does really succeed unless we already know which cases are genuine cases of knowing and which are not. *And so we are caught in a circle.*' The problem can be reconstructed as the following argument:

P1. If we are to theorise about some *X*, then we must be able to identify *X*.

P2. In order to identify *X*, we need to, at the minimum, be able to distinguish those which are *X* from those which aren't.

P3. However, if we can distinguish *X*s from non-*X*s, then we already have a set of criteria for what *X* is.

P4. If we have the criteria for what *X* is, then we have a theory of *X*.

C. A theory of *X* is the prerequisite for theorising about *X*, and so we are caught in a circle.

Prima facie, this argument seems reasonable. However, if we think about the actual activity of theorising, we soon realise that P3 and P4 are dubious. P3 presupposes that in order to distinguish *X*s from non-*X*s, explicit criteria are needed. This is false. Folks frequently use terms, the definitions of which they do not know. For example, a person can correctly assert the statement that 'I need more RAM in my computer' without knowing that 'RAM' is an acronym for 'random access memory',

[50] A related issue is the 'paradox of analysis'. However, I find Chisholm's version to be more illuminating.

or whatever that means. It is sufficient that the person can correctly *refer* with the term. In other words, what we need for correctly identifying an *X* is the ability to correctly identify the extension of *X*. Knowledge of *X*'s intension is unnecessary. Nonetheless, this does not mean that we lack the definition of *X as a community*. In most, if not all of the actual cases of theorising, we have established lexical entries of the terms that are under scrutiny in specialised fields. So even though P3 is false, it's still the case that we do *in fact* have explicit criteria for most *X*s, if we think of theorising as a public activity. After all, it's not the individual philosopher's conception of knowledge that is at issue, but the *community*'s conception of knowledge. So the question is now: does having explicit criteria for an *X* constitute a theory of *X*? P4 assumes that this is the case.

P4 presupposes that if we have knowledge of *X's* intension, then that is a theory of *X*. This is in one sense, trivial. However, this misses the point of theorising for a conjunction of two reasons. One, as a matter of fact, our pre-theoretic knowledge of an *X*'s intension are grounded on either the lexical entry for *X*, or from simply acquainting ourselves with how the community uses '*X*' to refer. Two, a theory is domain dependent. There is no theory of *X* without first specifying what *kind* of theory we want. One cannot *simply* theorise about water. One must specify whether one aims to study the *chemical* properties of water, or the *physical* properties of water, or the *gastronomical* properties of water. Therefore, having a lexical entry of 'knowledge' is insufficient for having an *epistemological* theory of knowledge. For example, when biologists studied the biological composition of fish, the dictionary definition would have been unhelpful. What the dictionary definition *did* do however, was to *fix the topic.*

One might point out that pre-theoretic intuitions about some *X* could differ in extension from the theories of *X*. For example, biologists revised the notion of 'fish' to exclude whales from its extension, although they began with the folk intuition that whales are fish. If so, didn't the biologists change the topic? There is a trivial sense in which this is the case, *if* we equate *topic* with *extension*. However, given the actual meaning of 'topic' and how scientists use the term, there is no reason to assume this.

The topic of ‘chemistry’ has evolved to exclude alchemy from its extension. The topic of ‘Italian cuisine’ has evolved over time to include large quantities of tomato based dishes after they were introduced into Europe from the Americas. The topic of ‘reasoning’ has evolved to include probabilistic reasoning. In non of these cases can we say we no longer were engaged in chemistry, or Italian cuisine, or reasoning when their extensions changed. Similarly, biologists never changed their topic of study when they excluded whales from the classification of ‘fish’. What happened was that they had decided to *revise* the notion of ‘fish’ to exclude whales for scientific/ pragmatic reasons such as theoretical cogency. For example: fish are (mostly) cold blooded while whales, like mammals, are warm blooded; fish use gills to breathe under water while whales, like mammals, must breathe in air; fish swim by moving their tales side to side while marine mammals, such as whales, swim by moving their tales up and down. If biologists were to classify whales as ‘fish’, then these properties can no longer be consistently classified as being part of ‘fish’, since they would then vary among the members. However, these properties are biologically salient and were shared among a significant population of animals called ‘fish’, such that it would have been more efficient to reclassify whales as not fish, but mammals. Furthermore, it is precisely because biologists cared about the topic of *fish* that they wanted a robust *and* consistent theory of ‘fish’.

To conclude, P4 of the criterion argument is false, since a set of criteria for some *X* does not constitute a theory of *X*. Nonetheless, they contribute toward our ability to pick out *X*s from non-*X*s. Thus, these criteria fix the topic for us in the following sense: when we wish to study an example of *X*, we look at the entities that satisfy the pre-theoretic criteria we have for *X*. In philosophical arguments, these pre-theoretic criteria are what we call ‘intuitions’. This naturally raises another burning worry: whether these intuitions are ‘tainted’ with expert opinions and whether this is desirable.

1.5 Theory-Free vs. Theory-Laden Intuitions

We like to think that science operates on *objective* evidence. However, are two major challenges to this assumption. First, quantum mechanics has experimentally demonstrated that observational evidence can be observer-dependent even without being theory-laden.[51] Second, philosophers of science have highlighted two ways for which observational evidence can be theory-laden.[52] The evidence can either epistemically *depend on* or be *influenced by* theoretical claims. In contrast, a theory-free evidence is one that is neither dependent on nor influenced by theoretical claims.

Epistemic dependence can be characterised as: a doxastic attitude B of a proposition p depends on a distinct proposition q iff $B(p)$ is valid only if q is true.[53]

[51] Proietti et. al (2019) has experimentally realised Wigner's thought experiment from Wigner (1961) where two observers experience irreconcilable realities at the quantum level. Granted, observer-dependent evidence need not entail observer-dependent quantum theories, since either locality or statistical independence could be violated by quantum mechanics, as demonstrated in Hensen et. al (2015). However, the point stands that we cannot in general assume observer-independency for observational evidence. Moreover, the point of the experiment wasn't that observations are observer-dependent, this is trivial (via being theory-laden). The point was that observer-dependency could be a feature of the measured reality rather than human quirks (including our beliefs etc.). This reinforces the QBist reading of quantum mechanics (and Proietti et al. support this view) where the very target of a quantum theory *just is* the reality *as experienced by an observer*. However, it does not entail anti-realism. In fact, the experiment demonstrates the opposite: the reality that exists as we know it may be necessarily observer dependent as a matter of physical reality. Since intuitions are trivially observer-dependent in virtue of having no obvious external reference, we need not delve further into this point.

[52] Of course, the 'theory' in 'theory-laden' is a misnomer. The concern about theory-laden evidence is whether an observation can be observer-independent. So in practice, 'theory' here simply refers to anything that potentially undermines the observer-independence of an observation.

[53] In the traditional epistemological literature: e.g. Audi (1983), Hardwig (1985), Oakley (2006), 'epistemic dependence' is often characterised as a relation between two doxastic states, with an emphasis on being justified. i.e. My belief that *p being justified* depends on my belief that *q being justified*. However, this kind of characterisation is too strong for dependence and insufficient for influence in the context of theory-ladenness. First, it's odd to talk of observational evidence (or intuition) as being *justified*. The observer do *not* typically hold doxastic attitudes about the presuppositions on which her observations depend. The dependence is between her observation and the presuppositions. Second, it's insufficient for characterising influence since one's observation can be influenced not just in a directly epistemic manner such as via believing a theory (in which case we can also say that the observation is *dependent* on the theory), but also by, say, a pragmatic manner via being a practitioner of a particular field (a physicist would simply see or intuit different propositions from that of a layman even if presented with the same sensory input). The standard characterisation fails to capture this. There's a related literature in social epistemology on 'epistemic dependence, e.g. Pritchard (2015) de Ridder (2014). Those works examine the phenomenon of interpersonal epistemic dependence, rather than a propositional epistemic dependence. Given that intuitions are community-dependent, the epistemic dependence can also be of an interpersonal nature. Being 'theory-laden' could mean being dependent on a community's practices – i.e. dependent on a linguistic framework. In any case, it's sufficient to keep in mind these features. We're not here to study what epistemic dependence is.

We say that an observational claim *o* is theory-laden if the *observation of o being valid* epistemically depends on some set *E* of background theoretical claims, whether *E* is explicitly held. Note that this does not mean that *o* depends on *E*, nor does it meant that *observing o* depends on *E*.[54] For example, if I look at the moon through a telescope, and record that 'the moon has brightens *b*', then *E* would include, say, the claim that the telescope's lens have refracted the moonlight in an expected manner under certain conditions. I could be completely ignorant of whether the telescope is working, but my observational evidence would still be undermined by the falsity of *E*. This holds true even for ordinary observations. For example, my observation that *the garden roses are red* presupposes that my vision is indeed working correctly, that I am not hallucinating, that I am not a brain-in-a-vat etc. Trivially, all observational evidence are theory-laden by dependence.[55]

Theory-laden by influence is less direct. While epistemic dependence relates the validity of a person's doxastic state to a proposition, epistemic influence holds between a person simply assenting to a proposition and that person's beliefs, dispositions, preferences etc.[56] We can characterise influence roughly as: an agent *Q*'s assenting to *p* is *influenced* by some relation between *Q* and a specialised domain D: *R*(*Q*, D) iff *R*(*Q*, D) would raise the likelihood that *Q* assents to *p*, ceteris paribus. *R* here could denote believing a proposition in D, that *Q* studies D, that *Q* espouses the priorities of D etc. In this second way, a set of observational claims *O* is theory-laden in case the person who assents to *O* with a certain degree of confidence would otherwise have been less likely to assent to *O* if she were to have a different set of

[54] After all, I could observe *o* even if not-*E*.

[55] Azzouni (2004); Chang (2005).

[56] In the case of the relation holding between doxastic attitudes, one can also say that this is a kind of epistemic dependence holding between the attitudes. i.e. My belief that *p* depends on my belief that *q*. But of course, influence isn't limited in this way. Note a potential ambiguity: while dependence is

background beliefs or dispositions or preferences etc.[57] For example, only a person who's had training in music theory would observe that a pianist is playing a certain harmonic progression. Furthermore, the trained person would *inevitably* see the harmonic progression. One could say that in this case the observation that 'the pianist is playing a certain harmonic progression' is *dependent* on the person's beliefs that certain combination of notes constitutes a certain chord etc. However, influence needn't be restricted to between doxastic attitudes. A person simply *being* a musician of a certain field (say, one who specialises in post avant-garde music) would influence that person's perception of certain sounds. This is irrespective of whether the person assents to any particular theory.

Our concern with theory-laden evidence is strictly with those that are laden via influence. We *want* the presuppositions on which our observations depend to be true. I want my telescope to work as intended. I want that I am not hallucinating when I make the observation that the roses are red. The worry about theory-laden by dependence is that the presuppositions on which my observations depend could fail, and that it's possible that at least one of these presuppositions is undecidable. This problem however could be mitigated via replication. On the other hand, we might see theory-laden via influence as something that undermines the reliability of our observations. After all, my measurements of physical reality *shouldn't* be influenced by my beliefs, dispositions, biases, desires etc. This essay does not deal with the question of whether being theory-laden via influence undermines the objectivity, or at least the reliability of observations. The point is to show how observations and intuitions could be theory-laden in an epistemically parallel fashion.

[57] Kuhn (1962) further distinguished between three types of theory-ladenness via influence: perceptual theory loading; semantical theory loading; and salience. In practice, our observations are almost always theory-laden via a mix of different of different *R* between ourselves and a specialised domain, and also a mix of distinct kinds of theory loading and salience. For example, the fact that I espouses the values of a particular branch of physics would affect how salience certain parameters are for me, and being a physicist would mean that I would be semantically loaded to describe an observation in a certain manner, and believing certain theories would entail a perceptual loading on what I might see as an anomaly for an observation or even a skewing of parameters. So it's not of interest for our purposes to discuss the details of how perceptual theory loading might differ from semantic theory loading etc. We need only a general 'influence' relation here for examining whether the intuitions we use are also influenced by our specialised domains in the way described here.

We already know that intuitions are community dependent. I argue that intuitions in philosophical arguments are theory-laden in the same way that scientific observations are. The ideal of a theory-free folk intuition is a red herring. Although there are contentions in analytic philosophy, the default position is to take intuitions in arguments to be theory-free, and see it as a virtue for using intuitions as evidence. Here is an example, where Kripke argues that the notion of a necessary or contingent property is meaningful by appealing to folk intuition: 'If someone thinks that the notion of a necessary or contingent property (… [and consider] just the *meaningfulness* of the notion) is a philosopher's notion with no intuitive content, he is wrong.'[58] In this argument, Kripke implicitly presupposes that an 'intuitive content' is *pre*-philosophical, or more generally, pre-theoretic/theory-free. He then goes on to say that theory-free intuitions are in some sense the most conclusive evidence one can have: 'Of course, some philosophers think that something's having intuitive content is very inconclusive evidence in favour of it. I think it is very heavy evidence in favour of anything, myself. I really don't know, in a way, what more conclusive evidence one can have about anything, ultimately speaking.' I think such attitudes about intuitions are deeply misguided. First, I argue that the notion of 'folk intuition' cannot be salvaged, and if we're to think of a distinction of some sort, we must talk of theory-free vs. theory-laden intuitions. Then, I demonstrate how the archetypical intuitions we use in philosophy are theory-laden just in the way that observational evidence are.

1.5.1 *The Problematic Notion of 'Folk Intuitions'*

Philosophers more often use the phrases 'folk intuition' and 'expert intuition', over 'theory-free intuition' and 'theory-laden intuition'.[59] Perhaps this is to avoid confusing it with the issue of theory-laden observational evidence in philosophy of science. However, the term 'folk intuition' can be problematically ambiguous. In the context of a philosophical argument, 'a folk intuition that *p*' can mean: 'the intuition

[58] Kripke (1980)

[59] Manley et al. (2013)

that *p* held by non-philosophers'; 'the intuition that *p* held by non-academics'; 'the intuition that *p* held by everyone when they're not in the context of a philosophical discussion'; 'the intuition that *p* held by philosophers when they're not in the context of a philosophical discussion'; 'the intuition that *p* held by anglophones'. The list isn't exhaustive, but it exhibits plausible interpretations. The ambiguity would be benign if the intuitions that philosophers consider to be 'folk' are shared across these 'folk' communities. However, as we know from even localised surveys on intuitions that are relatively anodyne in the context of their debates, this assumption cannot be true. For example, among non-philosophers, people from different social-economic backgrounds or different cultures can form different judgements on the Gettier cases.[60] More generally – as aforementioned on the same empirical ground – intuitions are domain/community dependent. Each of the above disambiguation of 'folk intuition' delineates a distinct community. Moreover, for almost every piece of intuition, there is some group of *non-experts* who do not share the intuition.[61] In the trivial case it would be simply because the group lacks a certain concept that is required for the target intuition.[62] For illustration, let's consider the intuition that water quenches thirst. For someone who has adipsia from birth, the intuition would be quite meaningless. You might think that my example is absurd.[63] Granted, the

[60] Weinberg et al. (2001). Furthermore, in private conversations with my mathematician colleagues, I have found that they also do not share the epistemic intuitions for which philosophers take the Getter arguments seriously. e.g. Many of them fail to see why the belief in the Gettiered case is justified. I myself share their intuition more so than I do with fellow epistemologists.

[61] This case differs from Steuer (1985), where two scientists conducting the same experiment under the same conditions can observe different results. However, the Steuer example exists for intuitions, since there are idiosyncrasies within a community, though it is not relevant to our discussion here.

[62] There could be various reasons for why an individual might fail to assent to an intuition that's shared by her community. Williamson gave an example for analytic propositions in Williamson (2007).

[63] If you dislike my example, think of deontological and political disagreements between different cultures, and even between groups of the same culture. Many of these disagreements are at the level of intuitions. For example, people can disagree about what fairness is in the context of taxation policies. Our priorities or psychological quirks such as confirmation bias aside, the very fact that we disagree at the level of intuitions – the lack of certain common grounds – is why it is very difficult to resolve political conflicts even for experts. One can also think of basic axiological or normative differences. These could lead to apparently incommensurable ethical differences. Some might see this as an argument for moral relativism, but that would hold only if one presupposes that intuitions cannot be revised or measured against each other. This is in fact the crux of this section, where I argue that being theory-laden is a good thing.

intuition that water quenches thirst isn't even *meaningful* to those with adipsia, since these folks have hampered experience of thirst. However, we must ask ourselves: to which communities should we apply our even more domain-dependent 'folk' intuitions, such as those regarding what knowledge is? i.e. Is knowledge a valid concept only for the upperclass western folk like thirst is for folks without adipsia?[64] The answer is a clear 'no', because unlike 'thirst' for those with adipsia, there is no clear reason to assume that certain communities have a diminished understanding of knowledge. This kind of segregation would be self-defeating for the very ideal of 'folk intuition' – to uphold the generality, or universality of these intuitions. For the more refined intuitions that are commonplace in philosophical arguments, which are about specific properties that are rarely considered outside of philosophical discussion, there is a high risk that such intuitions are idiosyncratic when examined *across* communities. Even the British and the American public, who both speak English, can have saliently different intuitions on simple matters such as what could be considered a 'forest'.[65] Or intuitions about 'time' between those who are familiar with general relativity and those who aren't.[66] And in the case of probability theory or the philosophy of probability, I'm certain even the most diehard ordinary language philosophers would shy away from permitting any kind of 'folk' intuition to guide their theorising.[67]

[64] Weinberg et al. (2001) found that certain Gettier intuitions are not shared by those from a lower social-economic class or folks with an east Asian cultural background. Although Machery et al. (2015) found no cross-cultural difference with respect to the Gettier Car intuition for four particular linguistic communities, the ball is on them to show that the Gettier intuitions are *in general*, shared across cultures. This is unlikely, however, because even within the philosophical community, certain kinds of Gettier intuitions are contentious. Bonjour (1985)'s True Temp case is an example of such a contention *among philosophers*.

[65] Britons distinguish woods from forests by area, whereas Americans distinguish the two by density of canopies. More generally, Americans would consider smaller areas of wooded lands to be forests in comparison to their British counterparts. https://sciencing.com/animals-forest-ecosystems-8056265.html; https://www.woodlandtrust.org.uk/blog/2018/03/difference-between-wood-and-forest/

[66] The former group might take for granted the relativistic nature of time, and see it as in relation to other physical entities such as mass, velocity etc. Whereas the latter group would find the idea of a relativistic time highly dubious. Resultant intuitions about time can thereby differ between the two groups, even if neither are 'experts' in any relevant way.

[67] Because non-expert intuitions about probability are notoriously unreliable.

You might think that I'm exaggerating. Sure, communities could diverge on specific intuitions. However, this doesn't make 'folk intuitions' an invalid standard for philosophical theorising. For instance, we could say what philosophers *aim* to do (although we often fail), is to use the intuitions that are indeed shard by *all* folk communities *for whom the intuition is meaningful*. For example, the fact that certain epistemically competent communities disagree with some of the Gettier intuitions is evidence that those intuitions shouldn't be used as evidence for epistemological theorising. However, the cost of this route is almost worse than losing generality. This very strategy assumes that somehow, perhaps magically so, the intersection of all folk intuitions on a certain topic is the correct set of evidence for that topic. There is absolutely no reason to assume this. The fact that certain epistemically competent communities might lack a certain epistemic intuition is no evidence against that intuition. This lack could be due to many accidental factors such as priorities, linguistic practice etc. For example, some languages have grammatical evidentiality, which means that a speaker *must* indicate evidential-type information such as 'I remember that', 'it seems that' etc. as a grammatical requirement. In European languages, evidential-type information is optional. Speakers of languages with grammatical evidentiality might exclude much of what we consider to be legitimate testimonial justification from their cognate of 'justification', simply because in their language the speaker would always convey where the information comes from, which would have an effect the listener's credence.[68] They might also have the intuition that epistemic justification always require an internalistic component, meaning that a person must be aware of the reasons for making a certain claim. Most anglophones aren't native speakers of English, and a significant minority could come from a linguistic background with evidentiality, which would affect their epistemic intuitions. Moreover, lack of certain intuitions about some X could simply indicate that the community in question has a defective understanding of X, and we have no way to determine whether this is the case upfront, unlike with adipsia. To make matters more

[68] And in case of a survey in the style of Weinberg (2001), their answers could display an interesting pattern that differs from the typical anglophone or European participant.

complicated, the size and number of communities that hold a particular intuition is no evidence for the evidential strength of that piece of intuition. For example, most communities in the world thought of whales as a kind of fish. Still, these communities had a faulty intuition, and this became only apparent after the act of scientific theorising. The same goes for most intuitions in philosophy – we cannot *pre-theoretically* determine who has a faulty intuition. In general, it's unlikely, and pre-theoretically undecidable that the intersection of all folk intuitions of X is a robust categorisation of X. Therefore, we absolutely should *not* consider only intuitions that are shared by *all* folk communities.

Then the question remains: If we want to hang onto the notion of a 'folk intuition', we have to decide *which* 'folk intuitions' are stronger and which are weaker evidence. Obviously, the flip side solution to the above – taking the union of intuitions from all folk communities also cannot work, unless one is ready to uphold some kind of paraconsistent framework for theorising. Here is one last hope for the ideal of 'folk intuitions'. Let's distinguish the intuitions of X from the intuitions of *whether something is an X*. For example, the former would be, the intuition that justification is fallible, while the latter would be the intuition that the agent in the Gettier case has a justified belief (this is an *exemplification of* justification). We might want to say that what we really mean by 'folk intuition' is the latter kind of intuitions. Surely these can be more universally shared by different folk communities, *even if they disagree on intuitions of the former kind*. In other words, folks can disagree on whether X is F while agreeing on which things are X. However, this won't help us. There is no evidence that different communities agree on even this restricted kind of intuitions. Again, consider the silly example of what counts as a 'forest'.

Therefore, the distinction cannot be about whether a piece of intuition is held by 'the folk', since there is a multiplicity of folk communities, with heterogenous intuitions and no way of ranking the credibility of conflicting or even minority folk intuitions. Rather, it's about whether a piece of intuition is theory-free. If we are modest, then we must admit that what philosophers *actually* mean (as opposed to what we think we mean, or what we would like ourselves to mean) by 'folk intuition' is often just intuitions that anglophone philosophers typically share, but appear theory-free.[69] And in light of the increasing mutual alienation between sub-disciplines, we should really take ‘folk intuitions’ to mean *theory-free intuitions of experts from a specific debate*. Therefore the alleged distinction is really between ‘theory-free’ and ‘theory-laden’, or ‘pre-theoretic’ and ‘post-theoretic’ rather than ‘folk’ and ‘expert’ intuitions. If we use the notion of ‘theory-laden’ from the context of talking about observational evidence, then a theory-free intuition is one that doesn’t depend on any theoretical claims. To exactly understand what a ‘theory-free intuition’ is, we must decide on 1. what this ‘dependence’ relation is for theory-laden intuitions and 2. what counts as ‘theoretical claims’ in the context of philosophical arguments.

1.5.2 *How are Intuitions Theory-laden?*

Recall that observational evidence could fail to be objective in the sense of being observer-independent by either being *simply* observer-dependent (without being theory-laden at the quantum level), or by being theory-laden. This subsection aims to demonstrate that intuitions are likewise observer-dependent by being theory-laden.

I argued that intuitions are community dependent in virtue of a person upholding certain values, beliefs, dispositions, categories etc. Crucially, both observations and intuitions can be evaluated against the respective realities that they measure. So the validity of both observations and intuitions are dependent on certain domain specific presuppositions. In this regard, both are trivially theory-laden by epistemic dependence.[70] As we have noted, this dependency does not undermine the reliability

[69] Here I mean by ‘anglophone philosopher’ philosophers who work in the tradition of anglophone analytic philosophy. Obviously, ‘intuition’ is a different topic in the domain of German Idealism or so-called ‘continental philosophy’.

[70] In fact, it’s easier to see why this is innocuous in the case of intuitions. In 1.3.2 I have pointed out how community-dependence is in effect a restriction of an intuition by the rules, axions, priorities of a domain. We can in fact say the same for observations. i.e. The presuppositions on which our observations depend are in fact also domain-specific. For example, in physics the presuppositions would relate to the correct measurement of physical quantities such as brightness, length, mass etc.

of an observational statement, since we usually can mitigate the risk via replication.[71] Now we ask: Is there a similar process in philosophy that mitigates the risk of idiosyncratic intuitions formed via a failure of adhering to a domain? First, I show that the dependency in the intuition case is indeed similarly desired to the observation case. Then, I show how a process of replication is similarly possible and in fact practiced in philosophy.

Consider again the intuition that 'torturing a sentient being is wrong'. We noted that the intuition *depends* on one's correct understanding of its constituent terms, and on holding onto certain values, categories, dispositions etc. that are shared within a certain community. i.e. The agent must have certain beliefs, values, categories, understanding of English etc. to have the intuition. Of course, these dependencies do not undermine the reliability of the intuition. Rather, they simply *become manifest* in the explicitly stated intuition. They demarcate the domain in which the intuition can be used as legitimate evidence. There should be nothing surprising about this. What applies at the macro level for an observational statement can easily fail at the quantum level. A correct observation is a product of a certain set of presuppositions and the physical reality; a correct intuition is the product of a certain set of presupposition and the social-linguistic reality.

So what happens when two epistemologists disagree with regard to a piece of epistemological intuition? This could be either because they are differently *influenced* by their background beliefs, dispositions, values etc. *or* because one of them has failed to assent to a certain core component of the domain of epistemology. Usually, the disagreements are due to distinct influences. For example, different intuitions regarding whether the agent in a particular Gettier though experiment has knowledge are likely due to distinct influences pulling the epistemologists in ever slightly opposite directions. In surveys that demonstrate how different socio-economic

[71] This is why replicability is so important for science, and not just the traditional empirical sciences. Lately, worries about research in AI and machine learning have centred on the lack of replicability due to the sheer financial inaccessibility of many results that have been produced via powerful machines owned by rich private companies like Google. I'll promptly show how a similar process is and has been equally important in philosophy (although not yet labeled as 'replication').

communities would judge whether the agent in the Gettier case has knowledge, the difference was certainly due to different background influences rather than one party failing to, say, understand the English language (in that case, the relevant community was *anglophones* rather than *epistemologists*).[72] We will come to this issue later when we examine how intuitions might be influenced. But for now, we need to ask ourselves how philosophers might mitigate the risk of specialists in a domain failing to adhere to the linguistic framework of the domain. So suppose that an epistemologist insists that the agent in the Gettier case *has* knowledge, not because she holds particular theoretical beliefs or values, but because she has failed to see, say, that the agent in the thought experiment *could have believed a falsehood*. The effect of such a failure is in fact much more easily mitigated in philosophy than in the sciences, because philosophers need no expensive and elaborate experimental setups to replicate an observation. Instead, we simply reflect on the thought experiment and cross-examine our intuitions with other epistemologists. This is what Frank Jackson meant when he said that each one of us is doing our bit of field experiment when we teach our students about the Gettier case.[73]

The rest of the section will try to show how intuitions could be *influenced* by one's commitment to a domain. Consider a version of Bonjour's True Temp thought experiment:

> Broken Thermometer: Lily forms beliefs about the temperature in her room by reading the thermometer on the wall. Unbeknownst to Lily, the thermometer is broken and fluctuating randomly within a certain range, while

[72] This also shows why a linguistic corpus study of English terms cannot help philosophers to theorise about the terms that are used in ordinary English. i.e The surveys on people's differing intuitions about knowledge is irrelevant for epistemology. It was however extremely insightful for how intuitions are community and domain dependent.

[73] See Jackson (1998). What he didn't mention is that the students in this case also should operate within the domain of epistemology (or in popular parlance 'the philosophy classroom'). Students or non-philosophers, including academics, who struggle to obtain the intuition are struggling due to a dislodging from the framework of epistemology. For example, when I first mentioned the Gettier case to a mathematician friend of mine, he struggled to understand the thought experiment simply because he refused to leave the domain of mathematics and understand what epistemologists were on about. He was still thinking like a mathematician.

there is someone in the room next door who would adjust the reading to the actual room temperature were Lily to consult it.

Now consider the intuition that (NLKT) 'Lily doesn't know the temperature in her room when she consults the thermometer'. NLKT is influenced by (though clearly independent of) the widely accepted Ability Constraint on knowledge (AC) in epistemology: an agent *S* knows that *p* only if *S*'s getting to the truth about whether or not *p* results from the exercise of *S*'s cognitive ability (or abilities).[74] However, it's unclear whether AC entails NLKT.[75] Therefore, we cannot say that NLKT is a conclusion in epistemology that is justified by AC. The *negation* of NLT (LKT) could be readily accepted if one accepts Hudson's argument while rejecting his 'true description' requirement.[76] Moreover, if we simply step outside of the domain of epistemology, it's not difficult to intuit that LKT. After all, the agent in the True Temp case has a much stronger case for having knowledge than other kinds of Gettier cases. So if folks outside of epistemology are ready to accept knowledge for standard Gettier cases, they would also be ready to accept LKT. So we can say that NLKT is a piece of theory-laden intuition in the domain of epistemology by influence.

Not all intuitions are influenced via having certain beliefs. As per the discussion on observational evidence, our measurements could be influenced in a variety of ways. It's easy to see how any expert from a non-philosophical domain, especially data science, would readily intuit that LKT. After all, Lily's belief isn't lucky, and is in

[74] Pritchard (2009) defended NLKT by appearing to a wrong 'direction of fit' between the thermometer and the environment. Pritchard (2010) (2012) use the True Temp case to illustrate that satisfying the Anti-Luck Constraint does not mean satisfying also the Ability Constraint. For Kelp (2013), the intuition that the agent has no knowledge is a given. Hudson (2014) argues that Pritchard should attribute knowledge to the agent given his anti-luck views. i.e. Pritchard intuits that Temp has no knowledge, Pritchard also likes his anti-luck epistemology, and so Pritchard has used his theory to provide a rationalisation of his intuition. On the other hand, Kelp and Hudson do not share the intuition that Temp clearly has no knowledge, and argues instead that Pritchard's anti-luck epistemology in fact supports the view that Temp has knowledge.

[75] Kelp (2013) and Hudson (2014) argue that Pritchard should in fact attribute knowledge to the agent given his anti-luck virtue epistemology. In the case of Kelp (2013), the intuition that Pritchard had about the agent was a given. In the case of Hudson (2014), this isn't so clear, as he seemingly leaves both options open.

[76] Hudson (2014) suggests that if Pritchard wants to insist that the agent in the True Temp case lacks knowledge, then he should accept what he calls the 'true description' requirement.

ordinary parlance, necessary.[77] For a data scientist who practices machine learning, what's important is that the measurement is reliable. Whether the measurement is true *because* of the measurement fitting the environment or the other way around is simply an irrelevant metaphysical concern for the data scientist. In this case, the influence is by preferences, values, or dispositions.[78] Moreover, in the real world, the measurement can often affect the environment. Examples include 'self-fulfilling prophecies', games such as chess, or William James's example that believing that 'the French Revolution would be successful' is necessary for its fulfilment.

1.6 Intuitions as Data with Statistical Significance

To summarise, I have argued that intuitions are evidentially basic but empirical, domain and community dependent, topic fixing, and theory-laden in an analogous manner to observations. Therefore, we can say that intuitions are defeasible and statistically laden. Now we have the necessary components to establish how observations and intuitions are evidentially analogous – in so far that intuitions are also to be used as domain-specific statistical evidence. The result is a shift in the structure of philosophical arguments: from using intuitions as premises of deductive arguments, to using intuitions as data with statistical significance.

For example, every piece of intuition about knowledge is a datapoint for our theory of knowledge. We ought to measure our models of knowledge against our intuitions about knowledge in a statistical manner. i.e. We should not jump to modifying a model at the first sight of a potential counterexample. Rather, we should examine and weigh every datapoint in a statistically rational manner. Assuming that the data is indeed verified to be valid within epistemology, the potential counterexample could also be rejected in favour of a combination of simplicity and overall better fit with the entire set of valid data on knowledge. Williamson's complaint that our theories of knowledge is becoming increasingly complex and

[77] Of course we don't mean this in the metaphysical or logical sense.

[78] Depending on how you would construe it.

untraceable could be dissolved with this new attitude towards intuitions.[79] In general, our models should aim to fit our intuitions in a statistically satisfactory manner. For example, perhaps the best theory on knowledge is one that would not perfectly fit over any of our intuitions. However, the theory neither overfits nor underfits, while maintaining a robust selection of parameters that allows us to fruitfully generalise the theory to new cases.

In preparation for the upcoming discussion in Chapter 2, I present a parametrised representation of intuitions. Consider some topic X in a domain D, let's call the set of all intuitions of X in D $I_D(X)$, and the set of all instances of X that satisfy $I_D(X)$ X_D. Suppose there are *n* number of D-specific features/parameters for X, let's represent these parameters as a vector $F_D = \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_n \end{pmatrix}$. Then for some datapoint x_{Di} belonging to

X_D, $x_{Di} = \begin{pmatrix} w_{i1} \\ w_{i2} \\ \cdots \\ w_{in} \end{pmatrix} \cdot F$ where *w* represents weights for each of the *n* parameters for x_{Di}.

i.e. Every instance of X is a weighed combination of the *n* number of *D*-specific parameters.

It is always easier to point at something than to articulate it. Theorising is the business of articulating our reasons for pointing at something. Intuitions and observations are our initial recording of the thing we point at, and theorising is the process of extracting useful information from that recording. In this respect it should be no surprise that we cannot do without observations or intuitions lest we theorise of empty fantasies, although we must caution against giving them absolute authority lest we suffer not only circularity, but irrelevance.

[79] Williamson (2000)

Chapter 2 On Revision in Descriptive Projects

'Generality is, indeed, an indispensable ingredient of reality; for mere individual existence or actuality without any regularity whatever is a nullity. Chaos is pure nothing.'[80]

2.1 Introduction

To answer the question 'What is water?', we may answer by either pointing to bodies of water, or by providing a description thereof. The former results in knowledge by acquaintance, the latter knowledge by description.[81] Much of scientific projects are about providing domain specific descriptions of entities that we already know via acquaintance. For example, chemists have provided a chemical description of 'water'; physicists have provided a physical description of 'star'; neuroscientists have provided a neural description of 'sadness'. Moreover, these domain specific descriptions are meant to generalise beyond observed data. So scientific theorising is also the activity of providing a general description from a finite set of data.

Given some X, I style questions of the form 'what is X?' as '*ontological questions*'. Philosophers take sides on whether X should be a term, a concept, a thing.[82] However, methodological concerns on how to theorise about X are indifferent to what the nature of X is. I stay with language because it is tractable and malleable – terms are what we see and work with when we define, explicate, analyse, and what have you. If my readers wish so, they may freely think in terms of concepts or

[80] Peirce (1905)

[81] The terminology is introduced in Russell (1910)

[82] Williamson (2007) claims that philosophy had undergone a conceptual turn from a linguistic turn in the mid twentieth century, and he claims that we should see philosophy as studying things, not concepts.

things.[83] Even in the case of a newly discovered entity[84], we must first name the entity some 'X' before we can ask what X is. If X does not have a description, then we must begin with the extension of 'X', via examining the cases in which we would consider 'X' to be exemplified. We could study a typical sample of the entities that 'X' denotes and delineate these entities from non-X via constructing a theory.[85] This theory would be a description of X. Using the terminology from Chapter 1, we can say that a D-specific description of some X is a theory of X in the domain D: $T_D(X)$. We can say that the goal of an epistemic enquiry about X in a scientific domain D is to produce $T_D(X)$.

Sometimes, existing descriptions are insufficient for our purposes and either new terms or revisions are required. This happens in both specialised and ordinary discourse. Scientists have revised the term 'fish' to exclude whales. Chefs have revised the term 'salad' to include fruit concoctions. When a revision occurs, people disagree on whether the revised term changes the topic. Those against revision are known as 'descriptivists' in the literature endorsing revision.[86] Descriptivists insist

[83] See Cappelen (2018) for a slightly more extended discussion of how conceptual engineering and conceptual analysis are not about *concepts* (in particular), but 'representational devices' – a general term for concepts, terms, things etc.

[84] By 'entity' I mean anything that can be named. This includes things, states, concepts, language etc.

[85] In some domains, we might draw a distinction between models and theories (and in some cases frameworks), and the distinctions can vary depending on the domain. In physics, a model would be something like a mathematical model of a planetary orbit around the sun; a theory would be something like the universal law of gravitation; while a framework would be something like Newtonian mechanics. Although even this is disputed. For example, while the standard view is that Newtonian mechanics is a theory (Frigg 2009), some physicists like David Grossman would count it as a framework like I have done above. In the computational sciences, one might treat 'model' and 'theory' as interchangeable (Diallo et. al. 2013). Within the scope of our discussions, we can safely use the terms interchangeably for most purposes. There are two reasons for this. One, philosophers have already been using 'theory' in a rather broad fashion to mean also *hypothesis* and *model*. Two, I see theorising as fundamentally a statistical activity of curve fitting. Although, on top of that, we still must abstract and translate from the mathematical models back into our domain-specific terms such as 'justification' in epistemology, or 'fairness' in ethics. One might rightfully see this as the distinction between a model and a theory in machine philosophy. However, this needn't concern us for now since having the distinction is orthogonal to the argument in this essay. Furthermore, it isn't clear whether the term 'theory' mean the same thing across different domains, and the distinction between a theory and a model isn't even all that clear or helpful within the sciences. To keep everything in order, let's stick to my general treatment of a theory of X as being a domain-specific description of X. In this regard, a 'theory' could be a model, a hypothesis, or a theory as shown above in my example from physics, demarcated by an adherence to statistical and domain-specific principles.

[86] Cappelen (2018) provides a up-to-date review of the two camps. See Strawson (1959) for an exegesis of the descriptivist position.

that because a revised term is not identical with the original term, it cannot answer the original ontological question, and by extension, questions regarding the original term. The most notorious example of such an argument would be P.F. Strawson's attack on Carnapian explication.[87] Carnap's proposal was to theorise about existing concepts via following the criteria of exactness, fruitfulness, similarity, and simplicity.[88] This was a revisionist proposal that aimed to *improve* existing concepts and make them more exact and fruitful, without changing the topic from the original concept. However, for Strawson, modifying or improving a target concept changes the topic, and therefore the result cannot answer the original question. Carnap replied by comparing the process of explicating to that of replacing a crude pocket knife with a more robust and perhaps specialised tool such as a machete. Carnap argues that once we have the machete, focusing on the pocket knife or its functions would be 'missing the point'. For Carnap, language is a tool, not a truth to be discovered.[89]

This exchange marked a conflict between the descriptive attitude and the pragmatic attitude regarding language, and by extension our concepts and categorisations of entities. The descriptivist wishes to leave the target of investigation as it is, and to be an observer who tries to accurately portray how things are. For the pragmatist, there is no 'final truth' to be discovered. Knowledge is a 'means of control'[90], a product of both reality and human effort.

Prima facie, the descriptive attitude seems sensible for answering ontological questions. After all, when we ask what knowledge is, we are asking about what knowledge is *as it has been conceived in this world*. Therefore, our theory of knowledge should capture accurately the extension of 'knowledge', which contains only entities that competence speakers of English would deem to be knowledge. But how we *represent* the world is up to us. Our terms and our corresponding conceptions of entities are human artefacts. Moreover, the meaning of a term changes over time. It

[87] Strawson (1963), with a reply from Carnap (1963).

[88] Carnap (1950)

[89] Carnap (1937)'s logical tolerance is an example of this attitude.

[90] Wheeler (2016)

would therefore be audacious to assume that our terms by default provide the best way to delineate the world, notwithstanding that what is 'best' would depend on the purpose of enquiry. Taxonomy isn't trivial. In philosophy, the primary source of evidence is intuitions. However, as I have argued in Chapter 1, intuitions are themselves domain dependent and theory-laden. The treatment of intuitions as a priori is a pitfall for philosophical theorising. The main response to this is the recent rise of functionalist epistemology and the abandonment of post-Gettier epistemology, which is an example of the pragmatic attitude replacing the descriptive attitude.[91] For an epistemic functionalist, what matters is the role we want knowledge to play, not what we have conceived knowledge to be. The question regarding topic change is not whether our pragmatically driven characterisation has changed the topic, but whether there was any clear and meaningful topic to begin with.

Following the Carnap-Strawson exchange, the literature has been treating descriptive projects and revisionary projects as going separate ways, or even mutually unintelligible.[92] The arguments appear to be at a higher-order level concerning whether philosophy should be descriptive. However, this is a mistake. Projects in the natural sciences are archetypically descriptive projects. Science ought to describe the world as it is, scientific theories strive to be true. However, more often than not – if not always – scientists engage in revision. In particular, scientists revise pre-theoretic terms. Sometimes, terms such as 'fish', 'water', 'star' as is used in scientific discourse might be treated as constructions rather than *re*constructions, under the banner of 'technical terms' as though they are independent of their ordinary language ancestors. While this is permissible for classificatory purposes, it is a horribly misleading

[91] 'Post-Gettier epistemology' refers to the debates on what knowledge is that was spawned by Gettier (1963). Outside of traditional epistemology, the pragmatist attitude has long been the default over the descriptive one. The notions of *knowledge*, *belief*, *justification* are often treated as tools of enquiry. For example, in formal and social epistemology, belief is often reconstructed as credence, represented with a probabilistic measure. Credence in turn could function as a parameter for computing expected utility (Resnik 1987). Justification could be reconstructed as the strength of confirmation, measured with conditional probability (Bovens & Hartmann 2003). Social epistemology was introduced by Goldman (1999b, 2010). See Hannon (2019) for an exposition of 'functionalist epistemology'.

[92] Cappelen (2018). The lack of contribution from non-traditional epistemology to post-Gettier epistemology, and the ignoring of formal accounts of belief or justification by the post-Gettier community is evident of this trend.

attitude to have when thinking about *descriptivity*. The scientific theories of 'fish', 'water', 'star' are still used to talk about the things we ordinarily think of as fish, water, star. Sure, there are changes in both the extension and the intension of our ordinary language term post-theory. However, these changes, if anything, clarifies our pre-theoretic conception and eliminate inconsistencies, rather than changing the topic in any meaningful sense of the phrase. Moreover, these revisions can be grounded on statistical principles.

The confusion on what counts as descriptive can be illustrated by the contradictory claims made by philosophers on whether a piece of theorising is descriptive. Steward Shapiro framed Tarski's theory of truth as a successful case of conceptual analysis.[93] In stark contrast, Hannes Leitgeb treats Tarski's theory of truth as a paradigm of successful rational reconstruction.[94] He also explicitly contrasted rational reconstruction against analysis, as a contrast between a normative project and a descriptive project, and also as a contrast between a revisionist and anti-revisionist project. This confusion over whether a descriptive project should involve revision can be seen all over the place. Grice and Strawson had opposing attitudes on whether ordinary language philosophy should involve rational reconstruction. Second, there is the more recent discrepancy between defenders of conceptual analysis such as Frank Jackson and the Canberra planners[95] and the actual practice of post-Gettier epistemology in their analysis of knowledge.[96] This indecision by the philosophical community on whether descriptive projects could involve revision has been in part noted by Cappelen. He remarked that although there are broad trends, perhaps no one is in full either a descriptivist or a revisionist, and most are a mix of both. This suggests that there is a gradient on which a descriptive project can involve more or less revision. This chapter aims to also clarify these confusions.

[93] Shapiro (2005)

[94] Leitgeb (unpublished manuscript). He also takes Kripke (1975) and Field (1994, 2008) to be improvements and elaborations on Tarski (1933, 1944), but their theories are still Tarskian.

[95] Jackson (2000), Braddon-Mitchell & Nola (2009)

[96] However, Jackson also argues that normative properties collapse into descriptive ones. See Jackson & Pettit (1996). Also Jackson (1998), (2001).

This chapter argues that successful descriptive ontological projects must involve revision. The aim is to clarify the ends of a descriptive project, and thereby improve our methodological framework via breaking down this false dichotomy between the descriptive and the revisionary. For this reason I think that using the term ‘descriptivist’ to refer to those who subscribe to the Strawsonian attitude is a misnomer. So I shall henceforth use the term ‘Strawsonian’ instead. The hope is that the so-called ‘descriptivists’ would see revision as a useful means for their ends. The issue of topic continuity itself would be dealt with in the next chapter.

Before moving on, I should clarify what the target of revision is. Within the empirical sciences, we often think of ‘revision’ in the context of revision over an existing theory. However, the debate between Strawsonians and its opponents is strictly on the revision over data. Epistemologists unapologetically revise over existing theories of knowledge, precisely because existing theories do not conform to our intuitions about knowledge. When we say that ichthyologists[97] have revised the term ‘fish’, we are making two claims: the modern scientific definition is a revision over, say, the Aristotelian definition of ‘fish’ *and* over ordinary discourse. The revision over Aristotelian taxonomy was insignificant; after all, descriptions could be false. However, the revision over ordinary discourse *was* salient, and warranted by scientific norms.

On a historic note, reconstructing existing terms or constructing new terms have been practiced for as long as theories have been proposed. Even ordinary language philosophers such as Austin advocated that ordinary language ‘is not the last word’, but could be ‘improved upon’.[98] Grice explicitly advocated for a rational reconstruction of terms to distinguish ordinary language philosophy from lexicography.[99] The age-old problem of epistemic scepticism existed upon allowing epistemology to deviate from our ordinary conception of knowledge. Moore’s appeal to common sense eliminated the very possibility for epistemic scepticism – because

[97] A sub-branch of zoology that focuses on the study of fish.

[98] Austin (1956)

[99] Grice (1989): 10, from a lecture in 1958

we do in fact claim that we have knowledge.[100] Revision has been practiced for far longer than it has been explicitly studied. The relatively recent Strawsonian attitude in the second half of twentieth century philosophy is testament that revision has always been the norm, not the exception.[101] This essay explains why this is so, and in effect why many ignore the Strawsonian complaint about topic change. On the other hand, the popular treatment of revision as a specifically 'normative' activity is likewise a misunderstanding.

The chapter argues that non-revisionary theorising is virtually impossible, via seeing the process of describing a term as curve fitting in statistical modelling. The explicit study of revisionist methodology isn't new. The most prominent examples are idealisation and abstraction in science. In 2.2 I introduce fundamental statistical techniques that are used to achieve idealisation and abstraction, which cumulates in the activity of curve fitting. I show that a good fit requires revision over data. I argue that this revision over data in statistical modelling is representative of the kind of revision in the Strawsonian concern. In 2.3, I argue that since data never exhaustedly represents an entity, we must measure the truth of a theory via indirect means. For this end, a descriptive project aims to successfully *generalise*. I demonstrate how the generality of a description is achieved via good curve fitting, and thereby argue that the very goal of a descriptive project requires revision. In section 2.4, I show that a Strawsonian attitude in the is a desire to overfit the data. Since overfitting acts against generality, the Strawsonian attitude acts against a fundamental goal of a descriptive project. I end the chapter with clarifications on what machine philosophy isn't.

[100] Moore (1939)

[101] Cappelen (2018) also makes a brief historic note on this point. This attitude began with the rise of ordinary language philosophy, and then the attitude began to die out at the dawn of the twenty-first century in various areas of philosophy: in philosophy of mind (Clark & Chalmers 1998; Chalmers 2010); in epistemology (Bovens & Hartmann 2003); in ethics (McKenzie 2007).

2.2 On Basic Statistical Norms and Revision

For centuries, scientists have been employing the statistical technique of *feature selection* to *abstract* over domain specific datasets, and various numerical methods to *idealise* over domain specific datasets. However, the techniques themselves have received little attention until recently, with the resurgence of interest in machine learning. Feature selection and methods to combat overfitting are fundamental statistical techniques for improving a model's predictive accuracy, which is equivalent to a model's ability to successfully generalise over its target dataset. Moreover, these techniques entail a model's revision over its dataset, known often as 'curve fitting'. This basic norm has been taken for granted in scientific theorising. So it's important for us to understand exactly what the business of curve fitting is, and whether the kind of revision it entails are those that the Strawsonians condemn. I begin by introducing the business of abstracting and idealising over data. I then go on to compare the revision over data entailed by these techniques with those that the Strawsonian would denounce.

2.2.1 Abstraction and Idealisation

Abstraction and idealisation are essential in scientific practice. Given parameterised data, abstraction is the *ignoring* of certain properties or parameters. For example, the model of the pendulum ignores parameters such as air resistance. Idealisation is the *distortion* of certain quantities, usually by means of rounding or distorting a measurement for the sake of uniformity in the distribution of data.[102] An example would be the elliptical model of our Moon's orbit around the Earth, which ignores the minor wobbles that occur in reality. The corresponding statistical techniques to achieve abstraction is *feature section*. Idealisation can be achieved via

[102] Frigg and Hartmann (2012) distinguishes between models of phenomena and models of data. I shall stick to models of parameterised data for the purposes of this chapter. Frigg and Hartmann (2012) also treats idealisation and abstraction synonymously as simplification. However, I maintain that the distinction is important, as we will soon see.

multiple techniques, depending on the situation. In statistical modelling, abstraction is also known as reducing the dimensionality of a curve, and idealisation as reducing the variance, or the smoothing out of a dataset. See figures 1 and 2.[103]

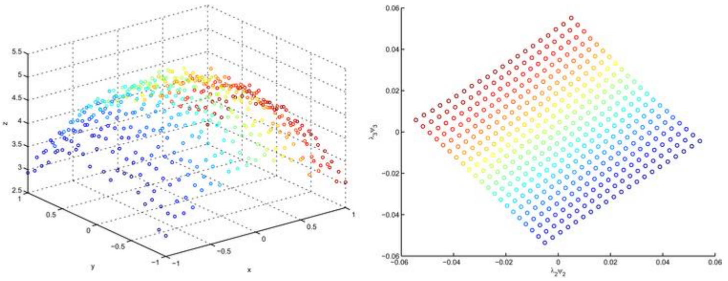

FIGURE 1
Reducing a 3-dimensional dataset to 2-dimensions

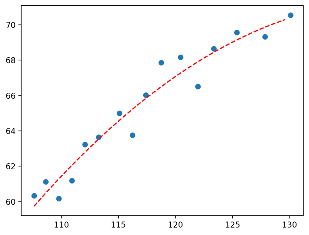

FIGURE 2
Idealising over 2-dimensional data

If one thinks of applying feature selection as simply one of many procedures in the process of curve fitting techniques, one might be inclined to treat abstraction as a kind of idealisation – the reducing of a parameter to zero at all points. However, this treatment has two issues. First, it ignores the differences in the statistical roles that idealisation and abstraction serve. Idealisation combats overfitting.[104] Abstraction serves a more complex role of reducing the dimensionality of a dataset. One reason for reducing dimensionality is to improve the statistical significance of a given dataset, since a model becomes increasingly more data hungry as more parameters are measured.[105] Another reason for reducing the dimensionality of a model is to eliminate parameters that reduce the robustness of a model. For example, if we include *colour* as a parameter for classifying swans, we would end up with a less robust model for classifying swans, since colour is an irrelevant parameter for *being a*

[103] In statistical models, a ‘curve’ can be n-dimensional, where ’n’ denotes a natural number. Images from https://3qeqpr26caki16dnhd19sv6by6v-wpengine.netdna-ssl.com/wp-content/uploads/2020/10/Plot-of-Second-Degree-Polynomial-Fit-to-Economic-Dataset.png; https://content.iospress.com/media/ida/2017/21-3/ida-21-3-ida150486/ida-21-ida150486-g003.jpg?width=755

[104] Breiman (2001). See Domingos (2012) for an overview of the issues of overfitting in statistical models.

[105] Known as the ‘curse of dimensionality’. See Domingos (2012). In the next section, we will see how this is also important for measuring the generality of a model.

swan.[106] Second, idealisation and abstraction are different statistical procedures. Idealisation is the business of deciding how to distort the values of a parameter, so that a distribution over that particular parameter would be more uniform. In the empirical sciences this is often done via more coarse-grained procedures. For example, physicists have decided that we should model the earth as a sphere even though it's in reality irregular like a potato. However, idealisation is a fundamentally mathematical, not analytical procedure achieved via weighing the cost of distortion: higher bias, against the benefits of a model's uniformity: lower variance. In machine learning, one common technique is to apply *regularisation*, which is a numerical technique that penalise an overtly complex model via the addition of a regularisation term to the model. One could also try to measure the variance of a model with respect to a given dataset.[107] Abstraction is the business of deciding which parameters (features) to keep (select), and is hence a primarily analytical procedure.[108] For the above two reasons I shall maintain the distinction between idealisation and abstraction.

2.2.2 The Aversion to Revision

Following on from Chapter 1, philosophical theories should likewise idealise and abstract over linguistic data – intuitions. Philosophers are already familiar with feature selection. For example, when theorising about knowledge *in epistemology*, we ignore psychological criteria for having knowledge. In fact, philosophers are pretty good at this analytic procedure. After all, this is part of the philosophical training. However,

[106] Another way to describe this is to say that having a certain colour is *inessential* for being a swan.

[107] There are multiple ways to do this, depending on the mathematical properties of the model. See the following reference for a summary: https://towardsdatascience.com/measure-variance-of-statistical-model-e3b4725095b6

[108] Although oftentimes the decision can be implicated by matters of statistical significance. In situations where it is unclear to us which parameters one should eliminate, we might even rely on comparing certain robustness measures between strategies. For example, the elimination of the parameter 'lives in water' for the category of 'fish' was because the resulting categories were statistically more uniform. Still, fundamentally the procedure is distinct from that of idealisation. In fact, the business of feature selection is still more or less an 'art' (Domingos 2012), as opposed to the crisp clear procedures for curve fitting.

philosophers struggle with idealisation. The Strawsonian desire for a theory of X to perfectly capture our proper intuitions is in essence a struggle to reduce variance for fear of bias. By 'proper' I mean data that the philosophical community has agreed upon.[109] The Strawsonian desire rests on the faulty view that while observations could be erroneous and entities in the world could have minor irregularities, our proper intuitions are both infallible and regular in virtue of them being the products of our socio-linguistic communities. In Chapter 1, I have discussed how this view is incorrect, and that intuitions are evidentially akin to observations. In particular, intuitions ought to be treated as statistical data. I will now demonstrate how the aversion to revision is an aversion to bias, and in the process, unearth the culprit – boolean reasoning.

Let's consider the case of post-Gettier epistemology. Our proper intuitions on whether a person has knowledge is taken as data to be satisfied fully by our theories of knowledge. This means that whenever the community of traditional epistemologists have an agreed on intuition about knowledge, every epistemological parameter of that piece of intuition must be satisfied by the theory. For example, the original Gettier case has revealed certain epistemological claims that epistemologists share. One of these is that justification is fallible. As a consequence, any theory of justification that entails an infallible account of justification would be rejected as failing to capture our data on what justification is. This is the desire to perfectly fit our description (of justification) to our intuitions (about justification). In Chapter 1 we have discussed the widespread obsession with *theory-free* intuitions. Underlying both our desire for theory-free intuitions *and* our desire to perfectly fit our intuitions is the fear of (statistical) *bias*. This fear is a consequence of the pervasiveness of boolean representations within traditionally analytic, ordinary language philosophy. A person's belief is either justified or unjustified; an action is either moral or immoral; a proposition is either true or untrue etc. Representing this in a statistical setting, we see that philosophers have been treating the parameters of their data as spanning over a

[109] Of course, Strawson did not mean to say that our theories should fit also the incorrect use of language.

scale of either 0 (untrue) or 1 (true). Moreover, philosophical theories themselves are boolean. Suppose we represent a typical theory of knowledge in post-Gettier epistemology as $T_E(K)$. Then $T_E(K)$ is not represented as a function wherein one can compute values that would provide one with varying degrees of knowing. Rather, $T_E(K)$ would be represented by boolean parameters such as boolean justification, boolean truth, boolean belief etc.[110] Since philosophical theories and data have been treated as boolean, any bias would have been devastating for the accuracy of a theory. The worry is compounded by the fact that we are terrible at identifying corrupt data.[111] So the main culprit behind the aversion to revision is actually a boolean paradigm of representation and theorising.

There is nothing wrong with representing beliefs or epistemic justification as non-boolean, and this is possible even in the domain of mathematics.[112] Even within post-Gettier epistemology, epistemic contextualism is evidence that our epistemic parameters shouldn't be represented as boolean. This is because it would be absurd to think that a change in context, however minor, should shift the criteria for a proper knowledge ascription so drastically, such that the requirement for a parameter swings from 0 to 1 or vice versa![113] More generally, what we need in philosophy is a mentality shift – not only do we need to think of intuitions as statistical data, we need to also think of properties (parameters/features) as generally non-boolean. Boolean parameters should appear as special cases. This is not a metaphysical point about the

[110] Of course, Bayesian epistemology rejects these assumptions from traditional epistemology. In this regard Bayesian epistemology engages in machine philosophy.

[111] This isn't unique to intuitions, of course. That's why empirical scientists have been relying on statistical norms rather than only analytic procedures.

[112] For example, even though my belief that 'there is no largest prime number' and my belief that 'every simply connected, closed 3-manifold is homeomorphic to the 3-sphere' (Poincaré conjecture) are both justified, they are not equally justified. My belief that 'there is no largest prime number' is based on my own first-hand practice of proving the proposition using clear, established logical methods. However, my belief in the Poincaré conjecture is based on my belief that Perelman's proof was correct and my trust in the mathematical community. Even less so justified would be my belief that the Goldbach's conjecture is correct. While there is plenty of computational evidence suggesting its truth, a mathematical proof hasn't yet been found. Still, I'd like to think that my belief of this conjecture (among many other unproven but widely accepted mathematical propositions) is justified.

[113] Of course, this point was never made within the contextualist literature itself, since the entire domain of traditional epistemology presupposes boolean parameters. It nevertheless is a previously unexpected consequence of the contextualist discovery.

things that we theorise about, but a representational and methodological point about how we theorise over the data we have.[114] We need to weigh the bias of our philosophical theories against variance, and respect the statistical norm of idealisation that is crucial for the generality and robustness of a scientific model.

2.2.3 Is Idealisation Revision?

I shall first provide an explicit representation of what 'revision' could mean. Suppose we have a topic X in a domain O, where 'O' represents ordinary discourse. A lexical entry of X in O is effectively a model of X in O, so let's call it $T_O(X)$. Now consider a specialised domain D, and a theory $T_D(X)$. Using the notation from Chapter 1, let's call the set of all intuitions of X in D $I_D(X)$, and the set of all instances of X that satisfy $I_D(X)$ X_D. Furthermore, let us specify similarly so for O: $I_O(X)$ and X_O. Suppose there are *n* number of D-specific features/parameters for X, let's represent these parameters as a vector $F_D = \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_n \end{pmatrix}$. Then for some datapoint x_{Di} belonging to X_D, $x_{Di} = \begin{pmatrix} w_{i1} \\ w_{i2} \\ \cdots \\ w_{in} \end{pmatrix} \cdot F_D$ where *w* represents weights for each of the *n* parameters for x_{Di}. Given our above discussions, a revision could happen in a number of ways. First, there could be revision of X_D over X_O.

Following on from Chapter 1, experts within D necessarily assent to $I_D(X)$ rather than $I_O(X)$. Trivially, $I_D(X)$ is distinct from $I_O(X)$ in virtue of D being distinct from O, and so X_D differs from X_O. So there is already revision at the level of data. Every expert, including philosophers, engage in this kind of trivial revision. The aforementioned kind of feature selection by epistemologists amount to this first type

[114] One might complain that I am simply doing something like applying fuzzy logic to the issue of vagueness. Instead of answering the question of 'at what point does X become non-X', I circumvent the issue by saying that 'X comes in degrees', and that our parlance of 'X or non-X' is simply an approximation of the reality of degrees of X. However, this is really an orthogonal issue, since…

of revision. However, this isn't the kind of revision that's of interest to anyone. It follow that a revision of T_D(X) over X_O or over T_O(X) also isn't of concern. Rather, the concern is over the revision of T_D(X) over data. This could itself happen via two steps.

First, there is the revision of X_D over itself via feature selection. This is when a revised X_D eliminates f_k from F_D such that the new F_D has $n-1$ parameters rather than n. A new x_{Di} would now look like $x_{Di} = \begin{pmatrix} w_{i1} \\ w_{i2} \\ \dots \\ w_{in-1} \end{pmatrix} \cdot F_D$. This procedure could be repeated until satisfactory. Let's call this kind of revision R(X_D, X_D).

Second, there is the revision of T_D(X) over X_D, after we have revised X_D via feature selection. This process is the computational procedure of *idealisation*, in order to balance bias with variance. Let's call this kind of revision R(T, X_D).

One might argue that R(T, X_D) does not constitute the kind of revision that concern philosophers. When we talk of 'revision' in metaphilosophical debates, we mean something like revising the term 'fish' to exclude whales, revising the term 'salad' to include fruit concoctions, revising the term 'marriage' to include same-sex couples etc. These examples don't look like idealisation over data.

This objection stems from a misunderstanding of what curve fitting means. A curve needn't only fit over large datasets, the curve needn't even be continuous. Taxonomy often involve certain pre-selected features, before curve fitting is applied. Consider the case of 'fish'. The parameter of *whether an animal lives in water* has been eliminated in the process of feature selection. This was necessary for the revision of 'fish' to exclude certain marine animals such as whales, starfish, cuttlefish. Since feature selection is generally an analytic procedure, one might say that this at best shows that revision might be a result of abstraction, not idealisation. But did I not say that philosophers are already pretty good at feature selection? To some extent yes. What I meant was that philosophers excel at *positively* identifying domain specific features. Our education has trained us to dig up as many features as possible for any given X within a specific domain. For example, epistemologists have correctly identified features such as whether a belief is true by luck, or whether a belief is formed via cognitive abilities as being epistemological parameters; while of course ignoring psychological or sociological parameters. However, we are hopeless at *eliminating* features that are *within the domain* and on which the theorised entity in question depends. And *this* is the difficult (and important) kind of feature selection that often requires a statistical analysis, simply because it's virtually impossible to analytically determine which features one should keep when that feature appears to be relevant. This was the case for 'fish' as it is for many other pieces of empirical theorising. Analytically, ichthyologists (a type of biologist specialising in fish) had no reason to think that 'living in water' is a biologically irrelevant feature for being a fish. After all, pre-theoretically, we categorised 'fish' on that very feature. What really forced the ichthyologists to give up that feature was a clear superiority of statistical uniformity, or, better fit *after* one eliminates that feature. For example, clear patterns which divided creatures that could breathe in water, from those that cannot, only formed when one eliminates the aforementioned feature. The revision of 'fish' itself happened as a result of the statistical distribution after the elimination of certain features, which was itself a result of trying to fit a curve over the dataset of all kinds of observed 'fish'. The 'curve fitting' in this case might be a simpler kind of pattern recognition that is significantly less involved than a handwriting recognition algorithm. But fundamentally, both are curve fitting, and for that matter, pattern recognition. In particular, the strive for a good fit is a strive for generality, and hence truth.

In other words, the revision $R(X_D, X_D)$ (abstraction) that Strawsonians worry about is oftentimes a result of trying to successfully execute $R(T, X_D)$ (to idealise). This is an unsurprising finding for statistical methods, since feature selection is in practice intertwined with trying to produce a model with a good fit. In the next section, I argue why the desideratum of generality is central to descriptive projects. I will return to this 'fish' example to show how the revised description results in a more general theory than the ordinary language description.

2.3 Descriptive Projects and their Aims

A descriptive project is marked by its aim to *truly* describe. However, the truth of a theory cannot be directly measured, but only via measuring how well the theory generalises. Building on our knowledge of statistical learning, I argue that descriptive generality is the measure of how well a descriptive theory fits independent datasets that are within the theory's intended domain of description. To begin with, we need an account of what a *descriptive* theory is. We often understand 'descriptive' in tandem with 'normative'. Consider some examples from distinct domains:

a. From decision theory:

'Descriptive decision theory is concerned with characterising and explaining regularities in the choices that people are disposed to make. It is standardly distinguished from a parallel enterprise, normative decision theory, which seeks to provide an account of the choices that people *ought* to be disposed to make.'[115]

b. From ethics:

'…the term "morality" can be used either

1. descriptively to refer to certain codes of conduct put forward by a society or a group (such as a religion), or accepted by an individual for her own behaviour, or
2. normatively to refer to a code of conduct that, given specified conditions, would be put forward by all rational persons.'[116]

c. From philosophy of logic:

'Stanovich' discussion of rules introduces a distinction between normative, descriptive and prescriptive rules. We give brief characterisations of the three kinds, followed by representative examples.

• normative rules: reasoning as it should be, ideally

– Modus Tollens: $\neg q, p \rightarrow q \vDash \neg p$

[115] Chandler (2017)

[116] Gert & Gert (2016)

– Bayes’ theorem: $P(D|S) = \dfrac{P(S|D)P(D)}{P(S)}$

• descriptive rules: reasoning as it is actually practiced

– many people do not endorse Modus Tollens and believe that from ¬q, p → q nothing can be derived

– in doing probabilistic calculations of the probability of a disease given a cluster of symptoms, even experts sometimes neglect the ‘base rate’ and put $P(D|S) = P(S|D)$

• prescriptive rules: these are norms that result from taking into account our bounded rationality, i.e. computational limitations (due to the computational complexity of classical logic, and the even higher complexity of probability theory) and storage limitations (the impossibility to simultaneously represent all factors relevant for a computation).’[117]

d. From metaphilosophy

‘For some, success is measured by a true description of, for example, what knowledge, belief, morality, representation, justice, or beauty is. For others, the aim [of philosophy] is figuring out how we improve on what there is: how can we improve on knowledge, belief, morality, representation, justice, beauty etc.?’[118]

Although the above examples reside in distinct domains of enquiry, they consistently mark the distinction between the goals of a descriptive versus a normative project. In each case, the descriptive project aims to provide a true description of a certain target. Descriptive decision theory would contain theories that tell us how people actually make decisions, what people actually perceive as being rational etc. Descriptive ethics contain theories that tell us what people in fact perceive as ethically good, fair, just etc. On the other hand, a normative project would produce theories that aim to improve upon a certain target. e.g. Normative decision theory would contain theories that tell us how people *should* make decisions, what people *should* see as being rational etc.

[117] Stenning & van Lambalgen (2008)

[118] Cappelen (2018)

2.3.1 Deal with the Data

Data is our gateway to facts. In specialised discourse, data is parameterised. For example, when epistemologists examine a piece of intuition about knowledge, they might look for specific features such as cognitive ability, luck, justification, truth, modal conditions etc. In philosophy, much of our data come in the form of intuitions. As I have argued in Chapter 1, intuitions are ungrounded, domain dependent propositions that a community would use as premises in arguments. If we want to theorise about *knowledge*, then our intuitions might include propositions such as 'a person *S* knows that *p* only if *p* is true', '*S* knows that *p* only if *S* believes *p*', 'In a certain scenario *e*, *S* knows that *p*' etc. These intuitions are propositions in which the term 'knowledge' properly ascribes. Now let K_E be the set of all intuitions about knowledge in the domain of epistemology. Then a descriptive project of what knowledge is aims to produce a theory of knowledge $T_E(K)$ such that $T_E(K)$ would satisfy propositions in K_E. More generally, given some X and the set of intuitions about X in D: X_D, a theory of X in D: $T_D(X)$ aims to satisfy X_D. In mathematical terms, $T_D(X)$ *maps* X_D. On the other hand, in a normative project, $T_D(X)$ would aim to improve upon X_D.

One might object that my account misrepresents what a theory in a descriptive project aims to do. It might seem that data isn't usually the target of our theorising. For instance, the theory of universal gravitation $T_P(UG)$ was never meant to capture the observed measurements of moving bodies, but the underlying physical relations that govern their motions. The measure of how well the theory captures observational data is an indirect way to measure whether the theory is true of the physical relations. $T_P(UG)$ is not a revision over the physical relations themselves.

However, the objection ignores that in practice, there is no way to theorise about something, to describe something *other than* via data. This methodological fact is indifferent to background metaphysical views about whether one thinks we have 'direct access' to the world or not.[119] T_P(UG) is ultimately stated in parameters represented by our data. Even if one maintains that T_P(UG) is a description of some underlying physical relations, they cannot deny the fact that the actual theory – down on paper – is expressed in terms that *strive to* accurately map our *data*. Another way to talk about T_P(UG) in physics is that it's an *explanation* of our data. Empirical verifiability itself requires that our theories are testable via observational data.[120] In this regard, the revision over data via idealisation and abstraction constitutes a revision over the target of theorising in the empirical sciences. Even if we ignore pragmatic goals as inessential for descriptive projects: such as being applicable for engineering, having explanatory power etc., physical theories still must revise over observational data even if merely for the sake of predictive accuracy, which is a fundamental epistemic goal for physical theories.[121]

Thus, the distinction between a descriptive project and a normative project lies not in the mere existence of revision. However, there is still a prima facie difference in their *goals*, which seem to hinge on the deviation from facts. Prima facie, a descriptive project aims to *describe*, whereas a normative project aims to *improve*. So the question is: How should we account for distinction between descriptive projects and normative projects, ignoring revision? The goal of a descriptive project can be stated in two segments. First, the goal of a descriptive project on some target X in domain D is to provide a true description T_D(X) of X in D: X_D, given that T_D(X) is statistically satisfactory. Second, since the truth of T_D(X) with respect to X_D is often

[119] For example, in the Kantian sense.

[120] Here I would like to make a remark I see as being constitutive of the whole debate on revision. Empirical verifiability via data is not a flaw, but simply a condition of our methodological investigations. The interpretation of this fact as a lack is due to a wishful picture of knowledge and justification, losing sight of the goals that knowledge serves. One psychological note for the traditional epistemologists before moving on is to see big data as bypassing the problem of induction, aerospace engineering as bypassing the problem of gravity. In both cases, what matters is that we are able to achieve our goals, rather than dwelling on whether we can deductively justify generalisations, or whether we can biologically fly. The hope of a meaningful, tractable diehard anti-revisionist project is the hope of finding a way to give humans superpowers. It's not an impossible project, but a terribly imprudent one for science. See a similar remark in Wheeler (2016) and Harman (1986).

[121] Elgin (2007) provides a detailed account of how trading representational accuracy for idealised and abstracted models better serves the epistemic goals of science.

tricky to measure, partly because we are stuck with data[122], the success of $T_D(X)$ is to be measured by how well $T_D(X)$ serves *the goals of being a true description of* X_D, which I argue, is how well $T_D(X)$ generalises to independent datasets.

2.3.2 On Generality

Typically descriptive theories – in the natural sciences, in economics, in linguistics – are motivated by the very desire to be general. First, generality is the indicator that a theory is *true* of its target, because its target often cannot be exhausted by the data that we use to derive a theory. Second, generality is crucial for a theory's fruitfulness, including explanatory power, predictive accuracy etc. In the previous section, I argued that a descriptive theory $T_D(X)$ aims to describe its target dataset X_D. In this section, I argue that the generality of $T_D(X)$ with respect to X_D is the fundamental goal for $T_D(X)$. This is because the generality of $T_D(X)$ with respect to X_D is the sole measure of whether $T_D(X)$ is *true* of X_D. I begin by explaining how generality in a statistical model can be evaluated by the technique of cross-validation, thus providing a characterisation of generality for descriptive theories. Then I demonstrate how generality is our best bet at evaluating the truth of a theory.

Consider a set of labelled data that we would use to generate $T_D(X)$. The labelled data consists of entities that we would judge as either X or non-X.[123] Let us divide this dataset into two independent datasets with roughly the same probability distribution: a training set $X_{training}$ and a test set X_{test}, where $T_D(X)$ is to be derived strictly from $X_{training}$. A general $T_D(X)$ would correctly identify data in X_{test}. At the limit, a true $T_D(X)$ would correctly identify all entities in the world as either X or not-X. Now the

[122] If truth is strictly the adherence of a theory to its target data, then underdetermination tells us that the truth of a theory would be impossible to measure. However, the point is not whether truth is granted, but whether the roles of truth are fulfilled. This is the point of a pragmatist approach to epistemology, i.e. Elgin (2017), Hannon (2019).

[123] Answering an ontological question in philosophy is easily understood to be a classification task in statistical learning. Of course, philosophers aren't only concerned with ontological questions, nor need ontological questions be questions of classification. I only use ontological questions as a case study in this chapter, since it's often at the centre of debates between Strawsonians and revisionists. Here I present it as a classification task for simplicity of illustration. The same statistical norms that results in idealisation and abstraction holds as much for classification as they hold for regression. So my choice of examples shouldn't matter.

issue is how we are to judge whether $T_D(X)$ correctly identifies new data. There are three strategies to judge whether $T_D(X)$ successfully generalises.

First, given a datapoint *n* in X_{test}, we measure $T_D(X)$ against *n*. $T_D(X)$ successfully generalises iff $T_D(X)$ correctly identifies *n* for all *n* in X_{test}. This strategy complies with the Strawsonian attitude. Gettier styled counterexamples work in this way: a proper intuition of knowledge in epistemology could singlehandedly refute a theory of knowledge. We have already established the problematic nature of this boolean attitude on intuitions in Chapter 1 and in 2.2. In the next section we shall see how this strategy in fact acts against generality. In statistical terms, this strategy results in low bias but high variance, or overfitting.

The second strategy is to construct $T_D(X)$ on $X_{training}$ with theoretical norms such as simplicity, and then evaluate any new data by measuring it against $T_D(X)$. The issue with this strategy is that it violates the factive element of generalising. There is no way of telling whether $T_D(X)$ *correctly* generalises to new data, because in this picture, $T_D(X)$ is the golden standard once it has been constructed. This strategy results in low variance but high bias, or underfitting. In practice, we rarely underfit our datasets by mistake. Instead, we use this strategy to hypothesise when data is insufficient and when aggressive feature selection isn't possible. To use the same example from epistemology, the traditional analysis of knowledge as justified true belief would be such a hypothesis.[124] Both strategies fail to achieve generality since

[124] Although, granted, Plato and his contemporaries might have practiced aggressive features selection as a matter of basic philosophical practice. In fact, such a methodological attitude in philosophy had been taken granted throughout history. Philosophers loved to provide grandiose generalisations that contained few parameters. This is however, not unique to philosophy, nor is it the fault of these philosophers. Scientists often had to hypothesise without sufficient data. This is not always bad science, since humans construct theories not only with data, but also with higher order criteria that are often pragmatic and domain dependent. These higher order methods are essentially ways to circumvent the hard issue of having insufficient data, and the methods have worked quite well (e.g. What Popper described as the endless process of conjecture and refutation is such a method). Still, ultimately these higher-order criteria are merely cruder forms of statistical learning – we evaluate the plausibility of our hypothesis against new data or hypothetical cases, even if sometimes we have only anecdotes (e.g. Bayesian epistemology). It was only recently that more refined statistical learning – specifically in the form of machine learning – became possible, and this was only because 1. computing power has rapidly improved and more importantly 2. data became abundant and cheap. It's important to note, however, that even machine learning isn't without higher-order norms, nor is 'pure' statistical learning possible (since we would need an infinitely large dataset). We still need feature selection, for example. Still, the recent computational techniques we have for employing large datasets is a huge leap in our scientific toolbox. Google's Alphafold and Caltech's algorithm for solving partial differential equations (Li et. al 2020), both notoriously difficult problems, are testaments to this leap.

they either overfit or underfit. So we need a middle ground. Now I introduce a standard technique in statistical learning as our third strategy.

The third strategy compares how well $T_D(X)$ fits $X_{training}$ with how well $T_D(X)$ fits X_{test}. We say that $T_D(X)$ correctly identifies *n* just in case $T_D(X)$ fits $X_{training}$ well and *n* satisfies $T_D(X)$ within or relatively near the distribution of how well $T_D(X)$ fits $X_{training}$. $T_D(X)$ successfully generalises just in case it fits both $X_{training}$ and X_{test} well, without its own parameters having been influenced by data from within X_{test}. Under this strategy, X_{test} cannot directly make $T_D(X)$ satisfy any *n*. Therefore, if $T_D(X)$ fits both $X_{training}$ and X_{test} well, then $T_D(X)$ would likely generalise to new data. This third strategy is the general strategy that is used throughout descriptive projects in modern times, from astrophysical models to machine learning algorithms. When we say that the characterisation of energy $E = mc^2$ does not generalise well for describing the energy possessed by a moving mass, we mean that when the velocity of a mass is non-zero, any measurement of the energy against the mass would lie beyond any reasonable distribution of $E = mc^2$. Here is a pair of plots to illustrate this description.

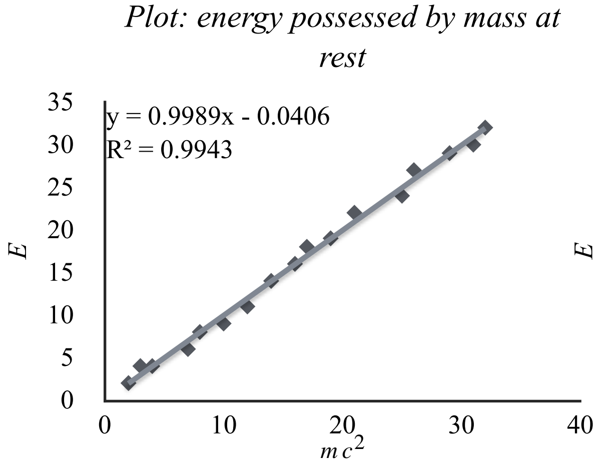

FIGURE 3

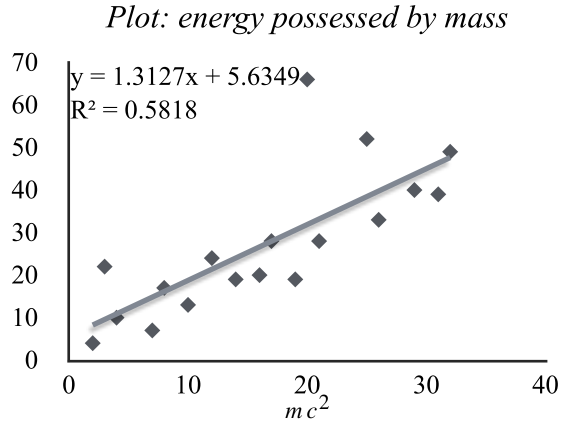

FIGURE 4

In the two figures, we plot the measured energy *E* of bodies against mc^2 where m is the mass of each body and c is the speed of light in vacuum. The lines are linear curves of best fit to the data, they represent our theory. In FIGURE 3, the line is roughly $E = mc^2$, and the line in FIGURE 4 is roughly $E = 1.3mc^2$. R^2 is the coefficient of determination, which lie between 0 and 1. It measures how well the lines fit their respective data. The greater the R^2, the closer the line fits its data.[125] If we consider the line $E = mc^2$ instead of the line of best fit for the given data, we would have even greater differences between the R^2 values of the two plots. This means that the theory which fits the data in *Figure 3* does not generalise to data in *Figure 4*. Now, why is our theory for the energy of a mass *at rest* $E = mc^2$ rather than whatever the line of best fit is? This is because $E = mc^2$ would fit better *on average* for all such datasets – measuring the energy of rest mass. It is impossible to gather all possible datasets for any domain of enquiry, we have only a limited dataset for deriving our theory. Therefore, the generality of a theory must be measured by considering how well it fits *independent* datasets. In data science, this is a fundamental procedure called 'cross validation'.

Cross validation is a technique that makes use of limited (though still relatively large) data to measure how general a particular curve is, the curve being the representation of our models/theories/hypotheses.[126] It works by partitioning the data that we have into independent subsets, and the minimum size of which would depend on how many parameters we are measuring. The more parameters we have, the more data hungry cross validation will be.[127] We compare a model's variance and bias across these independent datasets, via measuring the *fit* of the model in each of these datasets. The idea is to tweak the model (thus balancing bias against variance) so that the model would have a low predictive error across these independent datasets (see

[125] A close fit does not entail that the theory (the line) generalises well. A diehard descriptivist theory that fits every given datapoint perfectly would have an R-squared value of 1, this does not mean that the same theory would have a low R-squared value when considering new data. The point of a good fit is that the theory should have high R-squared values for all data within the descriptive domain of the theory.

[126] Of course, cross validation isn't the only technique for evaluating generality, though it is the best technique we have *given that* there is sufficient data and that we have done features selection properly. In the empirical sciences, other methods have been historically used due to the scarcity of data. For example, scientists might measure the deviation of newly discovered data from a model and make a decision, in combination with theoretical criteria, on how and whether to modify the model. The decision on what counts as an 'anomaly' in the empirical sciences has been an important part of the scientific method. However, we needn't worry about whether cross validation can actually be applied in every area of philosophy, since the point here isn't about cross validation, but *generality as a measure of truth*. Cross validation is simply used as an example to argue for this point.

[127] This is the curse of dimensionality. See Domingos (2012),

FIGURE 5).[128] There are various mathematical measures for fit in a statistical model, and choosing which one to use is often a matter of experience rather than hard rules.[129] Still, they all amount to the goal of improving the predictive accuracy of the model. This is another way to say that we want the model to generalise well.

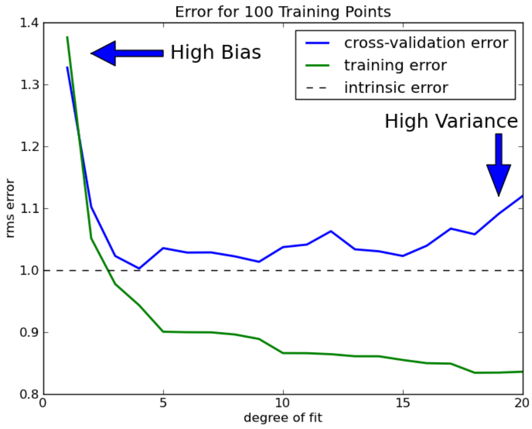

FIGURE 5

2.3.3. On Truth of a Theory and Two Notions of Generality

Now the question is: Is measuring predictive accuracy our best bet at evaluating whether a theory is *true*? We often hear physicists talking about theories as being ‘more true’ than one another. They could mean two distinct things by this, corresponding to two distinct propositions expressed by the phrase ‘theories aim to be general’. On the first reading, saying that ‘a theory $T_D(X)^*$ is more general/more true than $T_D(X)$’ is to say that $T_D(X)^*$ applies to a broader dataset than $T_D(X)$, where ‘broader’ refers to having more parameters. This is what we mean when we say, for example, that relativistic theories are more general than Newtonian theories. Returning to the above example, $E = mc^2$ is on this reading ‘less true’ than $E^2 = m^2c^4 + p^2c^4$, since the former cannot generalise to datasets containing bodies in motion. However, the former equation is *only meant to* describe bodies at rest, as mentioned earlier. In this regard, $E^2 = m^2c^4 + p^2c^4$ doesn’t actually have a better predictive accuracy than $E = mc^2$, since $E = mc^2$ applies to a different dataset – one that includes only bodies at rest (the parameter of ‘velocity’ set to 0 everywhere). This cannot be what ‘descriptive accuracy’ or ‘the truth of a theory’ means. We can make

[128] http://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/practical.html

[129] Domingos (2012)

true descriptive statements without make a *general* statement. For example, I can say the true descriptive statement that ‘there are black swans in Lake Griffin’.

On the second reading, saying that ‘$T_D(X)^*$ is more general or more true than $T_D(X)$’ is to say that $T_D(X)^*$ has a *better predictive accuracy in its target dataset* than $T_D(X)$ has in its target dataset. For example, consider the scientific classification of ‘fish’ vs. the ordinary English use of the same term. Suppose we were to statistically model both descriptions over their respective datasets. The curve representing the ordinary English classification would have a significantly worse variance and bias than the scientific classification, if we were to measure the fit of the curves to their respective datasets of what each description would consider to be ‘fish’.

One might point out that the ordinary English description and the scientific description capture distinct datasets, like in the case of $E = mc^2$ and $E^2 = m^2c^4 + p^2c^4$. So one might object that we cannot say that the scientific description is an improvement over the ordinary English one, but a change of topic. And so we’re back in the Strawsonian armchair. However, the difference between the case of ‘fish’ and the case of ‘energy’ is that in the case of ‘fish’, there is a significant improvement of statistical salience in the scientific description over its target dataset. $E^2 = m^2c^4 + p^2c^4$ on the other hand, does not improve the predictive accuracy of $E = mc^2$, when the latter is used only to measure the energy and mass of bodies at rest. The modern scientific description of fish has a much better chance at correctly classifying a new creature as either ‘fish’ or ‘not fish’, whereas the ordinary English one fails to do so since the corresponding curve must, for example, decide on what to do with features such as ‘being able to breathe in water’. The corresponding model for the ordinary English term must either have very few features (at the cost of losing robustness of the classification, or even making it outright pointless) or be extremely complex (e.g. specifying extra conditions under which a ‘fish' cannot breathe in water) in order for it to have a comparable predictive accuracy with our scientific description of fish. It is on this second reading of ‘generality’ as predictive accuracy that scientists use to evaluate whether a description is to be accepted (as being true).

2.4 Combatting Overfitting in Philosophy

In the previous section, I have provided an account of what it means for a description to be general, and have argued that generality is the measure of truth for actual theorising. In this section, I demonstrate that given a target dataset, in order to satisfy generality, a descriptive theory should obey the fundamental norms of statistical learning. I begin by introducing the problem of overfitting, and how it acts against generality. I argue that the Strawsonian attitude is an attempt to overfit. In the second half, I illustrate how philosophical theorising can be construed as a curve fitting exercise.

2.4.1 On the Problem of Overfitting

Given the task of describing a target dataset, overfitting happens when the description is more accurate of its training data than of its test data, when it could have been more accurate of its test data.[130] Since generality is a measure of how well a theory fits independent datasets, overfitting is a sufficient source for hindering generality. Now the question is whether the Strawsonian attitude amounts to a desire to overfit. Consider the example of post-Gettier epistemology.

The entire narrative of post-Gettier epistemology rests on counting every piece of proper intuition as valid counterexamples. In Gettier's paper, he provided a description of a hypothetical scenario, in which we would intuitively find a person who holds a justified true belief without knowledge.[131] Intuitively, we do not accept

[130] See https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76 for a very useful illustration of overfitting. The causes of overfitting can be multifaceted, and there is no single class of optimal solutions. Using too many features is a common cause of severe overfitting (Basheer & Hajmeer 2000). There is also no one best solution, but a multitude of standard methods such as cross-validation, optimal feature selection, regularisation (see Domingos (2012) for overview, see Wolpert & Macready (1997) for the no free lunch theorem). Thus, the technicalities of avoiding overfitting could only be fruitfully managed for specific problems, rather than at the meta level.

[131] Although Weatherson (2003) criticises the use of intuition, pointing to post-Gettier epistemology as the prime example, his core argument can be understood as a complaint about overfitting in the same sense as argued here.

that some *x* is F but at the same time deny that '*x* is F' is true.[132] Thus, to accept that a person *S* does not know that *p* is to accept that '*S* does not know that *p*' is true; and to accept that '*S* does not know that *p*' is true is to accept that *S* does not know that *p*. Thus, our intuitions of the Gettier-cases are our intuitions of correct ordinary discourse. After all, the entire enterprise was possible due to the rise of ordinary language philosophy, which tried to avoid injecting post-theoretic influences into pre-theoretic discourse.[133]

The methodology of post-Gettier epistemology can be thus stated: First, pit our best current descriptive theory $T_E(X)$ of some entity X against our intuitions X_E of X. Second, study any discrepancies between $T_E(X)$ and X_E. Finally, revise $T_E(X)$ until $T_E(X)$ satisfies every item in X_E, thus correctly describing X.

One problem immediately arise. As mentioned in the previous section, data cannot exhaust the topic of enquiry.[134] Thus, the actual dataset we have for deriving or revising X_E can only be a subset $X_{E\_training}$ of X_E. Mathematically, a $T_E(X)$ that is revised to fully satisfy an increasingly large $X_{E\_training}$ would also increase in complexity and variance. For simplicity of illustration, and without loss of generality, let us begin with SIZE($X_{E\_training}$) = 2 where the data considered measures only two relevant parameters.[135] The simplest model $T_E(X)_2$ that could fully satisfy $X_{E\_training}$ would be linear. If we add even just one datapoint so that now SIZE($X_{E\_training}$) = 3, then the simplest possible model $T_E(X)_3$ would still be linear, and only if the new datapoint already sits on $T_E(X)_2$, in which case $T_E(X)_3$ would just be $T_E(X)_2$ and no revision is required. However, if the new datapoint violates $T_E(X)_2$, then the

[132] Or if we accept the almost universally adopted *T*-schema that '*p*' *is true* iff *p*.

[133] Although this goal cannot come to fruition, as I have argued in Chapter 1.

[134] If ϕ is an entity encountered in ordinary discourse, then Φ would be infinitely large, since there would be infinitely many correct ascriptions of ϕ in ordinary discourse.

[135] Philosophical enquiries, including post-Gettier epistemology, often involve the discovery of parameters, such as adding 'sensitivity' (Nozick 1981) or 'safety' (Sosa 1999) to the theory of knowledge on top of modifying an existing parameter such as arguing that justification should be internal (Prichard 1950, Steup 1999, Bonjour 2010). However, many of the parameter such as 'sensitivity' are arguably not independent from 'justification', and thus can be treated as one parameter with different values (anti-luck justifications are better than ones that involve luck). In any case, adding parameters contributes to overfitting. Feature selection (Guyon & Elisseeff 2003, Koller & Sahami 1996) combats this source of overfitting.

Strawsonian would call for revising $T_E(X)_2$. In this case, the simplest $T_E(X)_3$ would be a degree 2 polynomial. In general, for any $T_E(X)_k$ of $X_{E\_training}$ where SIZE($X_{E\_training}$) = k, the complexity of $T_E(X)_{k+1}$ is at least that of $T_E(X)_k$. But what is wrong with complexity? Carnap thought of simplicity as only a weak criterion that applies in the face of competing theories that otherwise do equally well.[136] Williamson's complaint on the increasing complexity of the analysis of knowledge was due to the worry of losing informativeness.[137] However, the more fundamental reason for simplicity is because a simpler theory is more likely to generalise well, and this is because we can never have finite and exhaustive data.[138] This is what overfitting is – the increasing satisfaction of the training dataset at the cost of potentially new data. In post-Gettier epistemology, all available data has been treated as training data. The issue with this strategy is threefold. One, the theory would overfit the data. Two, we cannot verify our theory with an independent test set – if a new datapoint is violated by an existing theory, the it is *added* to the training set. i.e. We do not perform cross validation. Three, the process of producing a good theory becomes orders of magnitude more complex than necessary. In the post-Gettier case, the problem had been amplified by the fact that 'justification' was also a problematic term that had no exact description.[139] Now the question is, how should we then characterise *knowledge* or

[136] Carnap (1950b)

[137] Williamson (2000) had two complaints about the analysis of knowledge that were essentially observations of overfitting in the post-Gettier project. First, the analyses had become excessively complex. Second, there had been too many unsuccessful attempts and it did not look like the characterisations were getting any better at avoiding counterexamples. Williamson's suspicions can be verified with the analogy on curve fitting. His first criticism is the observation that the curve is taking on too many parameters and hence becoming overly complex. The second is that given ordinary language use as our data, there will be always more outliers, no matter how complex the curve becomes. Furthermore, Williamson made the observation that the increasing parameters were making the characterisation of knowledge opaque, rather than informative.

[138] If we were to have finite and exhaustive data, then the most general theory would be the one that would fit all of the data, in which case complexity or variance wouldn't matter at all for the accuracy of a theory, since prediction and hence predictive error would no longer even be possible.

[139] The internalism/externalism debate (Steup, Turri & Sosa 2013; Pappas 2018), and multiple forms of justification (a comprehensive list of positions can be found in Pritchard & Bernecker 2011) is evident of the undetermined parameter of *justification* in defining *knowledge.* Going beyond the immediate bounds of traditional epistemology, as previously noted, we find even greater varieties of accounts for what epistemic justification is.

justification as they have been understood in epistemology? Or more generally, how should we answer ontological questions of the form ‘what is X?’

2.4.2 Descriptive Project and Curve Fitting

Let’s continue with the knowledge example and elaborate on it. Our goal is to provide a description $T_E(K)$ of knowledge such that $T_E(K)$ would elucidate the epistemological properties of knowledge. Let’s name the set of epistemological intuitions on knowledge $I(K)_E$, and the corresponding dataset of all instances of knowledge K_E. The relation between $I(K)_E$ and K_E is one of equivalence: proper instances of knowledge reflect our epistemological intuitions about knowledge, and our epistemological intuitions of knowledge demarcate which cases count as knowledge. i.e. They are evidentially equivalent. However, it’s easier to consider $T_E(K)$ as trying to fit K_E.[140] Now, K_E would span some *n*-dimensional space, where *n* is the number of epistemic parameters about knowledge, specified in $I(K)_E$. The parameters would include ‘truth’, ‘belief’, ‘justification’, ‘anti-luck’, ‘cognitive ability’, ‘counterfactual conditions on belief’ etc. i.e. Everything that philosophers have thus far considered in the literature on the analysis of knowledge. If we map K_E onto an n-dimensional space, we would get some clusters, where each datapoint in K_E would take on some linear combination of the *n* parameters. i.e. Each instance of knowledge would contain varying degrees of ‘truth’, ‘belief’, ‘justification’ etc. Some of the parameters would be much more pervasive than others. For example, we should expect ‘truth’ to be satisfied to a strong degree, and much more universally so than ‘counterfactual conditions on belief’. Our $I(K)_E$ would include the proposition that ‘a person *S* knows that *p only* if *p* is true’, but not that ‘*S* knows that *p* only if were *p* not true *S* would not believe it’, but that ‘*In some cases*, *S* knows that *p* only if were *p* not true *S* would not believe it’. For mathematical knowledge, the ‘counterfactual condition on belief’ might hold to a strong degree. In other cases, such as for

[140] As is standard in classification tasks.

perceptual knowledge, the parameter would fail to hold universally, and to a lesser degree.

The exact cluster formation of our dataset would depend on finer details of the individual parameters of each datapoint, which would require an extensive examination of the epistemological literature. That is in itself a gigantic project. So we won't try to guess what the cluster/s might look like. However, what we can be sure of is that there would not be one single cluster such that $T_E(K)$ would fit the cluster with both zero bias and low variance. i.e. What Williamson has already observed about the increasing complexity and opacity of our theories of knowledge. If we were to split our datasets into randomised, independent validation sets, any $T_E(K)$ that would fit perfectly our training set would fail miserably on cross validation tests. Since our available dataset is always a limited subset of K_E, this entails that the Strawsonian attitude would fail to produce a $T_E(K)$ that would fit K_E well, validating Williamson's suspicion on the impossibility of producing a robust $T_E(K)$.

However, this is also the turning point – here is where machine philosophy steps in. We start with splitting our available subset of K_E into a minimum of three independent datasets, and use one of them as the training set. Let's call the training set $K_{E\_training}$. We should derive our $T_E(K)$ solely based on $K_{E\_training}$. We do so by identifying a pattern with respect to the *n* parameters within $K_{E\_training}$, while applying feature selection to perhaps eliminate certain parameters so that we have less than *n* parameters remaining for $T_E(K)$. In this case, feature selection would play on two major components. First, we need to look at how large is *n* versus SIZE($K_{E\_training}$), if *n* is too large with respect to SIZE($K_{E\_training}$), then we risk having a model that simply would not generalise. We need to then practice the tricky business of deciding which parameters to remove. Fortunately for epistemologists, in the case of knowledge, there can be fairly stark differences between the salience of certain parameters, such as the aforementioned 'counterfactual condition on belief'. So one obvious choice that would render our dataset more uniform is to remove this parameter. Still, *n* could still be too large for the amount of data that we have, and we might wish to hold on to all of the remaining parameters. In this case, we need to do what scientists have done for centuries, and hypothesise boldly and progress from thereon.

However, even if we do have a sufficiently large $K_{E\_training}$, we might still want to remove parameters simply on the basis of uniformity, as was in the case of 'fish'. This decision would depend on the details of the data cluster. Now, if we are happy with the cluster that we have, we can proceed with producing a $T_E(K)$ that fits $K_{E\_training}$ fairly well, but is at the same time, not overtly complex. Then we can test how well $T_E(K)$ fits the other independent subsets of K_E. Epistemologists might wish to measure fit by examining how accurately $T_E(K)$ describes the data in K_E. There might be datapoint at the tail of the distribution, such as the case of TEMP mentioned in Chapter 1. If our $T_E(K)$ would contain a strong 'cognitive ability' parameter, then TEMP cases would be excluded from $T_E(K)$'s classification. But if $T_E(K)$ would contain a weak 'cognitive ability' parameter or if the parameter is eliminated, then $T_E(K)$ would be included in $T_E(K)$'s classification. This isn't a bug, but a feature of good statistical practice. TEMP cases indeed aren't that clear cut, and we needn't force $T_E(K)$ to remove it simply because we like 'cognitive ability'.

2.5 What Machine Philosophy Isn't

In this final section, I want to make two important clarifications.

First, I want to clarify that machine philosophy doesn't entail computational philosophy. Computational philosophy is the methodology of applying computational methods to philosophical problems. Examples include using agent based modelling (ABMs) to study moral norms or social epistemological problems.[141] These have indeed worked quite well. However, machine philosophy simply calls for

[141] Goldman (2010) had laid the groundwork for using ABMs to study social-epistemological problems by setting out a system for social-epistemological settings with the parameters: choices; agent; evidence; norms. Mayo-Wilson (2014) uses ABMs of research networks to study the reliability of testimonial norms in scientific communities. Zollman (2010) uses network structures in ABMs to study the social epistemological problem of communication. These studies have been done in NetLogo, a high level language specifically designed for building ABMs. See also Weisberg & Muldoon (2009); Mayo-Wilson, Zollman, & Danks (2011). In ethics, Skyrms (2003, 2010) laird the groundwork by building a game-theoretic framework to account for cooperation. We have examples such as Alexander (2007) and Muldoon & Hartmann (2014), who used ABMs to study how moral norms can evolve out of basic game-theoretic norms.

philosophical theorising to be governed by statistical norms, and provides a foundation for doing so. As a result, philosophical problems *could lend themselves* to computational solutions. However, machine philosophy does not entail that philosophers *should* use computational methods. After all, scientific theorising as we know it had been practiced even before computational methods became feasible.

Second, machine philosophy does not entail formal philosophy. This is analogous to machine philosophy's relation to computational philosophy. Machine philosophy simply claims that philosophical methodology should shift from a priori, boolean reasoning to statistical methods like in the sciences. This does not itself entail that philosophers should use only formal methods. Again, scientists doesn't exclusively engage in statistical modelling, even if their *reasoning* are ultimately grounded on statistical norms. However, machine philosophy does entail that philosophers *should* strive to be as statistically fine grained as feasible. This means that, whenever a more refined statistical method is available, the philosopher shouldn't settle on less refined methods. If it's possible for a particular philosophical question to be answered by a machine learning model, there is no reason for us to stop at the archaic procedure of linguistic or conceptual analysis. This does not meant that we cannot *start* with simple analysis, just that we should *also* use the more advanced methods *in addition*, in order to refine our theorising. Moreover, there is absolutely no reason to assume that we need a lot of data to do machine philosophy well – the intuitions that we have considered in the literature are perfectly fine.[142] This is neither a pleasantry nor an annoyance, it's merely what the empirical scientists have been doing for quite a while.

In philosophy, we have the luxury of generating large (and relatively clean[143]) datasets by virtue of creating hypothetical scenarios that are intuitive within the philosophical community, we can even assign statistical significance to them given how much consensus we have. For example, in post-Gettier epistemology,

[142] Moreover, we have a ML techniques 'Less Than One'-Shot Learning from Sucholutsky & Schonlau (2020). This technique is much more data efficient than previously thought possible for statistical learning. It allows statistical learning to effectively generalise on what we might think of as anecdotal data. This is good news for philosophers, as we'd rather sit in our armchair than do corpus studies.

[143] Since it is in fact *easier* to eliminate idiosyncrasies from a piece of intuition than from an observation. It is far cheaper to examine community consensus for intuitions than for observations.

epistemologists have came up with a variety of scenarios where a person can be said to satisfy a variety of epistemological parameters without having knowledge, and these intuitions can differ in subtle ways. The virtues of experimental philosophy is that it has been engaging in this kind of *data collecting* for our theoretical work.[144]

In the light of this, it's also easy to see how machine learning is not such a radical deviation from standard philosophical methodology, but merely a necessary evolution. What philosophers have been doing are essentially engaging in the first step of constructing hypothesis with high bias and low variance, then improving their hypotheses with intuitions in order to decrease the bias. What machine philosophy points out is that we should go a step further – balancing bias with variance in order to achieve generality.

[144] Although sometimes these surveys may miss the target if they are of intuitions from beyond the domain of philosophy. e.g. Weinberg et al. (2001) have collected the intuitions of non-philosophers regarding the Gettier-like cases, however, as I have argued in Chapter 1, intuitions are fundamentally domain-dependent, and the topic that is fixed by these intuitions are domain-dependent. It wouldn't make much sense *epistemologically* to look at these extra-philosophical intuitions, although, it would be sociologically and linguistically interesting. Still, the methodological direction of the experimental philosopher can be praised, though machine philosophy is in no way a defence of experimental philosophy. It merely welcomes experimental philosophy as an important *segment* of philosophical theorising.

Part II. Concerns about Machine Philosophy

Chapter 3 On the Issue of Changing the Topic

'A natural language is like a crude, primitive pocket knife, very useful for a hundred different purposes. But for certain specific purposes, special tools are more efficient, e.g. chisels, cutting machines, and finally the microtome. If we find that the pocket knife is too crude for a given purpose and creates defective products, we shall try to discover the cause of the failure, and then either use the knife more skilfully, or replace it for this special purpose by a more suitable tool, or even invent a new one. Strawson's thesis is like saying that by using a special tool we evade the problem of the correct use of the cruder tool. But would anyone criticise the bacteriologist for using a microtome, and assert that he is evading the problem of correctly using a pocket knife?'[145]

3.1 Introduction

In philosophy, methodological proposals that endorse revision have thus far shared a common concern: Given that a theory $T_D(X)$ of X in a domain D revises the pre-theoretic 'X', how can $T_D(X)$ answer typical philosophical questions about the pre-theoretic 'X' that are not in D? Let's call this the 'Strawsonian concern'.[146] For example, if I want to know what knowledge is, how can a theory that entails certainty (in the domain of formal epistemology) – which would exclude most of our pre-theoretic ascriptions of 'knowledge' – tell us what *knowledge* is?[147] It's crucial to note that this concern is salient precisely because philosophy, as traditionally conceived, has no specialised domain of enquiry, and instead *aims* at answering pre-theoretic questions.[148] Questions like 'what is beauty?', 'what is knowledge?', 'what is the

[145] Carnap (1963)

[146] Strawson (1963) first raised it against Carnap's project of explication.

[147] A more elaborate statement of the concern can be found in Cappelen (2018) Chapter 9, with ample illustrations. So I will jump straight to my diagnosis and treatment of the concern.

[148] Although historically this is partly the fault of Wittgenstein and certain early 20th century Oxford philosophers.

meaning of life?' etc. *in ordinary discourse* are to be answered by philosophy. Machine philosophy is a rejection of this archaic and untenable view of philosophy. Therefore, at the metaphilosophical level, the Strawsonian concern is simply illegitimate. Nonetheless, outright dismissing the concern would prove unsatisfactory for both sides of the camp. Advocates of revision have spilled much ink on answering the concern. More importantly, these treatments misrepresent the concern. This chapter therefore aims to do two things: to clarify and restate what the concern is about, and to answer technical questions regarding topic change.

As a preview, my diagnosis is that the Strawsonian concern is about a narrow conception of topic, wherein the topic of X in a domain D would be distinct from a topic of X in E. However, philosophers have consistently assumed that Strawson was considering a broad conception of topic wherein a topic is domain independent. Consequently, this has resulted in simply talking past Strawson. Carnap's own response has also been misunderstood and taken to be a weak reply when it was in fact straight on point.[149] The Strawsonian concern should be divided into two distinct but related issues: 1. How can a specialised theory $T_D(X)$ adequately describe the pre-theoretic 'X'? 2. How can $T_D(X)$ answer questions about the pre-theoretic 'X'? So the primary concern is with the issue of distinct domains rather than with changing extension or intensions. After all, I could change the extension or intension of a term such that its application is unaffected for all the typical philosophical problems.

Given my diagnosis of the Strawsonian concern, my treatment on the technical questions regarding topic change amounts to answering the two issues of the Strawsonian concern. The first question can be answered by simply building on the results from chapters 1 and 2. The second question has been inadvertently answered in part by a standard view that philosophical questions could be illegitimate, and thereby should be revised. This is an approach that partially denies the desideratum of topic continuity *in the sense of the narrow conception*. I shall provide an account of

[149] e.g. Maher (2007) argued that Carnap's response was inadequate, and placed explication as specifically for formal philosophy. I argue that this is a misunderstanding and perversion of the Carnap-Strawson exchange.

situations in which topic continuity fails. Furthermore, I disagree with proponents of this view that the resulting philosophical theories are normative. In this chapter, I provide three solutions to cover the second part of the Strawsonian concern, all within a framework of *descriptive* theorising. This set of solutions include the standard revisionist approach, which will be further divided into two situations wherein we either *create* or *change* the topic, and one which preserves topic in the form of approximation or granularity. Here is a preview of my solutions:

1. It could be the case that there was never a proper topic of X to begin with.
2. There was a properly delineated topic of X and $T_D(X)$ revises X, in this case the pre-theoretic X delineated a faulty topic for D. In situations 1 and 2, questions about 'X' are likewise faulty and need revision.
3. $T_D(X)$ can answer all questions of the pre-theoretic X, via understanding X as approximating $T_D(X)$ in context, or X as a granular expression of $T_D(X)$ in context.

In 4.1 I reject the standard diagnosis of the Strawsonian concern, in particular, I reject the desideratum of topic continuity. In 4.2, I present examples where the pre-theoretic term has no established topic, and proceed to provide an account of how a $T_D(X)$ establishes the topic for X. In 4.3, I argue that even if the theory changes the topic, it still can be descriptive in so far as the revision was necessitated by statistical criteria rather than ad-hoc ones. The biological revision of 'fish' is such an example. In 4.4, I introduce pseudo-revisions. I begin by examining the example of the mathematical definition of circle. In 4.5, I distinguish between idealisation and abstraction, and argue that these are distinct from revision qua the biological theory of fish over the pre-theoretic 'fish'. In 4.6 and 4.7, I provide two distinct accounts of how a term can be truly exemplified by entities that do not exactly satisfy the term. These play on the idea of 'being true enough', and serve as bridges between a $T_D(X)$ and the pre-theoretic 'X'. 4.6 introduces an account of *approximation*, where an entity can be said to be properly called 'X' *post-theoretically* just in case it *functions* just like X for all the relevant purposes of a practical context. However, 4.6 requires the rejection of the well-received T-schema. To overcome this, 4.7 introduces a more robust bridging principle of *granularity*, where an entity can be said to *exactly* satisfy a post-theoretic 'X' just in case the resolution of the practical context dictates that the entity just is X. The bridging principle applies to each of the above three situations in slightly different manners, but overall, they all serve to bridge the gap between an exact theoretical term and the pre-theoretic term.

3.2 What is a Topic?

3.2.1 The Strawsonian Concern: On the Issue of an Equivalence

There is a prima facie equivalence between *topic continuity* and *the ability to answer the same questions*. i.e. $T_D(X)$ preserves the topic of X iff $T_D(X)$ could answer all the 'typical philosophical problems' about X:

> 'If these things[150] are true, it follows that *typical philosophical problems about the concepts[151] used in non-scientific discourse* cannot be solved by laying down the rules of use of exact and fruitful concepts in science. To do this last is not to solve the typical philosophical problem, but to *change the subject*.'[152]

However, in the tradition of discussing this issue, the concern is often framed as being exclusively about $T_D(X)$ changing the topic of X.[153] The concern is then answered by appeal to various accounts of topic continuity while inadvertently

[150] By 'things', Strawson was referring to the fact that we cannot simply *replace* non-scientific terms with scientific terms. e.g. We cannot replace the pre-theoretic term 'knowledge' with a philosophical reconstruction thereof for the same questions that we have asked about the pre-theoretic 'knowledge'.

[151] While the debate between Strawson and Carnap centred on 'concepts', we can harmlessly replace 'concepts' with 'terms' following the terminology from earlier chapters. Again, here I agree with Cappelen (2018) and Leitgeb (forthcoming) on dismissing the focus on 'concepts'. But our agreement stops here.

[152] Strawson (1963). Emphasis mine.

[153] Strawson himself used the term 'subject' in Strawson (1963). Haslanger (2000) uses the term 'subject' in her discussion of the very same issue. Ludlow (2005) uses the term 'point' (in the sense of 'missing the point'). Cappelen (2018) uses the term 'topic'. I will stick with the term 'topic' since the term is less ambiguous than 'subject'. On the other hand, using this term opens another can of worms by crossing into the tradition that investigates the notion of 'topic' independent of metaphilosophical considerations, but in the domain of semantics e.g. Hawke (2017). The semantic issues are orthogonal to the issues of 'topic change' a la Strawson. I will argue for this below.

rejecting the equivalence, often by those who engage in revisionary projects.[154] They want to say that $T_D(X)$ preserves the topic of a pre-theoretic ‘X’ while denying that $T_D(X)$ should answer the typical philosophical questions about the pre-theoretic ‘X’. So they reject the desideratum of answering the typical philosophical problems about the pre-theoretic ‘X’ with $T_D(X)$. Revisionists sometimes go on to claim that some of the previously considered questions on the pre-theoretic ‘X’ are in fact illegitimate (because ‘X’ was ‘faulty’ in some sense). For example, a question of whether two women can marry each other is trivialised when ‘marriage’ is revised to include same-sex couples. Consequently, a revisionary project that preserves the topic of X is a normative project, in virtue of ‘improving’ on the pre-theoretic ‘X’.

I do not endorse this treatment for two reasons. First, in Chapter 2, I have argued that this ‘improving’ is a core feature of descriptive projects (such as in physics or in biology). It’s therefore an insufficient reason for categorising revisionary projects as ‘normative’. Following on from Chapter 2, we can say that an X is faulty if it cannot generate statistically robust models. For example, the old term ‘marriage’ was faulty when it was restricted to heterosexual couples. This is because a model for marriage would have to map over parameters such as ‘love’, ‘relationship’, ‘sexual attraction’, ‘legal status’ etc. So the old notion would make a statistically poor model if we take these parameters seriously, since, for example, there are plenty of non-heterosexual couples who love each other in the very same way that heterosexual married couples do. This is what people fundamentally mean when they say that a notion is ‘inconsistent’, ‘unfair’, ‘useless’ etc.[155]

Second, I think the rejecting of Strawson’s equivalence is itself a change of topic. As a result, conceptual engineers have been addressing an issue that was never raised (by Strawson). The next section will elaborate on why the notion of ‘topic’ in the

[154] e.g. Cark & Chalmers (1998); Chalmers (2012); Haslanger (2000); Cappelen (2018).

[155] e.g. Scharp (2007) argued that the reason for why we have the liar paradox was because ‘truth’ was an inconsistent concept. In the same way that old notions of ‘fish’ or ‘marriage’ cannot generate statistically robust models, the old notion of ‘truth’ had the liar paradox (since in the latter case the background inferences are logical rather than statistic). In Chapter 4, I will use ‘coherence’ to similarly refer to datasets that generate more robust models. Here I’m more so providing an improved theory of what an ‘inconsistent’ concept is rather than arguing against the notion.

Strawsonian concern isn't equivalent to the notion of 'topic' in the conceptual engineering[156] literature. For now, I argue that if we accept the conceptual engineer's understanding of 'topic' as aboutness or samesaying in a broad sense,[157] then the Strawsonian concern wouldn't be fundamentally about topic. The primary concern is whether $T_D(X)$ can contribute to providing answers to philosophical questions about the pre-theoretic 'X', not whether $T_D(X)$ preserves the topic of 'X'. To sacrifice the former in order to save the latter is to put the cart before the horse. Rather, the very point of wanting to talk about the same thing in the context of philosophical theorising is so that 1. our theory truly describes and 2. we could answer the same questions. Therefore, in the context of philosophical theorising, samesaying is satisfied just in case that points 1 and 2 are satisfied. The first point was answered in Chapter 2 – revision doesn't entail non-descriptivity, in fact revision contributes to descriptive accuracy via boosting generality. This chapter focuses on the second point.

There's nothing wrong with revising a question without worrying about whether we're still talking about the same thing. What matters is that the questions we ask are legitimate[158] and fruitful. It would be absurd to complain that scientists have not answered the question: 'Which kind of fish are warmblooded?' when they had revised the notion of 'fish', such that no fish would be warmblooded. What happened was that the scientists have pointed out that one of our questions about 'fish' was trivial, because our pre-theoretic notion of 'fish' was faulty/unfruitful (in the way that, for example, no robust *kind* of fish could be picked out). Even if we insist that the

[156] Also known in modern practice as 'rational reconstruction' e.g. Leitgeb(2020), 'explication' e.g. Maher (2007), 'amelioration' Haslanger (2000). The first two terminology follows Carnap (1934 & 1950b).

[157] 'Aboutness' reads from the philosophy of language literature in the tradition of Lewis (1988a, 1988b) . They are concerned with providing metaphysical accounts of 'aboutness'. The question here is 'what it means, metaphysically, to be about some X?' This is distinct from the notion of 'samesaying' discussed in the metaphilosophical literature re the Strawsonian concern. The latter asks 'under which conditions, regarding extensional and intensional change, is a topic of X preserved?' The two enquiries are orthogonal. e.g. A Lewisian account of 'topic' as being a set of propositions doesn't implicate whether the identity of this set changes if the extension changes. This is similar to our discussion in Chapter 1 on why metaphysical accounts of intuitions cannot help us to answer epistemological questions on using intuitions in philosophy.

[158] By 'legitimate' I mean dependent of. This is a weak notion of a 'legitimate question', but it's sufficient for our purposes here.

scientists have changed the topic, given some special use of the term ‘topic’, the reasonable response would simply be that ‘the previous topic was faulty, inconsistent, impractical etc.’, not that the change of topic *fails* to answer our initial questions. Whether a question is a good question comes *before* concerns about samesaying.

Before moving on, I should clarify what I mean by a ‘legitimate question’. The point for the Strawsonian concern is that $T_D(X)$ can contribute to answering a question *q* about X in D. For that end, we can define a ‘legitimate question of X in D’ as a question the answer of which *probabilistically depend* on $T_D(X)$. It’s important to note that dependence need not entail decidability. $T_D(X)$ needn’t decide *q*, but only affect it (makes it more or less probable). After all, most of philosophy is about the world, not about formal languages. I also do not count trivially true or false propositions as illegitimate. There is no point in excluding proposition like ‘all vixens are female foxes’ from the domain of biology, even if it’s trivial. We’ll come to see later the significance of why we need questions to be domain specific. For now, the point is to fix our use of the term ‘il/legitimate question’.

Since I maintain that ‘answering the same legitimate questions’ grounds topic continuity, I shall stay faithful to the Strawsonian equivalence, and use ‘topic of X’ in this chapter as a shorthand for ‘a particular set of legitimate questions about X’. Likewise, the question of ‘what is a topic’ can be treated as being equivalent to ‘what constitutes a set of legitimate questions about X?’[159] The next subsection reinforces and elaborates on this equivalence, and clarifies the Strawsonian concern by distinguishing a *narrow* conception of topic from a *broad* conception of topic.

[159] Of course, it grounds topic continuity jointly with descriptivity (truth/generality/predictive accuracy). But we have already established that descriptivity isn’t violated by revision. That of course doesn’t mean that revision cannot lead to a violation of descriptivity. In a case where a model has a high bias, the model would fail to be descriptive. But that is beyond the scope of the Strawsonian concern. No one doubts that such a drastic revision (that leads to high bias) changes the topic. The kind of revision under discussion are the likes of Haslinger’s ‘marriage’, Tarski’s ‘truth’, Chalmers’s ‘belief’, which are all examples of aiming to improve the robustness of the respective descriptions. So we can ignore the point about descriptivity in this chapter.

3.2.2 The Strawsonian Concern: Broad and Narrow Conceptions of Topic

Let's say that a 'narrow conception of topic' treats topics as domain dependent, and a 'broad conception of topic' treats topics as domain independent. They correspond to two distinct notions of 'topic', both of which have clear uses. Under the narrow conception, the topic of 'fish' in biology is distinct from the topic of 'fish' in culinary practice. When a biologist walks into the kitchen and tries to 'correct' the chef that what she prepares isn't fish, the biologist would be changing the topic, as Strawsonians would rightfully suspect. Under the broad conception, the topic of 'fish' in biology is the same as the topic of 'fish' in every other domain. A biological description of 'fish' and an ordinary English description of 'fish' are about the same topic. In the latter case, we can even compare how well the two descriptions fit the dataset of fish in their respective domains. In this case, as shown in the last chapter, the biological description is a more robust description. A mere change in extension and intension is insufficient for changing samesaying.[160] Now the question is, was Strawson complaining about a change of topic in the sense of a narrow conception or in the sense of a broad conception?

In the literature that discusses the Strawsonian concern, philosophers have consistently focused on the broad conception of topic.[161] I argue that this is a mistake. In fact, the Strawsonian concern cannot even get off the ground with a broad conception of topic. The very complaint was because Strawson wanted philosophical theories to respect ordinary language. Specifically, he claimed that philosophy shouldn't be a 'specialised' domain like the empirical sciences. So *by his own preaching*, he would have been well aware that when we use 'topic/subject' under the 'broad conception', *in ordinary discourse*, 'fish' in the culinary tradition and 'fish' in

[160] Cappelen (2018). But even without theoretical discussions, we know this to be true in ordinary discourse, which is what the Strawsonian concern is based on. Furthermore, note that samesaying holds within both the narrow and the broad conception. So this isn't about whether 'topic' is samesaying. i.e. Here 'samesaying' is as ambiguous as 'topic'. The biologist wouldn't be talking about the same thing as the chef in the first scenario, but they could be talking about the same thing that's known as 'fish' in a different scenario.

[161] See Maher (2007), Cappelen (2018)

biology would be about the same thing – the same topic. Instead, the Strawsonian concern is on domain shift:

‘For however much or little the constructionist technique is the right means of getting an idea into shape for use in the formal or empirical sciences, it seems prima facie evident that to *offer formal explanations of key terms of scientific theories* to one who seeks philosophical illumination of essential *concepts of non-scientific discourse*, is to do something utterly irrelevant – is a sheer misunderstanding, like offering a text-book on physiology to someone who says (with a sigh) that he wished he understood the workings of the human heart…’[162]

Despite the misleading example, the illustration of the failure for a physiology textbook to grant insight on the ‘workings of the human heart’ is a direct complaint of a domain mismatch – a failure of samesaying in the narrow sense. The example is misleading precisely because it *seems as if* Strawson was sketching a failure of samesaying *in the broad sense*.[163] However, this ambiguity can be easily resolved by simply reading Strawson in the proper context. Carnap’s explication (and his earlier ‘rational reconstruction’) was to work in tandem with his idea of a linguistic framework (what I call ‘domains’). A question within a domain D is only decidable with the language of D. So explication entails that a philosophical theory would sit within a specialised domain, with specialised uses.[164] Strawson was aware of this, and it would be strange for Strawson to *merely* complain about a change of extension rather than the much more jarring issue of a specialised philosophical theory. Indeed, Strawson’s explicit complaints were about the mismatch between scientific language being *specialised* vs. ordinary language having general uses: ‘The scientific uses of language, whether formal or empirical, are extremely highly specialised uses. [Ordinary] language has many other employments.’ For Strawson and his allies, the

[162] Strawson (1963), emphases mine.

[163] There was an equivocation between the literal use of ‘heart’ and a metaphorical use. I give more examples below that don’t suffer from this mismatch of reference. For the impatient reader, think of the mental state ‘knowledge’ as described by neurology vs. by epistemology; or, romantic love as described by chemistry vs. by psychology.

[164] This is also in line with Carnap’s own reply to Strawson about replacing the ‘pocket knife’ with the ‘microtome’ in Carnap (1963).

domain of philosophy is that of ordinary discourse. i.e. Philosophical questions derive their very meaning from ordinary language terms, not scientific terms. The concern about replacing an ordinary term ‘X’ with a specialised ‘X_D’ has not much to do with samesaying in the broad sense. After all, a neurological theory of pain still *is about* pain, but it cannot answer the philosophical question of ‘whether pain is physical’.

> ‘And it seems in general evident that the concepts used in non-scientific kinds of discourse could not literally be replaced by scientific concepts *serving just the same purposes*; that the language of science could not in this way supplant the language of the drawing-room, the kitchen, the law courts and the novel.’

Strawson was confident that a domain dependent theory could not sufficiently answer our typical philosophical questions. He was making the point that philosophy as a discipline shouldn’t produce theories in specialised domains. Of course Strawson wasn’t concerned about the fact that physical or chemical theories fail to answer some ordinary questions about stars or about water, that’s trivial. Strawson’s point was that the domain of enquiry for philosophers is that of ordinary discourse, rather than a specialised scientific domain. Strawson’s point was metaphilosophical, not technical (whether changing the extension changes the topic). Therefore it’s a mistake to get hang up on the issue of changing extensions and stray from the crux.[165] Since the concern is on answering questions about a pre-theoretic ‘X’, the corresponding notion of ‘topic’ should also be domain-dependent. So the narrow conception of topic is correct here.[166]

[165] In any case, if Strawson was really thinking of ‘topic’ in the broad sense, he would also agree with every conceptual engineer that changing extension does not entail a change of topic. That’s just how we use the term ‘topic’ in the broad sense. Strawson wasn’t stupid.

[166] This would explain why some philosophers believed that Carnap didn’t really address Strawson’s attack. It’s due to a misunderstanding of both Strawson and Carnap. Strawson complained that Carnap’s explication meant that philosophical theories would fail to answer typical philosophical questions. Carnap’s reply was that philosophical theories needn’t answer all of those questions. To go deeper, Carnap’s view on philosophy is that philosophy should be like the sciences – a theory should be specialised (domain dependent), and it should be fruitful in the domain. Carnap didn’t provide an explicit reason for this view, but it’s simple enough to fill in a plausible explanation – Carnap thought that domain independent theories are simply impossible. A strong piece of evidence to support this reasoning is Carnap’s idea of a linguistic framework. He argued that questions within (the domain/topic of) physics cannot be determined beyond the linguistic framework of physics. This would at least entail that only data, models, theories within physics (in the specialised language of physics) would be capable of answering questions in physics, whether those questions are about light, gravity, water.

In other words, philosophers have been looking at the wrong kind of dataset. In one sense it's obvious that the neurological theory of sadness and the psychological theory of sadness are about the same topic. However, also in the realm of ordinary discourse, it's paramount for students of psychology to understand that their topic of study on sadness is a *distinct topic* from that of neural science. It is this second sense of 'topic' that is of prime concern. A topic change or change of samesaying in the Strawsonian concern is therefore about a change of domain, not of extension/ intension. In principle, it's possible for both the situation that a theory changes the domain without changing the extension, and the situation that a theory changes the extension without a change of domain. So the two issues are also independent. The second situation by itself wouldn't be of concern for ordinary language philosophers, it is the first situation that challenges their metaphilosophical view.

This would also explain why philosophers believed that Carnap didn't really address Strawson's attack. It's due to a sheer misunderstanding of what they were even debating. Strawson complained that Carnap's explication meant that philosophical theories would fail to answer typical philosophical questions. Carnap's reply was that philosophical theories needn't answer all of those questions. To go deeper, Carnap's view on philosophy is that philosophy should be like the sciences – a theory should be specialised (domain dependent), and it should be fruitful in the domain (statistically robust). This metaphilosophical view was supported by Carnap's idea of a linguistic framework. He argued that questions within (the domain/topic of) physics cannot be determined beyond the linguistic framework of physics. This would at least entail that only data, models, theories within physics (in the specialised language of physics) would be capable of answering questions in physics, whether those questions are about light, gravity, water. The debate between Strawson and Carnap was precisely about *whether philosophical theorising should be specialised like in the empirical sciences*, not about whether a theory that changes the extension or intension of its data changes the topic.

Then what about revisions that don't appear to change the domain? For example, consider the revision of 'marriage' to include same-sex couples. Wouldn't

Strawsonians still complain? My answer is this: If the revision happens purely within ordinary language, just like how most of our words would change meaning over time, then it simply isn't philosophy. The question of whether such a revision changes the topic of 'marriage' wouldn't be of philosophical concern, but of lexicographic concern. In any case, the revision of 'marriage' has indeed been executed within a specialised domain. The particular description of marriage in question would be dependent on *ethical* concerns, and therefore would be within the domain of ethics. The fact that we could adopt a specialised description in ordinary discourse is besides the point. As I have already mentioned, samesaying can be either satisfied or denied depending on the conversational context. If one person uses the term 'marriage' or 'fish' derived from specialist concerns, and another person uses the terms in a pre-theoretic manner, then we still cannot say whether they're talking about the same thing unless we have further context that would make domain-dependence salient.

One might argue that revising a term upon propositions from a specialised domain D isn't sufficient for considering the revised term itself as belonging to D. In particular, ethical concerns are particularly common for the revision of many terms in ordinary discourse, or in other domains such as law or medicine. This objection relies on a weak supposition that most terms in ordinary discourse belong to no specialised domains. However, this supposition is clearly false. For example, common terms we use in ordinary discourse such as 'virus', 'engine', 'data' all belong to specialised domains (biology, engineering, statistics). Moreover, one of the most common arguments against the revision of 'marriage' itself relies on the insistence that 'marriage' is a specialised term within religion.[167] The complaint is essentially a Strawsonian move – by constructing an ethical or legal theory of 'marriage', one is

[167] They might, say, distinguish 'civil marriage' from 'religious marriage' (as celebrated in a church). Proponents of this particular argument would want to maintain only heterosexual marriage within churches while allowing homosexual civil marriages, treating them as distinct events. In this particular line of argument, there is also a change of about-ness, but that is not an essential feature, since the very point of the debate rests on 'marriage' being the same thing. The domain change itself is however a fact, and that is ok. Oftentimes we discover new features in particular domains, which leads us to change our views on the corresponding theory in another domain, and that is perfectly fine.

changing the topic from marriage as a religious notion. It's not about extension at all. The change of extension is an accident, a casualty, a side-effect.

To seal the deal, I argue that no questions can be asked without specifying a domain. In ordinary discourse, the context of discourse provides the domain. For example, even if I ask a simple question such as 'where is the school?', you could provide the correct location in terms of latitude and longitude, map coordinates, address, distance and heading etc. The example might look silly, but for philosophical questions, this is even more evident. In fact, it is precisely because philosophical questions are usually raised without context, that philosophers have the illusion that philosophical questions are asked as if they're still debating in ordinary discourse. Suppose I ask the question: 'what is love?'. I would of course be unsatisfied if you respond by giving me a chemical description of what happens in my brain when I fall in love. This domain mismatch was what Strawson had in mind. But that would really be my fault for failing to specify the domain. In fact, when I ask the question 'what is love?', I could be asking the question in the domain of behavioural psychology, or cognitive psychology, or sociology etc. I want to know what love is *in terms of* my behaviours, or in terms of my emotional states, or in terms of how two people who love each other interact etc. The answers needn't even differ in extension. But there is simply no way to answer the question 'what is love?' *without* a domain. Consequently, the change of topic for most terms we are concerned with cannot start with the domain of ordinary discourse, because the linguistic framework that governs ordinary discourse is just socio-linguistic rules.[168] It's merely a euphemism for 'I don't know for what I'm asking'. Rather, there must be a specialised topic for any legitimate question. This consequence is in line with our formal discussions in 2.2.3. When we talk about revision of a theory over data, we mean the two-step procedure of $R(T, X_D)$

[168] Unless, of course, you're explicitly interested in discussing grammar, or the use of terms etc. i.e. What children learn to do. Lexical entries are provided in the most commonly associated domains. 'Love' for example is defined as being a 'feeling of …', which would be a *psychological* description. However, the very same thing that we call 'love' can be described with a set of behavioural conditions – 'love is when a person…' etc. The domain of psychology pre-theoretically can also be called 'folk psychology'. In general, we can add 'folk' as a prefix to all the domains that we do engage with in ordinary discourse: 'folk physics', 'folk engineering' etc.

and $R(X_D, X_D)$. $R(T, X_D)$ represents idealisation, and $R(X_D, X_D)$ represents abstraction, or the process of feature selection.

Fortunately, the domain is usually specified in modern practice. Philosophers are already sensitive – even if only for practical reasons – to the fact that a question must be specialised. Consider the example of 'belief'. We can ask a question like 'Is true belief knowledge?' or 'What constitutes a rational belief?'. Both questions concern the same notion of belief. However, they belong to distinct domains, and to be answered by distinct investigations. The first question belongs to the domain of traditional epistemology.[169] It's to be answered by examining the data on valid cases of belief. The second could in the realm of rational choice theory or Bayesian epistemology. It could be answered by constructing a probabilistic model of belief where a belief can have varying degrees of strength, by examining the conditions for rationality, or even by looking at theories within economics or psychology.[170] In effect, we needn't concern ourselves with what a 'typical philosophical question' is. We need to only look at the *domain* for the X in question.[171]

Finally, I should clarify the key qualifier 'pre-theoretic' in this context. As I have argued in Chapters 1 and 2, when epistemologists talk about, say, their 'pre-theoretic conception of knowledge', they really mean the intuitions of knowledge *in* epistemology. So the qualifier of 'pre-theoretic' refers to the dataset of instances of X in D: X_D, corresponding to our intuitions about X. In other words, two experts from D and E starts with distinct datasets X_D and X_E. Within the broad conception of topic, we say that the topic of X includes both X_{D1} and X_{D2}, so that X_D and X_E are *partial* topics of X. That would entail that the theories $T_D(X)$ and $T_E(X)$ are both partial descriptions of X. $T_D(X)$ and $T_E(X)$ can also persevere samesaying. Under the narrow conception of topic, $T_D(X)$ and $T_E(X)$ are of distinct topics: X_D and X_E respectively.

[169] And traditional epistemology is considered by many to be simply a sub-domain of ordinary discourse. Anyone who took Gettier seriously in fact did so.

[170] Of course domains could overlap.

[171] Philosophers can sometimes engage in non-philosophical domains, and that is perfectly fine. Metaphysics and the philosophy of science overlap with theoretical physics. Formal epistemology overlaps with economics. Philosophy of mind overlaps with the cognitive sciences etc.

Suppose that D represents the domain of ordinary discourse, and E a specialised discourse, then the Strawsonian complaint is that philosophical theorising should be within D and D only. But as we have argued just then and in Chapter 2, $T_E(X)$ cannot possibly theorise over X_D. In particular, revision of the kind under discussion happens specifically via idealisation and abstraction: $R(T_E(X), X_E)$ and $R(X_E, X_E)$. Of course $T_E(X)$ can still *talk about the same thing* as $T_D(X)$ under the broad conception of samesaying, but that was never Strawson's point. However, 'pre-theoretic' for philosophers already refer to some X_D where D *isn't* the domain of ordinary English discourse. So *in practice*, Strawsonians have been averting from merely revising over already specialised X_D. Let's call this refinement the 'revised Strawsonian concern'.

So, the original Strawsonian concern (as Strawson himself stated) consists of the following two complaints:

S1. a theory $T_D(X)$ such that X_D is distinct from 'X' from ordinary discourse (OD) cannot adequately describe 'X' in OD, and S2. $T_D(X)$ cannot answer questions about the 'X' in OD.

In addition, here is an additional variation following from the above discussions:

SV1. a philosophical theory $T_D(X)$ that revises X_D cannot adequately describe X_D, and SV2. a $T_D(X)$ that revises X_D cannot answer all the questions about X_D.

I have refuted SV1 in Chapter 2: if our goal is to provide descriptively satisfactory theories, then $T_D(X)$ must revise over our X_D.[172] With regards to S1, we can then say that $T_D(X)$ is a *partial (domain specific) description of* 'X' in OD. The points of interest are S2 and SV2. S2 represents Strawson's complaint that a specialised philosophical theory cannot answer questions presented in ordinary discourse. SV2 represents the complaint that a theory that revises over its dataset

[172] Plenty of philosophers who do revision ignore the Strawsonian concern, but they haven't provided a detailed argument on why the Strawsonian concern is trivial (perhaps some don't think that it's trivial, but merely unimportant). The literature either ignores it or argues against it directly on false presuppositions. I hope that this chapter provides a satisfactory, explicit basis for philosophers to henceforth ignore the Strawsonian concern.
The actual concern for philosophers should be feature selection: which parameters I could ignore when theorising? That is a difficult problem and there isn't thus far a general solution from the statistical sciences. It's a case by case issue.

cannot answer all the questions that were raised about the dataset. For example, suppose that I eliminate the parameter of 'cognitive ability' for knowledge in the process of doing feature selection, then my resulting theory might be unable to answer questions on the relation between cognitive ability and knowledge. So although SV2 wasn't Strawson's own worry, it's related and perhaps more interesting. The rest of this chapter is devoted to solve S2 and SV2. Before moving on, I shall fix the notion of topic in question via a formal description.

3.2.3 Topic as Set of Propositions

Given that I use 'topic' as a shorthand for 'the set of legitimate questions', it would be natural to adopt the Lewisian formalism of treating a topic as a set of propositions in a logical space.[173] To start, we can define *a topic X* as *a set of legitimate philosophical questions about X*. However, we have just seen that topic is domain-dependent. There is no one set of legitimate philosophical questions about some X. Rather, there are multiple such sets, each associated with some philosophical domain. The topic of 'knowledge' in epistemology is associated with a set of questions distinct from the topic of 'knowledge' in information theory.[174] So we should modify the definition: A topic X *in a domain* D is a set of legitimate questions about 'X' *in D*. Or more succinctly: A topic of X_D is a set of legitimate questions about X_D. We have just circumvented the need to specify what counts as 'philosophical', because we have already restricted a topic of X to a specific domain. What matters is no longer whether a theory $T_D(X)$ is able to answer 'philosophical questions' (whatever that means), but whether $T_D(X)$ is able to answer legitimate questions of X_D.

[173] Lewis (1988a, 1988b). Here we needn't focus on the semantics of the Lewisian picture. As I have argued, the point isn't on what 'topic' (in the broad sense) is, nor can a semantics of 'topic' affect our argument, since Strawson had specified what he meant. It would be a change the topic to focus on the semantics of 'topic' when Strawson is talking about something very specific when he used the term 'topic'.

[174] Of course there could be overlaps.

We can now specify three kinds of topic change within the Strawsonian concern. Suppose we have X_O representing instances of 'X' in ordinary discourse, $Q_O(X)$ representing the set of legitimate questions on X_O, and $Q_D(X)$ representing the set of legitimate questions on X_D *before* we revise X_D over itself, i.e. $R(X_D, X_D)$, and $Q_{R\_D}(X)$ the set of legitimate questions on X_D *after* we do feature selection on X_D, and $T_D(X)$ a theory that fits over the revised X_D. The first kind of topic change (TC1) corresponds to concern S2:

TC0. $T_D(X)$ changes the topic of X where 'the topic of X' mean $Q_O(X)$.

The second kind of topic change (TC2) corresponds also to S2:

TC1. $T_D(X)$ changes the topic of X where 'the topic of X' we mean $Q_D(X)$.

The third kind of topic change (TC3) corresponds to SV2:

TC2. $T_D(X)$ changes the topic of X where 'the topic of X' mean $Q_{R\_D}(X)$.

First, we need to see that TC0 and TC1 are equivalent. It matters not whether we think of $T_D(X)$ as revising over X_O or X_D before $R(X_D, X_D)$. To see this, consider the following response to TC0:

For TC0, the questions concerned are ones with an ambiguous domain. For example: 'what it is to say that one thing is conditional upon another?'[175] We can replace such questions with domain specific ones like 'under which circumstances in reasoning do the material conditional apply?' etc.[176] It would be absurd to complain that by looking at distinct kinds of conditionals in different formal systems, we'd be evading the question of 'what it is to say that one thing is conditional upon another?'. Rather, the sensible response is that we are answering the question by dissecting it into specialised sub-questions, as empirical scientists do. In general, answers to specialised questions contribute to answering general domain-unspecific questions. Sometimes this happens by dissecting an ambiguous question into multiple questions, such as in the above example on conditionals. At other times, if the question isn't

[175] This was Strawson's own example from Strawson (1963)

[176] This would be in the domain of philosophy of logic, or more specifically, the normativity of logic for human reasoning. For example, the relevant literature has several examples of the applicability of different logics. e.g. Stenning & van Lambalgen (2008) This is beside the point of whether logical pluralism holds.

ambiguous, we simply need to explicitly specify the domain, such as asking whether a fish is warmblooded. If I ask a question: Is X *f*? without specifying a domain, then given that *f* belongs to features described by $T_D(X)$ for some D in which *f* is a member, $T_D(X)$ would be able to at least answer this particular question of X in D-specific parameters. If the concern of TC0 is for specialised theories to answer specific *philosophical* questions which are asked in ordinary discourse, then those philosophical questions should also belong to some $Q_D(X)$ where D is a philosophical domain. If in addition, we assume that $T_D(X)$ doesn't eliminate certain parameters from X_D such that certain questions of interest from $Q_D(X)$ are eliminated, then $T_D(X)$ would be able to answer the original questions from $Q_O(X)$.

However, we already know that the assumption I made in this initial response to TC0 is false. Here the issue isn't exclusively about $T_D(X)$ revising over X_D, but the fact that $T_D(X)$ makes illegitimate[177] certain questions in $Q_D(X)$ after an intertwining two-step procedure of curve fitting. The first step is $R(X_D, X_D)$: feature selection before fitting $T_D(X)$ onto $X_{D.}$ The second step is $R(T, X_D)$: fitting $T_D(X)$ onto X_D to balance bias against variance. I assumed that X_D did not eliminate D-specific parameters with feature selection. As we have seen from the 'fish' example, this assumption fails. The very point of feature selection is to eliminate D-specific parameters in X_D in order to reduce the dimensionally of our dataset. One of the very issues philosophers have is that we are terrible at feature selection, and this is also the key point of the Strawsonian concern. The worry about TC0 isn't on the questions that are in $Q_O(X)$ but not in $Q_D(X)$. Epistemology isn't concerned with psychological questions about knowledge. Rather, The worry about TC0 is that $T_D(X)$ doesn't account for questions that are shared by $Q_O(X)$ and $Q_D(X)$. This is equivalent to the worry about TC1. Since TC1 is more concise, I shall henceforth consider TC1 as the concern for S2.

Now, TC2 isn't usually understood as an issue of 'topic change', even by Strawsonians. Rather, it's an issue of 'what counts as a good theory' or 'which

[177] trivialises, falsifies, or simply make undecidable

questions are important', of which the revision of X_D is one aspect. So it's a red herring to worry about topic change *simply because* one is doing revision. By Strawson's own account, TC2 can be understood as his complaint that we cannot replace imprecise ordinary language terms with precise scientific terms. I shall provide two candidates for bridging imprecise ordinary discourse terms with precise scientific terms. But first, I shall respond to TC1.

3.3 On the Notion of a Faulty Topic

To recap, the worry about TC1 is: $T_D(X)$ fails to answer certain questions from $Q_D(X)$ before X_D is revised to exclude certain D-specific parameters. Feature selection doesn't entail that a removed parameter *f* is itself removed from $Q_D(X)$. The best parameters to remove are actually ones that obviously hold for every member of X_D. In this case, other previously salient parameters could become trivialised or falsified. The biological theory of 'fish' is such an example. By excluding the parameter of 'living in water', biologists make salient the biological differences between the marine animals we call 'fish', such that certain other biological parameters such as 'being warm blooded' or 'being able to breathe in water' become falsified or trivialised via a statistically robust classification. As a result, certain questions are answered (such as 'whether all fish are warmblooded') and other are made illegitimate (such as 'which fish are warmblooded?').

Now, Strawsonians are welcome to insist that this kind of revision constitutes a change of topic. However, I have a better story than simply calling $T_D(X)$ a change of topic for X. This is to say that pre-$T_D(X)$, the topic $Q_D(X)$ was itself faulty.[178]

In Chapter 2 we argued that revision via the two interdependent procedures of abstraction and idealisation are essential for $T_D(X)$ to *truly* describe X_D. So $T_D(X)$ is our best bet at answering questions about X_D. Oftentimes, the procedures would involve revising X_D itself via feature selection. This results in a revision of also

[178] Remember, we're not using 'topic' as samesaying anymore. If we consider also the condition of 'adequately describing X', then of course there wouldn't be a change of topic, since to adequately describe X, such revisions are not only harmless, but necessary. But that clearly isn't the Strawsonian concern.

$Q_D(X)$. By modus tollens, if we insist on an unrevised $Q_D(X)$, then we must also give up providing a true $T_D(X)$. So we cannot have both a descriptively adequate $T_D(X)$ *and* an unrevised $Q_D(X)$.

Now, it follows that there are questions within an unrevised $Q_D(X)$ that cannot be answered by a descriptively adequate $T_D(X)$. Let $S_D(X)$ be the set of statements corresponding to positive answers for $Q_D(X)$, and *s* as belong to $S_D(X)$. Suppose that $T_D(X)$ has no bearing on whether *s* is true. This means that two statements of X *in D-specific terms* are epistemically independent of each other.[179] It follows that *s* and $T_D(X)$ are about different topics under the narrow conception of topic. Since $T_D(X)$ is descriptively adequate of X_D, *s* cannot belong to the topic of X_D.[180] We can then say that the unrevised $Q_D(X)$ was a faulty characterisation of the topic of X_D.

Finally, let's return to our example of 'fish'. Consider another question: 'Which fish are warmblooded?' This is a question within the domain of biology.[181] One corresponding statement might be 'A whale is a type of warmblooded fish.' However, the biological description of 'fish' trivially falsifies the statement.[182] We say that

[179] This has nothing to do with Gödel's Incompleteness theorems. Incompleteness tells us that there are proposition in a formal language that are undecidable within the language. That has nothing to do with whether those undecidable propositions can still be corroborated. The Axiom of Choice is undecidable within set theory, however, it is still deemed as true since it does contribute towards proving many theorems that are intuitively true in mathematics. i.e. the Axiom of Choice is statistically corroborated by other true statements in mathematics, and it isn't *independent* from them. This is why I used dependence rather than decidability to determine the legitimacy of a question.

[180] i.e. $T_D(X)$ answers a distinct set of questions from $M_D(X)$ *as a result of* more accurately describing X_D.

[181] Note that 'warmblooded' is no longer a term used scientifically, since it is a rather unfruitful category in the light of the various kinds of thermoregulation between species. In other words, 'warmblooded vs. 'coldblooded' is too crude or coarse-grained for scientific purposes. Nonetheless, it's clear that no fish are warmblooded, so we can stick to the term without any issues. Note that this doesn't meant that 'warmblooded' is no longer a term in the domain of biology. The term still derives its very meaning from within the domain of biology, and there is no other domain in which the term could belong. The term simply becomes 'informal' as opposed to being 'formal' or 'technical', but that has nothing to do with its domain. For illustration, consider the example of the term 'aether' in physics. Even though scientists have discovered that there is no evidence that 'aether' refers to anything, the term doesn't suddenly get booted out of physics. It is still a term that derives its very meaning from physics. It just happens so to be an empty term.

[182] I didn't include 'trivialised' statements as part of being 'illegitimate', since they are technically answered by the theory. However, note that in this case, the statement is trivialised because of the new definition simply excluding 'whales' from its classification. This shows that revising a term doesn't simply entail that certain old questions cannot be answered. Moreover, it reinforces my point that Strawson's complain about answering questions about the 'heart' with a 'physiological textbook' is about domain change, not extension or intention.

biologists, in providing a descriptively adequate theory of fish, have not changed the topic of fish in biology, but salvaged it, or discovered it from a previously faulty topic. We can also say that the biologists created the topic of fish in biology where there was none. The reader is free to pick either characterisation. Either way, the Strawsonian worry about TC1 is dissolved.

3.4 On Bridging Imprecise and Precise Languages

A key Strawsonian concern is that we cannot replace imprecise ordinary language terms with precise scientific terms. I argue that while this is true under the narrow conception of topic, it's untrue under the broad conception. Moreover, specialised theories do ultimately aim to describe that which we already talk about in ordinary discourse. This point is important, and it needs a story. In the remaining of this chapter, I provide two accounts that bridge the gap between imprecise and precise terms. In 3.5, I provide an account of *approximation*, and in 3.6 I provide an account of *resolution shifting*. These accounts are not meant to be mutually exclusive nor are they meant to exhaust all the ways that we might bridge a precise scientific term with an ordinary language term. They are meant to serve two purposes. One, as ways to dispel the Strawsonian concern. Two, as a methodological tool for philosophers to theorise without worrying about making imprecise terms precise (a la Carnap).

3.4.1 The Example of the Circle

We already bridge imprecise and precise terms in mathematics, philosophy, the empirical sciences. For example, mathematical definitions of concepts are often understood as ideals towards which ordinary entities that are ascribed as exemplifying such descriptions approximate. Consider the mathematical description of the circle $T_M(C)$:

> A circle is the set of points in a plane that are equidistant from a given point
> O. The distance r from the centre is called the radius, and the point O is called

> the centre. Twice the radius is known as the diameter d = 2r. The angle a circle subtends from its centre is a full angle, equal to 360° or 2π radians.[183]

Folks often speak of circles, and ascribe circularity to ordinary physical objects. The mathematically trained would be well aware that nothing in the physical world fully exemplifies $T_M(C)$, since physical units are discrete. If not all, most cases of ordinary ascriptions of circularity fail to satisfy $T_M(C)$. Nonetheless, our mathematically trained would not forbid the ordinary folks from using the word 'circle', nor would they themselves refrain from using the word 'circle' in ordinary contexts. The reason is that it's proper to use the language in this way. The reason why it's proper is because in such contexts, the utterances of the word 'circle' denote entities that *approximately* satisfy $T_M(C)$ for all relevant practical purposes.

A Carnapian might object: Shouldn't one distinguish between 'circle' as a *mathematical* term and 'circle' as a *physical* term? Accordingly, there are (at least) two kinds of spaces with different kinds of geometries: mathematical spaces with mathematical geometries, and physical space with physical geometry; the former are studied in pure mathematics on a priori grounds, the latter is studied in physics on a posteriori grounds.[184] Understood in this way, 'circle' as a mathematical term applies to mathematical entities, while the 'circle' as a physical term applies to empirical entities.

My reply: First, 'circle' as a physical term just is $T_M(C)$. Physicists adopt the mathematical definition of a circle (and other mathematically defined objects) for their physical descriptions. There is no distinct study of 'circle' in physics. This is an example of an overlap of domains I mentioned in earlier chapters. Physics overlaps with mathematics in so far as physical descriptions are made in mathematical terms. Now, the Carnapian might deviate from Carnap and forgo the remarks on physics to opt for a modified objection: that by 'physical' she meant 'the physical world' rather than 'physics'. In this case, the suggested distinction is between $T_M(C)$ and $T_O(C)$, the

[183] Weisstein, Eric W.

[184] e.g. In Carnap's discussion of this in his doctoral dissertation "Der Raum".

latter representing an ordinary language description of ‘circle’ as ‘a perfectly round plane figure’[185]. We can all agree, whichever camp you’re in, that $T_M(C)$ and $T_O(C)$ talks about the same thing – ‘circle’ under the broad conception of samesaying. i.e. What the mathematician (or geometrician) studies just is the plane figure that ordinary folks would refer to as ‘circle’. After all, the studying of circle in mathematics has been motivated by their application in the real world. The diehard Carnapian might now point to my own assessment and claim that $T_M(C)$ and $T_O(C)$ are *not* about the same thing under the narrow conception of samesaying. But that’s just a repetition of Strawson. The second part of the concern was that we cannot replace $T_O(C)$ with $T_M(C)$ since $T_M(C)$ is from a specialised domain. I’m pointing out that this is beside the point. First, I have argued in the last section that we *can* answer legitimate mathematical questions regarding $T_O(C)$ with $T_M(C)$. Moreover, the issue here is now on accounting for the *continuity* of topic (the broad conception) between $T_M(C)$ and $T_O(C)$: how can $T_M(C)$ describe C_O despite $T_M(C)$ being within a specialised domain?

Another commentator might object that we can and perhaps should be more careful by instead uttering that ‘the surface of this tabletop is *approximately* circular.’ In most contexts, it would be outright misleading to say the sentence ‘this tabletop is *approximately* circular.’ When one enquires whether something is circular, the enquirer generally has in mind ‘circular for a particular purpose’, and so it would be misleading to respond to an enquirer that a tabletop is approximately circular, even if that’s true. Anyhow, even if we do use language in this hypothetical fashion, approximation still relates the ordinary ascription with the concept for all relevant pragmatic purposes, albeit in a more explicit manner. Furthermore, if we think of a ‘point’ as having a low enough resolution, like pixels on an old screen, then we might even have a circle that fully satisfies $T_M(C)$.

[185] "circle, n." *OED Online*, Oxford University Press, March 2021, oed.com/view/Entry/33187. Accessed 22 April 2021.

3.4.2 Important Observations

I want to make three clarifications following my discussion of the 'circle' example. First, it seems that I have been endorsing a kind of Platonism regarding mathematical entities. However, that assumption would be unwarranted. Although there is a resemblance to Platonism in the sense that ordinary entities approximate ideal Platonic Forms, there need be no such Platonic Forms for an object to approximate being a circle. The point here is purely formal and linguistic – under which conditions can we properly say that something is 'X'?

Second, there is no one single bridging relation for every term, and the measure of how well something satisfies a description doesn't depend only on the description. In different contexts, the requirements for a particular entity to satisfy the target description can differ, and in a variety of ways. Consider a complex description $T_D(X)$ with a certain number of individual parameters $F_D(X)$. First, the minimum number of parameters in $F_D(X)$ that an entity must satisfy in order for it to be properly ascribed 'X' can differ depending on the context. Second, the weighing of the distinct parameters in $F_D(X)$ can differ due to different contextual desiderata. So, to measure how well something satisfies $T_D(X)$, the function must be binary, with both the bridging relation and the practical context as arguments. Here we understand context as a set of practical desiderata. The desiderata might contain, for example, margins of error. The context mapping is in turn a weight function that maps $T_D(X)$ to a vector, each element of the vector corresponding to a distinct parameter in $F_D(X)$. The weighing would depend on how well each parameter satisfies each of the practical desideratum. What the weighing function does is looking at the parameters that the context wants, look at how much of each parameter it wants, and outputs a vector of all the wanted parameters in $F_D(X)$, each with an associated margin of error or degree of importance. We can then compare this with our actual referent, and see how well the referent matches the output. This comparison yields a degree of how well the thing we call 'X' satisfied $T_D(X)$.

I'll illustrate this point of context dependence with a single parameter for simplicity. For example, suppose that we have agreed to meet at 8:00 p.m. in front of the opera house, and suppose that my margin of error for 'being on time' in this context is ±1 minute. In another context, where we have agreed to a dinner reservation, the margin of error might be ±10 minutes. If we have agreed to catch a train in Japan, it might be ±2 seconds. In chess, the margin goes down to ±10 milliseconds. Suppose that we have a description of 'being on time' as 'arriving at location *l* exactly at time *t*'. Now, even the location could have margins of error (in kilometres, meters, millimetres etc.) If we take locations as points in space and time as measured in the smallest measurable unit – zeptoseconds, then we'd never 'be on time', and every utterance of the phrase would be a falsehood. Of course, that's not how we use language, even taking the phrase 'being exactly on time' literally. We have two ways of accounting for this discrepancy between exact descriptions and what we are referring to, or, between the context dependence ascriptions and context independent descriptions.[186] One is to say that what we said is technically false, but 'true enough'. So if I arrive 30 seconds late to the opera house, then I'm not on time, but I am *on time enough* such that we can *say* that I am 'on time'. Another way is to say that I am indeed on time, if we measure the smallest units of time in the opera scenario as being 1 minute, so that 30 seconds isn't even measurable. These two ways to bridge the apparent discrepancy corresponds to *approximation* and *changing resolution*.

In general, there can be no single measure with which we can measure how well an instance satisfies a description, even for the very same pair of description and

[186] So far, the story is just another example of what epistemic contextualists have noticed (e.g. DeRose (1992); Lewis (1996)). Hereon though, the issue diverges from that of the contextualist. In one respect, my bridging relations are ways to bridge contextualist accounts of knowledge ascription with context independent theories of knowledge, rather than having to make the terribly problematic claim that *knowledge* itself is a context dependent entity (which won't work since we'd have to make almost every theory a context dependent one). Furthermore, the bridging relations allow us to maintain topic continuity between exact descriptions and inexact language use, which is what Carnap and his followers wanted. The point here, again, is to use this as a new bargain to buy over the Strawsonians.

entity.[187] Furthermore, formulating a concrete measure and a context weight function for a particular case is by large very difficult, even for simple descriptions with few simple parameters. There is often no precise measure with which we can work out, say, how circular something is with regard to a particular set of purposes, since it would require you knowing exactly what you want in every situation. However, we do have ways of judging correctly whether something will fulfil an exact description (if there is one[188]) closely enough for the relevant purposes, ignoring cognitive defects. Here we have hit second-order approximation, since perhaps for different purposes, scales with different resolutions for the measure's codomain are adequate for different contexts. To illustrate, suppose for a particular context, description $T_D(X)$ and bridging relation, we have a measure of how well something satisfies $T_D(X)$ with a codomain of real-values between 0 and 1. However, in this context, we don't need such a fine-tuned measure, for whatever reason. Thus, we could work with an adequate *approximation/lower resolution of the measure* — perhaps a measure with only rational outputs, or outputs that range over $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

Two further observations follow. First, we don't apply bridging only at the first-order between descriptions and entities, but also at the second-order between our judgements of how well something satisfies a description with the minimal degree of satisfaction for a given purpose, although we may not know what this minimum is. Second, bridging appears to be needed at the third-order, to answer the question of when the bridging of the measure is adequate. In other words, we need a bridging of the bridging of the measure that measures first-order bridging. This may begin to sound like an infinite regress. However, the nature of this regress is such that as the order increases, it becomes easier to identify the salient features we want. Thus, eventually, it would be obvious enough in practice to make a sound judgement,

[187] Of course, one may argue that the same entity cannot exist in two different contexts. The point is however to emphasise that even for indiscernible entities, there is no single measure to measure how well it satisfies a target description across different contexts.

[188] Trivially, when there is no definition or theory for a description, then there is nothing to relate with the ordinary language term. The task for the philosopher is then to first construct a definition.

perhaps at the third or even the second order. There would be no infinite regress after all.

Third, one might notice that I did not attribute a measure of degrees to circularity and flatness, but instead, attributed it to the adequacy of approximation. Since I argued against boolean representations, so it might look as if parameters already come in degrees within my view. For example, Steven Hetherington argues that knowledge permit degrees in order to dissolve the Gettier problem.[189] However, that's besides the point. The point isn't whether a description consists of discrete or continuous parameters. It's on the issue of *given a description* $T_D(X)$ fitted onto some X_D, what do we make of entities that we call 'X' but don't fully satisfy $T_D(X)$. Saying that $T_D(X)$ comes in degrees cannot help with this question of proper ascription.

One might now wonder about relations. We can of course talk about the relations of 'being redder' or 'more circular'. However, relations themselves do not come in degrees. Consider the example of the relation less-than in arithmetic, denoted by '<'. In the natural numbers system, $1 < 2$ and $1 < 3$. One might say that '1 is less than 2 to a lesser degree than it is less than 3'. What this 'lesser degree' means just is that $(2 - 1) < (3 -1)$; i.e. the difference between 1 and 2 is *less-than* the difference between 1 and 3. This has nothing to do with the meaning of 'less-than' or '<'. In this example, talk of 'degrees' is simply a misleading way to express the relation less-than at the second order. Similarly, if we say that some e_1 is 'more circular' than some e_2, we simply mean that the degrees to which e_1 satisfies the description of circle $T_M(C)$ is *less than* the degrees to which e_2 satisfies $T_M(C)$. Remember, $T_M(C)$ itself is fixed and context independent. There is however a genuine measurement of degree involved in both examples. In the arithmetic example, the measurement of 'how-much-less-than' would be the differences between the values of $(2 - 1)$ and $(3 - 1)$, which is 1. However, this measurement is never expressed nor contained in the relation less-than, it's in the *measurement* of *how-much-less-than,* which is a quite distinct notion from less-than: the former is the binary function of subtraction; the latter is the binary

[189] Hetherington (2005, 2011); Gettier (1963)

relation of less-than. In general, anything that generates a measurement of degrees must be a map that yields a value. A good example of a measure that already exists in philosophy is one that could be used also for approximating belief, usually designated by ‘credence’ or ‘degrees of belief’. Whether a particular instance of credence in a proposition *p* is an adequate approximation to warrant the utterance of ‘believing that *p*’ is a contextual matter. Furthermore, applying rational choice theory on top of utility measures can provide the criteria for when a measure (in this case, credence) is adequate (for being a belief).

Granted all that, an important question remains: ‘Why does an entity approximate a particular description rather than a competing one?’ For example: ‘Why does the surface of this canvas approximate the description of being flat rather than being bumpy?’ Of course, for fine liners, the canvas would be bumpy; for an atom, it’s like the Swiss Alps; but for a dinosaur, the canvas is as flat as Australia. The answer to the above question is that the entity approximates a particular description because it acts like that description more than any other competing descriptions for a particular practical context. If we try to draw with a fine liner, then the canvas isn’t flat enough. If we paint with acrylic, then the canvas would be perfectly flat. The meaning of ‘flat’ didn’t change, the criteria for its proper ascription changed. What about for compatible descriptions such as *being flat* and *being rectangular*? Well, both could be approximated by the same entity in the same context, nothing mysterious about that — ‘This canvas is flat and rectangular.’ In other words, bridging relations not only relate an entity and a description for a particular context because the entity functions sufficiently like something that fully exemplifies the description in the context, it relates them because no other incompatible descriptions are better exemplified by the entity for the given context. Of course, this latter negative condition is not by itself sufficient, or else we would have many vacuously true claims such as ‘Peter is a philosopher.’, since being a philosopher is not incompatible with any other entities that Peter manifests.

3.5 On the Approximation Relation

Approximation relates a precise description $T_D(X)$ and an ordinary referent that we call 'X'. For example, the description of true sentences as per Tarski is approximated by ordinary instances where sentences or propositions are said to be true. Confused debates arise in all kinds of enquiries, from the analysis of knowledge in traditional epistemology to concerns about thought experiments, due to an inadequate understanding of the role of approximation in such theories or methods. Approximation is important because it explains and reconciles the relation and apparent discord between theoretical domain-specific descriptions, models etc., which are precise, and their imprecise ordinary language counterparts. In this section, I aim to provide an account of approximation. The criteria for approximation are essentially pragmatic, and the measure for measuring the adequacy of an approximating entity would thus contextually vary. Therefore one should not expect to find a universal measure that would be applicable for all scenarios. Furthermore, it seems to be very difficult to find a precise measure for any given case, especially if the theoretical description is complex. Regardless, folks often have no problems with categorising entities; this is because approximation is applicable to itself, and at a higher order, it bears a low enough resolution for ordinary folks to discriminate between whether an entity approximates a theoretic description adequately enough for all relevant pragmatic purposes.[190]

I shall now piece together an account for approximation. To have a full treatment, we must answer two questions. The first is: Under what conditions would *x* satisfactorily approximate *y*? The second question is: Under what conditions may one properly judge *x* to bear *y*? The reason for this divide was discussed in the previous paragraph — although folks generally judge correctly whether an *x* approximates a *y* in a given context, their method of doing so is essentially a higher-order approximation of a measure on approximation. Answering the first question gives us a

[190] This doesn't mean that the folk is aware that she's doing approximation nor does it mean that she's aware of any theory. It just means that she can competently apply terms correctly, given a theoretic description holding true in the community in which she communicates.

real-valued measure with which there is a contextual minimum in the range for *x* to satisfactorily approximate *y*. This is the part that provides the account for approximation. The second part is an application of approximation to itself, and would need to draw upon decision theoretic tools for a proper answer, as mentioned earlier in the brief interlude on credences. This chapter attempts only to answer the first question, and it should provide a frame with which we may answer the second question.

Here is a preliminary account for the relation of approximation:

(Approximation) Given a precise description/theory/model $T_D(X)$*;* an ordinary entity *e;* a hypothetical entity e_T that fully exemplifies T_D[191]; the set ~X consisting of all the rival descriptions that are incompatible with T_D in that no entity can bear both T_D and something from ~X (the 'rival' qualifier is required for relevance — so that the description of a circle is inconsistent with that of a square, but not with that of a sphere since the latter applies to a different domain); the set of entities ~XE that fully exemplifies the descriptions in ~X; a practical context *C* that is defined as being a set of duples, each duple consisting of a practical end and a weight for that end, with all of the weights summing to 1*:*

e sufficiently approximates T_D in *C* iff *e* functions in *C* adequately similar to how e_T would function in *C*, and operates more similarly to how e_T would operate in *C* than it would with any entity in ~XE.

This sketch is fine for a first pass, however, it doesn't tell us exactly what it means for *e* to operate adequately similarly to an entity e_T in *C*. Elaboration is needed, and thus I advance the following four criteria for *S* to sufficiently approximate *C* in *P*:

1.1 $W_C(T_D)$ is a weight function that maps the singleton partitions[192] of T_D onto a labeled vector.[193] Each singleton represents a D-specific parameter of

[191] I henceforth abbreviate $T_D(X)$ as T_D.

[192] A 'singleton partitions' means T_D is partitioned into singletons, not that the partition is a singleton.

[193] One could equally take S_c as the argument instead of *C,* since S_c fully exemplifies *C* and thus it has all the partitions of *C,* and extra ones since S_c might take on other properties such as existing in the physical world, to take a trivial example. In the case of using S_c as the argument, simply leave out all the extra partitions of S_c not contained in *C*. i.e. We take only the relevant partitions of any such *S*.

T_D. The values of the vectors add up to 1 and their values are assigned according to the rules for weighing partitions in a pragmatic context, which will be specified below. In other words, $W_C(T_D)$ assign weights to the parameters of T_D depending on how much utility each parameter of T_D contributes to the practical ends of *C*. If T_D is already simple (is itself a singleton), then $W_C(T_D) = 1$. Making the values of $W_C(T_D)$ add up to 1 is to scale everything in relation to T_D, so that the efficacy of *e* in *C* could be measured as a percentage of T_D.

For example, if T_D is the concept of the circle, then T_D is a singleton. It is partitioned into the element: 'the set of points in a plane that are equidistant from a given point *O*'. Then $W_C(T_D) = 1$ since there are no other simple constituents in T_D. For entities that approximate T_D, we take the cross-sections and already abstract the entities away from their other physical parameters; alternatively, we could also take the cross-sections themselves as the entities. Mathematically it would be simpler to take the latter route.

1.2 There is a binary function M_X that takes the partitions of *e* and the weight function $W_C(T_D)$ as arguments, and returns a real value between 0 and 1. *e* contains imperfect exemplifications of T_D or a lack of particular exemplifications. More specifically $M_X = Sum(Part(e) \cdot WC(T_D))$ where *Part*(*e*) takes the partitions of *e* that are imperfect exemplifications of T_D and leaves everything else out, while assigning 0 everywhere *e* fails to exemplify T_D. *Part*(*e*) should thus have the same dimensions as $W_C(T_D)$, with the order of the labels corresponding to T_D. The dot product multiplies the values that have the same label in a pairwise fashion, so that the result is a vector of the same dimension as the original two vectors. The "*Sum*" function adds up the values in the resulting vector to yield a value between 0 and 1.

For a general example to illustrate how M_X works, consider a partition of T_D as $\{\{\gamma_1\}, \{\gamma_2\}, \{\gamma_3\}\}$, each γ representing a simple constituent of T_D. Suppose that $W_C(T_D) = [\ 0.2\ \ 0.3\ \ 0.5\ ]$ each element corresponding respectively to how efficacious

$\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$, and $\boldsymbol{\gamma}_3$ is in a context *C*. Again, for simplicity of presentation I shall leave how to obtain $W_C(\mathrm{T_D})$ from *C* and $\mathrm{T_D}$ below. Now, suppose that *e* could be partitioned as the following simple constituents $\{\{0.9\boldsymbol{\Gamma}_1\}, \{\boldsymbol{\Delta}\}, \{0.8\boldsymbol{\Gamma}_3\}, \{\mathbf{B}\}\}$. We say that $\boldsymbol{\Gamma}_1$ is the exemplification of $\boldsymbol{\gamma}_1$ and so on. The relevant partitions of *e* are thus $\{0.9\ \boldsymbol{\Gamma}_1\}$ and $\{0.8\boldsymbol{\Gamma}_3\}$, with nothing exemplifying $\boldsymbol{\gamma}_2$. So *Part*(*e*) = [0.9 0 0.8]. Therefore M_X = *Sum*([0.18 0 0.4]) = 0.58. Let us refer to this as the 'degree' with which *e* approximates $\mathrm{T_D}$ in *C*.

To get an idea of how M_X might work for a concrete example, consider a simple description such that of the circle. Let *e* be the appropriate cross-section of a wheel, and suppose that the wheel has 1/2 of its circumference[194] at *r* metres from the centre *O*, and 3/10 of its circumference at *s* m from *O*, and 1/5 of its circumference at *t* m from *O*. The average radius would then be $avg_r = \frac{1}{2}r + \frac{3}{10}s + \frac{1}{5}t$. Let *C* be the pragmatic context of the wheel being able to roll forward with some energy efficiency requirement *E*. Then *E* determines the minimum value: M_C presumably by applying one's knowledge of physics. The set of ~XE in this case would be the set of all concepts of two-dimensional Euclidean figures, since these figure and only these are attributable to the same entities as a circle would and are incompatible with being a circle. Now:

$$M_X(e, W) = \frac{1}{2}(1 - \frac{r - avg_r}{avg_r}) + \frac{3}{10}(1 - \frac{s - avg_r}{avg_r}) + \frac{1}{5}(1 - \frac{t - avg_r}{avg_r}).$$

The terms $\frac{r - avg_r}{avg_r}$, $\frac{s - avg_r}{avg_r}$, and $\frac{t - avg_r}{avg_r}$ represent the deviation from the average radius. Suppose that $M_X > M_G$ for all $G \in {\sim}X$, then *e* adequately approximates being a circle iff $M_X \geq M_C$. For complex concepts, the weight function W_C becomes crucial.

1.3 M_X is greater than or equal to the minimum M_C for which *e* functions *adequately* similar to e_T in *C*. M_C is a binary function that takes *C* and $\mathrm{T_D}$ as

[194] 'Circumference' here just is a shorthand for 'the edge of the appropriate cross-section of the wheel'.

inputs, and returns a real value between 0 and 1. In other words, M_C is the minimum threshold below which e can no longer be said to adequately approximate T_D for the purposes of C. M_C is to be computed as follows:

Suppose that $C = \{\langle \mathbf{E}_1, 0.6\rangle, \langle \mathbf{E}_2, 0.4\rangle\}$ In other words, C has two purposes $\mathbf{E}_1$ and $\mathbf{E}_2$ that it requires T_D to fulfil, with $\mathbf{E}_1$ being 1.5 times more important than $\mathbf{E}_2$. Again, let T_D be partitioned into $\{\{\boldsymbol{\gamma}_1\}, \{\boldsymbol{\gamma}_2\}, \{\boldsymbol{\gamma}_3\}\}$, suppose that the *minimum requirement* for $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$, and $\boldsymbol{\gamma}_3$ to fulfil $\mathbf{E}_1$ is to degrees 0.2, 0.3, and 0.3 respectively. Note that this is not how much $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_2$, and $\boldsymbol{\Gamma}_3$ would actually fulfil $\mathbf{E}_1$, because we wish to compute the minimum threshold rather than how e_T would fulfil the desiderata of C. Now, do the same for $\mathbf{E}_2$ and suppose that the values are 0.1, 0.4, and 0.2 respectively for $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$, and $\boldsymbol{\gamma}_3$. M_P is then computed by adding up the satisfaction of $\mathbf{E}_1$ by the conceptual constituents, and multiply that by its degree of importance in C, which is 0.6; and doing the same for $\mathbf{E}_2$ and sum the values together.

i.e. $M_C = (0.2 + 0.3 + 0.3)(0.6) + (0.1 + 0.4 + 0.2)(0.4) = 0.76$

The computation can be generalised and operationalised by treating C as a vector without the **E**s and take the dot product of $Min(T_D)$ and C, then taking the sum of the elements. $Min(T_D)$ is of course the minimum satisfaction requirement for the singletons of T_D. In general, $Min(T_D)$ is to be determined empirically.

To make sense of the formalities above, one could think of some toy interpretations. Say that we are in a context of constructing a classical guitar, and that $\mathbf{E}_1$ is the end of accurate intonation, and $\mathbf{E}_2$ is the end of pleasant timbre. Suppose that T is the cooked up notion of *being a good guitar*, where a musical instrument is a good guitar iff it has (1) a light soundboard; (2) dry solid woods; (3); precise fret placements. Now suppose that the minimum requirement for the instrument to satisfy being a good guitar for the purposes of accurate intonation and pleasant timbre is for the soundboard to contribute 20% towards accurate intonation, the wood to constitute

30%, and the fret placement to contribute 30%. If any one of these three constituents contributes less, then the guitar has failed its purpose of maintaining accurate intonation and producing a pleasant timbre. Anyhow, I hope that this much is enough for understanding my formal representations, the reader may fill in the rest of the story as she or he pleases.

1.4 For all $G \in \sim X$, $M_G \leq M_X$.

This condition tells us that for any rival description G that is incompatible with $T_D(X)$, the degree with which a target e approximates G is no greater than the degree with which e approximates $T_D(X)$.

The condition 1.3 takes care of the necessary component of e being adequately similar to e_T in C. The condition 1.4 makes sure that e functions more similarly to e_T than any entity that would have fully exemplified a description from $\sim X$.

As promised in 1.1, I devise two rules for *weighing partitions in a pragmatic context*:

(Rule-C): Let H be the partition of T_D with only singletons. Let N be the set of all elements of H that are necessary for C. Let V be the sum of the elements of a vector.

Then for any $A \in N$, $V(W_C(H\backslash A)) < V(W_C(N))$.

(Rule-Mn): Let e_N be an entity that satisfies all and only elements of N.

Then, $M_C \geq M_X(e_N, W_C(T_D))$.

Consider the situation where an entity e satisfies, say, every element of its target T_D except for one crucial element that would be necessary for its context C. Rule-C ascertains that $M_X(e, C)$ in this scenario will be less than $M_X(e^*, C)$, where e^* is an entity that would have satisfied all of the necessary elements of T_D but nothing else. Rule-Mn tells us that the minimum value M_C for e to adequately approximate T_D must be no less than the degree of adequacy for a e_N that satisfies all of the necessary elements of T_D for C. To determine M_C is a much more difficult task in reality, without cooked up numbers, for all the complexities involved in telling when M_X is large enough to be an adequate degree. Nonetheless, Rule-Mn is a significant step closer to our goal. Furthermore, if N is non-empty, due to the fact that $V(W_C(N)) > 1/2$ and that usually it's very large, in practice, we often need only $M_X(e_N, W_C(\mathrm{T_D}))$ rather than M_C. Of course, in the difficult case of $N = \varnothing$, the determination of M_C cannot be helped by Rule-Mn. To see why $V(W_C(N)) > 1/2$ and that it's generally expected to be very large:

If card(N) = 0, then trivially $V(W_C(N)) = 0$.[195]

If card(N) = 1, then $(\forall A \in N)$, $V(W_C(H\backslash A))$[196] $+ V(W_C(N)) = 1$.

Since $(\forall A \in N)$ $V(W_C(H\backslash A)) < V(W_C(N))$; $V(W_C(N)) > 1/2$.

If card(N) = 2, then $(\forall A \in N)$ $V(W_C(A)) + V(W_C(H\backslash A)) = 1$

By Rule-P: $(\forall A \in N)$ $V(W_C(A)) + V(W_C(N\backslash A)) > V(W_C(H\backslash A))$.

so $(\forall A \in N)$ $V(W_C(A)) + V(W_C(N\backslash A)) > 1 - V(W_C(A))$,

so $(\forall A \in N)$ $2(V(W_C(A)) + V(W_C(N\backslash A)) > 1$

The smallest value $V(W_C(N))$ could bear is when $V(W_C(N\backslash A)) = V(W_C(A))$.

To see this, suppose $V(W_p(N\backslash A)) < V(W_p(A))$, since A is arbitrary, we need only consider this one direction.

We want both $2(V(W_C(A)) + V(W_C(N\backslash A)) > 1$

and $(V(W_C(A)) + 2V(W_C(N\backslash A)) > 1$

Now $1 - V(W_C(A)) < V(W_C(A)) + V(W_C(N\backslash A)) < 2(V(W_C(A))$,

and $1 - V(W_C(A)) < V(W_C(A)) + V(W_C(N\backslash A)) < 2V(W_C(N\backslash A))$,

so $1/3 < V(W_C(A))$ and $1/3 < V(W_C(N\backslash A))$

Thus $V(W_C(N)) = V(W_C(A)) + V(W_C(N\backslash A)) > 2V(W_C(N\backslash A)) > 2/3$

If $V(W_C(N\backslash A)) = V(W_C(A))$,

then $V(W_C(N)) = V(W_C(A)) + V(W_C(N\backslash A)) = 2(V(W_C(A))$,

so $3(V(W_C(A)) > 1$, and $2(V(W_C(A)) > 2/3$

Thus $V(W_C(N)) > 2/3$

The reasoning for card(N) = 2 goes through for any card(N).

The smallest value $V(W_C(N))$ could bear is the value when

[195] 'card' returns the cardinality of a set.

[196] The set-theoretic notation of 'N\A' means N without A.

$$\forall A \in N, \frac{V(WC(A))}{V(WC(N))} = \frac{1}{card(N)}^{+}$$[197]

From previously, ($\forall A \in N$) $2(V(W_C(A)) + V(W_C(N \backslash A)) > 1$ (This depends not on the value of card(N)).

Therefore $2\frac{V(W_C(A)))}{V(W_C(N))} + \frac{V(W_C(N \backslash A))}{V(W_C(N)} > \frac{1}{V(W_C(N))}$

Therefore $\frac{2}{card(N)} + \frac{card(N) - 1}{card(N)} > \frac{1}{V(W_C(N))}$

Therefore $V(W_C(N)) > \frac{card(N)}{card(N) + 1}$

It's thus established that a mapping that takes card(N) to the minimum value of $V(W_C(N))$ is the function $F(\mathrm{x}) = (\mathrm{x} / (\mathrm{x} + 1))^+$ for integers x ≥ 0. Thus $V(W_C(N))$ would approach the asymptote at 1 very quickly. Now to show that we often need only $M_X(e_N, W_C(\mathrm{T_D}))$ rather than M_C, we must show that $M_X(e_N, W_C(\mathrm{T_D}))$ also approaches 1 very quickly as $V(W_C(N))$ approaches 1. In other words, we show that the margin of error for M_C becomes rather small as card(N) increases.

One natural way to construct M_X is by extending the ideas from my previous example with the simple concept of being a circle. In the general case, we can treat the argument *e* as partitioned into its singletons. Take the computation from the circle example as one particular correspondence, M_X then computes how much each element of *e* approximate the corresponding elements from $\mathrm{T_D}$. Since by definition e_N satisfies *N* perfectly well, that means $M_X(e_N, W_C(\mathrm{T_D})) = V(W_C(N))$. Furthermore, for any e_N* that satisfies *at least N* perfectly well, $M_X(e_N^*, W_C(\mathrm{T_D})) \geq V(W_C(N))$. From the previous result we know that $V(W_C(N)) >$ card(N)/(card(N) + 1). Therefore $M_X(e_N, W_C(\mathrm{T_D})) >$ card(N)/(card(N) + 1).

[197] Just one unit greater than 1/card(N), since the inequality is strict.

3.6 On Changing the Resolution

3.6.1 Problem 1: At the Limit

A serious worry emerges for the approximation relation: ‘A perfectly circular wheel on a perfectly flat surface would have exactly one contact line perpendicular to the direction of motion, and thus would fail its function of rolling forward, let alone do so with a minimum efficiency requirement. Then how can a state of being the appropriate cross-section of a wheel approximate being a circle for the function of rolling forward with a minimum efficiency requirement?’

Reply: One obvious fault with the enquiry is that it has both a perfectly circular wheel and a perfectly flat surface under consideration. For assessing entities, we would always fix the context, leaving only one entity for consideration at a time. In this case, if we are to consider the entity of being a wheel, then we must fix the road surface as part of the context.

However, even if the road isn’t perfectly flat, a perfect circular wheel would still be too smooth to efficiently roll forward (for example, it would skid). This wouldn’t be enough to resolve the difficulty. A better answer would be that when we consider the pragmatic context of wanting the wheel to roll forward, we consider each unit of *sufficient length for motion* along the circumference (of the cross-section) *as a point* on the two-dimensional cross-section of the wheel. In that case, we see that given this minimal unit of measurement, indeed the perfect circle is the ideal towards which the cross-section of this wheel should approximate. This is the bridging technique of *changing the resolution*.

To give another example, think about graphic calculators with discernible pixels. We are satisfied with identifying a graph of a circle as being an adequate representation of a circle even if not all of the pixels are of equal pixels distant from the centre. Note that this is not the same as saying that all of the points on the representation are not equidistant *per se* from the centre — the smallest units of measurement differ between the two claims. Here the reader may be suspicious of a triviality result, namely, that anything can approximate its target adequately enough if we just modify the unit of measurement to our likings. Returning to the wheel example: Can one really just use arbitrary units for the practical context of wanting it to roll forward along the road efficiently? The answer is obviously no, because we are considering a practical context, not just playing language games — the smallest unit must have a sufficient amount of contact with the ground for grip to occur. Of course, the minimal unit of measurement could change depending on how large the wheel is.

Another reason why we say that wheels are circular is because (the cross-section of) the wheel is closer to being a circle than any other clearly defined geometrical figure. This is one of the conditions both bridging relations must satisfy.

3.6.2 Problem 2: Truth Claims

A second benefit of changing the resolution over using approximation is that we can preserve the T-schema: a sentence '*p*' is true iff *p*. Under the approximation relation, we bridge the gap between a precise description $T_D(X)$ of X and an imprecise term 'X' in a context *C* by saying that 'X' refers to some *e* such that *e* behaves just like e_T *in C,* where e_T fully exemplifies $T_D(X)$. As a result, '*e* is X' is true, or at least a proper ascription but *e* isn't X (it *approximates* X in *C*). By changing the resolution, we no longer need to think of e_T, instead, we simply say that *e* satisfies $T_D(X)$ in *C*.

The way that *changing the resolution* works is by changing the scale on which we measure the parameters of $T_D(X)$. Recall the time example from above: the context specifies the smallest significant unit of time, such that 'being on time' is true in different contexts, even though 'being on time' can refer to *being 5 minutes late, being 1 second early*, being *1 nanosecond late* etc. As a result, we have '*e* is X' is true *and e* is X, since given our specifications. The wheel just *is* a circle rather than approximates a circle, since the smallest point on a wheel and the smallest salient unit of length is such that all of those points are equidistant from its centre. Being 5 minutes late just is being on time for dinner reservations since the smallest salient unit of time is 10 minutes. That means, being 5 minutes late and being 1 minute late makes absolutely no practical difference for all the purposes of a dinner reservation. *The* *context of the dinner reservation has a resolution that cannot distinguish between 5 minutes and 1 minute.*

The key formal difference between this technique and approximation is in change the way we measure contextual desiderata against the parameters of $T_D(X)$. We still partition $T_D(X)$ into its sets of features, and every other procedures and results stay the same. In approximation, we measure how closely e approximates e_T with the partitioned features and the desiderata of a context C. This required C to specify *how much* of each of T_D's parameters it requires. This is where *changing the resolution* differs – now C specifies the *minimum unit of significance* for each of T_D's parameters that it wants. This one simple change simplifies the bridging idea and fixes the two bugs mentioned above. So we can see this changing of resolution technique as a revised version of approximation, rather than as a separate technique.

3.7 Summary

Tarski's theory of truth is a descriptively adequate theory of truth. Tarski tried to account for the notion of truth that was already vaguely conceived of before in ordinary language and in philosophy. After all, the point of the description was to make the notion of truth *more cogent,* rather than creating something completely new. Of course, sometimes theorists invent new terms, but that isn't what scientists or philosophers primarily do.

We have seen that the Strawsonian concern is about two intertwined issues. One, it complains that philosophy as a discipline shouldn't theorise in specialised domains. Two, it complains that imprecise terms cannot be replaced by precise terms. The entire enterprise of machine philosophy is a rejection of the first complaint. Philosophy should indeed specialise, since there is otherwise no way to do better than folk-psychology, folk-epistemology etc. In particular, we must specialise and revise in order to provide *true* and relevant descriptions. Moreover, what we think of as legitimate questions before we have a precise description might really be questions that cannot be answered by any true description. As for the second complaint, we have provided a bridging technique to account for the discrepancy between imprecise ordinary language terms and precise scientific terms. This is important since the precise descriptions are after all, aimed at providing insight on our ordinary language referents.

Chapter 4 On Normative Theorising

'*… the distinction of vice and virtue is not founded merely on the relations of objects, nor is perceived by reason.*'[198]

4.1 Introduction

Machine philosophy makes two core claims. One: intuitions are fallible evidence in philosophy, which reflect *objective* facts about our *socio-linguistic realities*. Two: good philosophical theories are *descriptive* of our socio-linguistic realities in virtue of being *statistically adequate models of our intuitions*. A question naturally arises: Are normative theories such as those within normative ethics descriptive of our socio-linguistic realities? The worry is that if normative theories are not part of our socio-linguistic realities, then specifically normative disciplines such as normative decision theory or normative ethics cannot be within the scope of machine philosophy.[199]

I argue that ethics or other 'normative' disciplines are squarely within the scope of machine philosophy. In 4.2, I distinguish normativity as a monadic linguistic feature from normativity as an epistemic relation. I use the examples of physics and logic to illustrate this distinction, and argue that theories are normative *in relation* to a certain domain. In particular, normative theories are *not* normative statements. While normative statements are normative in virtue of making normative claims, normative theories are normative in virtue of relating to another domain in a specifically epistemic manner. In 4.3, I introduce a parallel distinction between the grounding for a statement's normativity and the grounding for a theory's normativity. I argue that in

[198] Hume (1739)

[199] Not that this is a pressing issue. Machine philosophy isn't logical positivism. It doesn't claim that there is one correct way to do philosophy. However, it also doesn't support a crude methodological pluralism. As mentioned in Chapter 2, machine philosophy is simply a natural evolution from conceptual analysis and conceptual engineering. If a domain allows for more advanced methods, then they should use those methods. If a domain cannot but for now be content with simple analysis, then they have all the right to stick with analysis. The very notion of epistemic justification that machine philosophy endorses is a pragmatic one, wherein the method that justifies within a domain just is the best performing method available for the domain. i.e. If you have more data for some X, make use of it. If not, well, do the best you can with what you have!

the latter's case, the normative components are irrelevant for the grounding of the theory itself. Therefore, the Humean is/ought argument cannot apply to normative theorising. I argue that a successful normative theory must also be successfully descriptive, and should abide by the same statistical norms that govern good descriptive theorising. As a corollary, I argue that having normative force in a related domain is a good criterion for testing whether a descriptive theory is fruitful beyond the domain in which it's factive. I use ethics as an example here, and in the process, provide the groundwork for examining ethical theorising in line with machine philosophy. In 4.4, I provide an account for normative ethics as a descriptive domain. In particular, normative ethics aims to describe the socio-linguistic reality of the domain of ethics, via intuitions that ethicists have about what kind of actions are ethical. I argue that ethical theorising is descriptive in so far that it aims to truly describe our morally relevant social-linguistic reality. I argue that theories in normative ethics are normative just in case they are descriptively adequate models of our moral intuitions. At the end of the chapter, I provide clarifications on traditionally contentious issues in meta-ethics: specifically, on the conflict between moral relativism and universality.

4.2 Normativity as an Epistemic Relation

In 1966, Peter Wason devised an experiment to show that people tend to commit the withholding of the contrapositive,[200] so human reasoning do in fact violate elementary logical norms.[201] He concluded that the behaviour was irrational and required correction. Gilbert Harman argued against Wason's conclusion, disassociating human reasoning from logic.[202] He argued that reasoning is the process of forming, revising, and maintaining beliefs. However, logic studies the relations among propositions, not how beliefs should be adopted or abandoned. For example,

[200] The failure to consider the contrapositive of a conditional statement when considering the veracity of the conditional. i.e. the failure to see that 'if not-q, then not-p' is logically equivalent to 'if p then q'.

[201] Wason (1968)

[202] Harman (1986)

we should avoid cluttering our limited memory with petty information, therefore we should at least violate inferential closure. Harman's theory rationalised the fact that reasoning violates logical norms. However, the result that reasoning do in fact violate logical norms played no role in grounding Harman's claim that reasoning *ought not to* follow logical norms. Rather, the core disagreement between Harman and Watson is on what reasoning *ought to do*. This example illustrates the Humean view that 'there is no value neutral argument for an evaluative conclusion'[203], even if two parties disagree on whether the descriptive and the normative correlate in some domain.[204]

While there is nothing wrong with having the Humean worry for normative *claims/sentences/propositions*. I argue that applying the Humean view to theories is a result of conflating normativity as a monadic feature with normativity as an epistemic relation. The former notion is familiar to us and straightforward: a proposition *p* is *normative* just in case it makes a normative/evaluative/value claim, as opposed to making a *factive* claim. The claim that 'humans ought not to reason in accord with logical rules' or 'pleasure is good' are normative claims about what *ought to be the case*. Contrasting these claims with 'humans do not reason in accord with logical rules' or 'humans want pleasure', which are claims about what *is the case*. The distinction here points to normativity as a monadic feature of a sentence. In particular, normativity of a sentence of the form 'it ought to be the case that X' is a feature of X. The question arises: Is this the kind of normativity when we talk of normative theorising? I'll answer this question by introducing a distinct notion of normativity as an epistemic relation:

A theory T_D in D is *normative in a certain domain* Q iff D is normative *for* Q.

Using the above example, D would be the domain of logic, and T_D would be a logical theorem, Q would be the domain of human reasoning. If logic is normative for human reasoning, then human reasoning should abide by logical theorems wherever applicable. Conversely, if human reasoning should abide by logical theorems

[203] This is a weak reading of Hume's is/ought distinction.

[204] A major challenge comes from Jackson (1998), (2000) on moral properties. e.g. a sentence containing a moral property can be re-expressed in a sentence that contain only descriptive properties.

wherever applicable, then logic is normative for human reasoning. This notion of theory being normative *for* some domain Q expresses a *relation* between a theory in a domain and a distinct domain. In the next few sections, I argue that this relation of normativity isn't reflective.[205] I also provide the conditions for which a theory T_D is normative for Q. As a preview, the two conditions are: a normative claim *N* in Q, and the descriptive adequacy of T_D in D. T_D is normative for Q just in case T_D helps Q to satisfy *N*.

When we say that a theory T_D 'is normative'. We can mean one of two things. First, we can mean that 'It *ought to be the case* that T_D', which treats normativity as a monodic argument over T_D. Second, we can mean that 'T_D is normative with respect to some domain Q.', which considers normativity as a relation between T_D and Q. On the first reading, we might think of T_D as having a different epistemic grounding from descriptive theories. However, I argue that the first reading cannot hold. Only the second reading makes sense. Moreover, the very claim that T_D is normative for Q must be grounded on the *truth* of T_D *in D*. Normative claims such as 'logic is normative for human reasoning' isn't a claim *about* logic, but a claim about the *relationship* between logic and human reasoning. In 4.2.1, I illustrate this point by showing how paradigmatically descriptive theories such as those in physics are normative with respect to other domains. In 4.2.2, I compare the truth conditions for a normative claim *about* some X with the truth conditions for a normative claim about the relation between a T_D and a Q.

4.2.1 Physics is to Engineering as Logic is to Reasoning

Let's consider *the theory of universal gravitation* (UG): $F_1 = F_2 = \dfrac{Gm_1m_2}{r^2}$, which states that every pair of masses in the universe attract each other with equal force, and the gravitational force between two bodies is proportional to the product of their mass, and inversely proportional to the square of the distance between them. We

[205] This is equivalent to arguing that a theory T in D isn't normative for D.

can use UG to compute the amount of force F that acts on an object of mass α by another object of mass β, given a distance d between them: $F = \dfrac{G\alpha\beta}{d^2}$. Furthermore, we can compute other parameters such as acceleration via theories such as Newton's second law: $F = ma$. Such computations enable aeronautical engineers to gauge how a spacecraft would behave under non-negligible gravitational influence. Engineers would then be able to design optimal flying routes and use parameters that would raise the probability of a successful mission. On the other hand, if engineers were to violate UG, or some other relevant physical theory, say, by computing $F = \dfrac{G\alpha\beta}{d^3}$, then they would work with a value of F that would lower the success rate of their mission. This is because UG is a *true* description of the relevant physical reality. In other words, UG governs the correct design of space flights, in virtue of being a true description of physical reality. I argue that this 'governing' is in the same sense as how the law of the contrapositive in propositional logic governs correct inference.

For the sake of argument, suppose that classical propositional logic (PL) is normative for reasoning tasks in which only binary truth-values and true premises are concerned, and where truth is the only goal. So if a proposition or schema S is a theorem of PL, then our reasoning tasks should abide by S and not violate S wherever S applies. The question is: what are the grounds on which our reasoning should abide by S? Here is a preview of the answer: because S is a true description of the relevant logical space. Here is the argument.

Consider the following theorem of PL: $\vdash (A \to B) \to (\neg B \to \neg A)$, which says that if a sentence A implies a sentence B, then the negation of B implies the negation of A.[206] By deduction theorem, we have the law of the contrapositive (CP): $(A \to B) \vdash (\neg B \to \neg A)$, which says that given the premise 'if A then B', we can

[206] To be sure, the material conditional ('→' in classical PL) does not capture all valid inferences in ordinary reasoning. However, there are two reasons why this consideration can be deemed as innocuous, at least here. First, UG too does not capture all of the viable models in engineering, but nonetheless we do use UG for the cases where it *does* capture a viable model. Second, the material conditional does indeed capture a genuine set of cases where only binary truth-values are concerned and where the premises are true.

validly infer that ‘if *not-B* then *not-A*’. By the soundness of PL, we could say that the truth of ($A \rightarrow B$) guarantees the truth of ($\neg B \rightarrow \neg A$). Other theorems of PL could also be turned into rules of inference in virtue of the deduction theorem. Soundness tells us that such rules are semantically valid – a true premise in a valid reasoning must result in a true conclusion. Therefore, a reasoner who follow these theorems would reason from true premises to true conclusions. On the other hand, if a reasoner were to violate CP where it applies, then that reasoner could reason from truths to falsehoods.[207] This is because in the domain of PL, CP is a true description of a valid inference, in the same sense that UG is a true description of a physical relation in physics. The target of physics is physical reality, while the target of PL is logical space. PL is a body of theories that truly describe the logical space of PL, just as physics is a body of theories that truly describe physical reality.[208] Violating CP would mean reasoning with a logically invalid rule. Inferring from true premises using a logically invalid rule cannot guarantee the truth of the conclusion. In other words, CP governs reasoning in virtue of being a true description of the logical space, in the same sense of how UG governs engineering in virtue of being a true description of physical reality.

Before moving on, it’s worth noting the normative scope of a theory. One could say that in cases where UG or CP return the wrong results, one ought to violate UG or CP. In such cases, UG or CP simply aren’t normative. However, this cannot change the fact that they are normative for tasks in which they do apply. i.e. A theory’s scope of descriptively determines its scope of normativity.

For example, UG should be violated in cases where general relativity overrides Newtonian mechanics, such as when an object’s mass approaches the density of a black hole. In such cases, UG returns the wrong results, because it simply doesn’t

[207] By ‘violating CP’, I mean specifically using a rule that conflicts with the results of CP. For example, denial of the antecedent, affirmation of the consequent, withholding the contrapositive are all violations of CP. I don’t mean using a more expressive or weaker logic instead of PL, just as we don’t mean violating UG in the sense of using relativity instead of Newtonian mechanics. Using a probabilistic conditional is not a violation of CP in my picture.

[208] Whether this logical space exists in the platonic sense, or whether it is nominalistic is irrelevant. The only thing that matters is that it is a genuine

describe matters and interactions near a black hole. The cases for which general relativity applies are cases where variables that are neglected in Newtonian mechanics become non-negligible.[209] As I have argued in Chapter 2, such considerations do not make UG a false description. It likewise cannot affect UG's normativity for the tasks in which it does apply. In cases where general relativity is needed, UG would be simply inapplicable, because it doesn't describe the relevant features of reality for such cases. In other words, UG is normative for those and only for those situations wherein UG truly describes the relevant parameters of the situation. In a similar spirit, one could say that the crux of Harman's argument was that logic simply does not capture all the relevant factors in human reasoning, such as belief revision.

4.2.2 Normative Theories and Normative Projects

Philosophers ascribe normativity to entire projects. We say that normative decision theory is distinct from descriptive decision theory. I argue that we can only ascribe normativity as a relation between a theory and a domain, rather than as a feature of a domain. We cannot ascribe normativity as a monadic feature to domains or the theories within like we do to sentences. Moreover, we cannot systematically divide a domain D into projects that are consistently normative for a domain Q. A claim that Newtonian mechanics is normative for civil engineering but not for aeronautical engineering would simply be false.[210] It seems to follow from the consideration of a theory's restricted applicability that in general, there is no body of theories that are consistently normative for a given domain. The question is then, what does it mean when someone claims to be engaged in a normative, not descriptive project? My reply is that there is no such thing as a normative project, although there are certainly normative *relations* between the theories of a domain and a distinct domain. The illusion of an essentially normative project is due to the fact that we have

[209] Variables such as momentum, speed etc.

[210] This sometimes work, for example, in the normativity of logic for mathematics – classical logic is normative for classical mathematics, while constructive logic for constructive mathematics. However, this is due to a peculiarity with logic and mathematics, and fails for other domains such as decision theory for decision making, or logic for reasoning.

no good accounts of what particular domains such as ethics describe.[211] In 4.4. I will argue that all such theories describe our socio-linguistic realities – there is no difference between epistemology and ethics in their grounding, including the kind of evidence they use. Let us begin by examining the example of expected utility theory (EUT).

Normative decision theory is a paradigmatic normative project. The goal of normative decision theory is not to describe how we actually make decisions, but how we ought to make decisions. For example, expected utility theory (EUT) is nowadays considered to be a normative theory within the project of normative decision theory. EUT tells us that in the face of uncertainty, we should decide based on the principle of maximising our expected utility. This example is illuminating because EUT had been considered to be a descriptive theory in classical economics. This was until Kahneman and Tversky pointed out in 1974 that the classic models in economics falsely assumed that humans are perfectly rational, or rather, *homo economicus*.[212] So a descriptive model of decision making should treat its subjects as having *bounded* rationality. So what makes EUT *normative* for human decision making is a conjunction of three claims. One, humans are *not* perfectly rational. Two, EUT describes how *ideally rational agents – homo economicus – would* make decisions. Three, *homo economicus* represents ideally rational agents.

Suppose we accept the notion of a distinctively normative project with distinctively normative methodologies. We can say that a project on X is normative just in case the goal of the project is to describe the ideal features of X. By contrast, a descriptive project on X aims to state the actual features of X. So is not 'describing

[211] Any domain in which philosophers have realist vs. anti-realist debates are ones where we are yet to find a good ontology. In this regard, machine philosophy is like a new microchip architecture that inadvertently solves an entire class of problems. Indeed, this is the very spirit of machine philosophy or machine epistemology a la Wheeler (2012) – circumventing issues by designing better methodological architectures rather than trying to tackle those problems head on with an old (boolean) architecture. In this case, the aspect of the new architecture responsible for dissolving the realist/anti-realist debate the understanding of intuitions as reflective of our socio-linguistic realities, which entails moral cognitivism while accounting for non-cognitvistic intuitions. Still, I must elaborate on this in the next section for heuristic efficacy.

[212] Tversky & Kahneman (1974)

the ideal features of X' the very same activity with describing the actual features of X, just with a different target?

One might object on two grounds. First, the source of description is fundamentally different in the two cases. In the case of a descriptive project, the target of description is objective reality. In the case of a normative project, the target of description is an idealised reality, dependent on a normative claim. In this case, the very target of description is dependent on accepting the third claim that *homo economicus* is indeed ideally rational. However, it is an open question why we take, say, *homo economicus* instead of, say, *homo reciprocans* as the *idealised* version of rational agents?[213] However, in Moore's own spirit,[214] I highlight an equivalence between the objectivity of the idealised space and the open question argument. The open question argument *relies on* the supposition that what counts as an *idealised* reality of X isn't objective. On the other hand, the legitimacy of the open question argument *entails that* what counts as an idealised reality of X isn't objective. So we cannot use the open question argument as grounds for the lack of objectivity of an idealised X. i.e. We need independent grounds on either side in order to argue against the other. In this chapter, I provide the independent ground for establishing the objectivity of an idealised X, thereby kicking the open question argument out of the room. As a preview, the solution is simply to point out that the requirements of statistical adequacy provides an objective ground for accepting a certain choices of datasets. In order to establish that, we need to claim that to adequately describe idealised features of X, we need to follow our standard statistical norms for descriptive theorising.

I provide two independent responses. In the next section, I argue that the open question is orthogonal for the process of theorising itself. The normative claim that rational agents strive to be *homo economicus* and the truth of EUT as a description of *homo economicus* jointly ground the normativity of EUT for decision making.

[213] A variation on the 'open question argument' from Moore (1903)

[214] a la Moore (1925)

However, the normative claim that rational agents strive to be *homo economicus* is independent of whether EUT is descriptively adequate of *homo economicus*. In 4.4, I argue that these idealised realities are objective, in the spirit of machine philosophy.

Before proceeding, I'd like to make a clarification. I'm not attacking the distinction between, say, descriptive decision theory and normative decision theory. I am arguing that the normativity of a theory is grounded on the relation between that theory and a distinct domain, and that this relation has no bearing on the activity of producing and validating a theory.

4.3 Ethics and Objectivity

First, recall the distinction between normative theories and normative statements. Consider a theory $T_D(X)$, a distinct domain Q, and the claim that $T_D(X)$ is normative for Q. $T_D(X)$ is therefore a normative theory for Q, and the claim that '$T_D(X)$ is normative for Q' is a normative claim. I argue that the descriptive adequacy of $T_D(X)$ in D and a set of normative claims about X in Q jointly ground the normativity of $T_D(X)$ for Q. Requiring a normative premise in the argument for a normative claim is trivial. That was Hume's very point: $T_D(X)$ alone cannot tell us whether it is normative for Q. This section argues that the normativity of a theory $T_D(X)$ for some Q is also substantially grounded on the *descriptive adequacy* of $T_D(X)$ in D. While Hume is right that we cannot argue for the normatively of $T_D(X)$ for Q on the descriptive adequacy of $T_D(X)$ *alone*, the production and validation of $T_D(X)$ does not at all depend on the evaluative claims for which $T_D(X)$ is normative. i.e. There is no distinct methodology of normative theorising.

4.3.1 Grounding the Normative on the Descriptive

Let us turn to normative ethics for the remaining of this chapter. One form of act utilitarianism claims that an action is ethical just in case it increases the total happiness of all sentient beings. Let's denote it as $T_{nE}(X)$, where X stands for 'moral actions' and 'nE' stands for 'normative ethics'. Act utilitarianism is a theory in normative ethics on what a moral action is. It is a normative theory for the domain of actual human actions: aHA. $T_{nE}(X)$ is not descriptive for aHA because it does not accurately portray aHA. However, $T_{nE}(X)$ does aim to capture how an *ideally moral* agent *would* act. The normative component, and also our moral intuition here is of course that the parameter of 'producing joy' is treated as being the most important parameter for a moral action. I will argue in the next section that this normative component is in fact a reflection of our socio-linguistic reality. For now, we can safely hold off on this point. Now, T_{nE} is a description of an idealised space of X (given the claim that 'producing joy' is the most important parameter). Following our practice from the previous chapters, we construe this idealised space as a set X_{nE} where X_{nE} is the set of all instances of 'moral action' that satisfy the moral intuitions of ethicists. i.e. X_{nE} demarcates the actions of perfectly moral agents. The moral intuitions would include propositions such as 'pleasure/happiness is good' etc. X_{nE} can be derived from our moral intuitions via various means of rational inference. One is to use the help of game-theoretic models, which examine the distribution of utility in settings that involve multiple agents interacting with each other.[215] For example, consider a two-player iterated prisoner's dilemma game (IPD). IPD models a particular kind of interaction where there are two actions *cooperate* and *defect*, with the following payoffs: if one player *cooperates* and the other *defects*, then the player who *cooperates* would be very upset and the player who *defects* would be very happy; if both *cooperate*, then both players would be mildly happy; and if both *defect*, then both would be mildly upset. Suppose that the optimal strategy for maximising happiness for each individual player is 'tit-for-tat': *cooperate* on the first round, and then copy whatever the other player did in the previous round.[216] Suppose that both players are perfectly rational, so that both would play the optimal strategy. Then both

[215] This is a method that has been used to do descriptive ethics a la McKenzie Alexander (2007). Notice that game-theory is now playing a normative role for the domain of normative ethics.

[216] Even though Axelrod (1980) have presented tit-for-tat as an optimal strategy, from his own tournament. He later in Axelrod (1984) proved that there is in fact no optimal strategy. Nonetheless, tit-for-tat is one of the simplest strategies, requiring very little computational resource relative to more sophisticated strategies. So if we put efficiency constraints on being a rational agent, then we might end up with tit-for-tat after all. But that is a discussion for another topic.

players would always *cooperate*, and they would effectively be acting in accord with $T_{nE}(X)$ since they'd be increasing the total happiness of all agents. In this case, IPD tells us that if you're rational, and if you want to maximise your own happiness, then you would also act in accord with act utilitarianism. In other words, $T_{nE}(X)$ *describes* X_{nE}: what ideal agents would do, given our moral intuitions and rational presuppositions.[217]

Consequently, any proper argument against $T_{nE}(X)$ must be an argument against the way ethicists have constructed $T_{nE}(X)$ upon X_{nE}. For example, one could complain that while 'producing joy' is an important parameter for a moral action, there are other parameters that should be weighed against 'producing joy'. In particular, there could be situations in which 'being just' might come into conflict with 'producing joy'. This complaint could be about a bad feature selection (ignoring 'being just'), or about a high bias in $T_{nE}(X)$ (in which case we need to increase the complexity of $T_{nE}(X)$ to account for the variance in X_{nE}). One could also complain that tit-for-tat isn't the optimal strategy across all situations.[218] There are situations in which the optimal decision won't maximise happiness. One could also point out that the value of happiness is logarithmic, and that 'equality' and 'fairness' are important parameters for moral actions, such that no 'utility monster' should take all the happiness there

[217] Rational presuppositions are those that are shared for all means of theorising. These include logical and statistical reasoning, and derivative methods thereof.

[218] Since there is no optimal strategy for all situations (Axelrod 1984), not to mention 'happiness' isn't the only final value to which an individual strive towards. It would therefore be quite a complicated task to generate the dataset of perfectly ethical actions for the sake of ethical theorising. This complexity has certainly contributed to the illusion that ethics isn't grounded on objective facts of reality.

is.[219] All of these arguments against $T_{nE}(X)$ comes down to an argument against how we massage the data in X_{nE}, or how $T_{nE}(X)$ fits mA_{nE}. They are not arguments against our moral intuitions. Moreover, they cannot be arguments against our moral intuitions, since there is no further grounding for that. Arguments against how we do feature selection and how we idealise aren't arguments against the data itself. This is an important distinction.

In general, every normative theory $T_D(X)$ is descriptive of D. Suppose that $T_D(X)$ is normative for a domain Q. This means that $T_D(X)$ need not describe Q, and Q ought to abide by $T_D(X)$. It follows that abiding by $T_D(X)$ is good for Q, as per some normative claims *N*. For $T_D(X)$ to have such a guiding role, it must truly describe X in D, where Q would satisfy *N via X*. For example, a logical theorem truly describes the inference rules of a logical space such that reasoning with them would lead to true beliefs. Physical theories truly describe the relations of the physical world such that doing engineering with them would result in success; normative ethics truly describes the rules of an ethical space such that acting in accord with them is ethical. The fact that physics is incomplete, or that ethical theories are contentious has nothing to do with their normative force being grounded on their descriptive success.

4.3.2 Ethical Space is Part of our Socio-Linguistic Reality

Even though we can say that an ethical theory $T_{nE}(X)$ describes a dataset X_{nE} of X in normative ethics, one might insist that this dataset is unlike the dataset in, say, epistemology or physics. In particular, moral intuitions are fundamentally different

[219] Nozick (1974) introduced the idea to argue against utilitarianism. However, this kind of wild argument in philosophy is really quite ludicrous and unfortunate. It's as if philosophy isn't constrained by objective facts. In physics, thought experiments not only obey physical laws, they are efficacious in virtue of obeying physical laws. Philosophy shouldn't think of itself as special in this regard. In this case, the positing of a 'utility monster' ignores the fact that happiness or pleasure isn't additive, let alone the diminishing marginal value of commodities (Broom 1994). The supposition cannot even hold for perfectly rational and moral agents. At the very least, I don't think a perfectly moral agent should also at the same time have the capacity for an infinite number of neural fibres firing for an infinite amount of time, or violate the axiological fact about diminishing marginal values. It's like telling a mathematician 'what if there are a finite number of prime numbers?' Nozick's 'utility monster' isn't an argument against utilitarianism, it's an exercise of counterfactual questioning. This kind of wild arguments lead us astray, into all the wrong kind of premises. It's akin to metaphilosophical relativism and methodological anarchy, undermining the scientific rigour of philosophy.

from epistemological intuitions. Moral intuitions are *normative*. It is in this sense, my interlocutor claims, that nE is normative. So we can attribute normativity to theories in nE in virtue of this axiomatic feature of nE, contra my claim that normativity can only be attributed to the relation between ethical theories and a distinct domain.

However, there is no reason to think that our moral intuitions are distinct from epistemological or metaphysical intuitions in an epistemic manner. In particular, moral intuitions reflect our social-linguistic realities just as epistemological intuitions do. For example, the intuition that 'pleasure is good' is reflective of the fact that we *want* pleasure. The Humean complaint that we cannot *argue for* or *justify* 'pleasure is good' on the fact that we want pleasure is not only innocuous, but necessitated by my account of what intuitions are. Remember, intuitions by definition cannot be justified. They reflect or capture our domain-dependent socio-linguistic realities, in the same way that observations reflect or capture domain-dependent physical realities. The fact that 'pleasure is good' is a domain dependent intuition does not undermine its objectivity. Its objectivity is grounded in the fact that a community (in this case, anglophone ethicists) agree to the intuition. It is in this agreement that 'pleasure is good' captures the socio-linguistic reality of ethicists. The normativity of 'pleasure is good' does not change this fact.

Now one might complain about universality: Isn't normative ethics meant to reflect universal human values? I have two responses to this. One, I don't see why we should assume that our values are shared across distinct communities. This seems to me like moral imperialism that's painfully pervasive in traditional Catholic thinking: that everyone, Christians or not, should abide by the moral code embedded in Christianity.[220] Two, it would be a numerical fallacy to infer that just because two communities can have distinct values, their values must differ *substantially*. The fact remains, that even if communities could have different intuitions, whether on ethics, epistemology, ichthyology, they can still talk about the same topic in virtue of

[220] That's why Nietzsche even made the claim that 'God is dead!' in the Gay Sciences.

contextual factors.[221] The last three chapters, especially the last one on topic change, have made this clear. So one shouldn't be encouraged into moral relativism from my metaphilosophical claims. On the contrary, my account should provide a basis for why ethical arguments are objectively grounded in the way shown in 4.3.1. For example, one could discuss how to classify moral actions given the parameters of 'pleasure', 'justice', 'fairness', 'equality' etc. Moreover, this isn't a conflation with descriptive ethics, which describes the codes of conduct of different communities. Even if we stick with the community of anglophone ethicists. The code of conduct for these ethicists would differ from the actions of ideal agents generated by their moral intuitions. *Even if* the ethicists were to 'practice what they preach', the two theorising would still be of distinct datasets, since to practice what you preach, you must first find out what you preach. i.e. The task would be first to discover the theories that best fit their moral intuitions, and then act in accord with it. However, the parallel programme of descriptive ethics in this case wouldn't be the programme of describing how they would act *after* the theorising in normative ethics, but of how they acted beforehand. After all, descriptive ethics or descriptive decision theory aren't projects that try to describe 'how we would act *if we were to be* perfectly rational and moral agents'. They are projects that aim to describe how we *actually* act. Slavery in the anglophone world became abolished long *after* people in the anglophone world have argued for its immorality.

The above discussion commits us to the position that normative theories aim to be true within the theory's domain. Machine philosophy therefore commits us to cognitivism for all domains. In particular, a theory $T_{nE}(X)$ in normative ethics is meant to be a *true* description of X_{nE}. Where X_{nE} is all the instances of X that satisfy the intuitions of X in nE. The crux is that theories in normative ethics are to be constructed and evaluated in the same way as theories about physical reality – by following the methodological norms of statistical adequacy. In the remaining of this

[221] This also explains the apparently conflicting evidence we get from debates on whether Gettier intuitions are universal: See Weinberg et al (2001); then Nagel (2012) and Seyedsayamdost (2015); then Stich (2013); then Machery et al. (2015).

chapter, I provide a concrete account of how we might justify moral beliefs, and how they can bear truth values.

4.4 Epistemological Challenges of Ethics

I provide an account of moral claims in order to answer the following two epistemic challenges of ethics: How can moral beliefs be justified? Can moral beliefs bear truth values? The machine philosophy answer is a two-step procedure. First, we fix on the epistemic conditions for a moral *theory*, and then, we employ the bridging mechanism from Chapter 3 to account for ordinary moral utterances, just like how we would account for any other ordinary utterances. An account of moral claims should also provide an explanation of the dataset for ethical theorising. Hence the following descriptive questions should be accounted for: Why do we have the moral intuitions that we have? Where do the normative force of moral beliefs come from? Why do we have the intuition that morality is an objective matter? Why do different societies uphold different moral beliefs? I focus on the beliefs about moral actions. In particular, with regards to the question ‘What is a moral action?’.

In accordance machine philosophy, I introduce the notion of a value framework, and specify how one might go about theorising about what a moral action is. I use my account to explain contentious moral data, in particular, moral relativity and normativity. I conclude by summarising how my account answers the two epistemic challenges of ethics and explains the descriptive facts about moral claims.

4.4.1 The Semantics of Moral Claims

Let’s begin with some background. There are various accounts on what moral claims express. These accounts affect the normative scope of ethical claims. For example, if moral claims express necessary truths such as those of mathematics, then moral claims would not only play a universally normative role for guiding actions, but also for correcting beliefs that are inconsistent with these principles. For example, Kant's categorical imperative entails that moral claims are necessary.[222] If moral claims express objective but contingent facts grounded on sense observation, like those of physics, then moral claims are objective but fallible. Moral naturalism presents such an attitude towards the epistemic status of moral claims.[223] On the other hand, if moral claims express conventions, then while they maintain their normative force like laws in guiding actions, their normativity is of a limited scope. Often those who study moral norms from a game-theoretic perspective maintain this.[224] If ethical claims express mere personal opinions, then they would be just that — expressions of opinions, and hence ought not to have normative force. Moral non-cognitivism contain various accounts that amount to the claim that moral statements have no substantial truth conditions.[225]

The epistemic conditions of moral claims depend on their semantical content. If moral terms refer to natural entities, then ethical claims are to be justified by examining the *physical* relations and features of those entities. If moral terms are expressions of approval or disapproval, then moral claims cannot even bear truth values. Debates between various camps on the referents of moral terms have noticeably diverged following Moore's work on what 'good' means.[226] This divergence is largely due to the seemingly conflicting moral intuitions that we have, and the lack of a method to sort it out neatly. On the one hand, we make moral claims that appear to bear strong normative force. In fact, some 'fundamental' claims such as 'murder is wrong' or 'pleasure is good' just seem plain obvious. On the other hand, we observe that moral beliefs could differ in salient ways across communities. Moreover, a particular community's moral norms could change over time. Prominent examples include our change in attitude towards slave ownership, discrimination and intolerance, animal welfare. In some cases these changes occurred gradually over time

[222] Kant (1785), see Johnson and Cureton (2016) for a modern discussion.

[223] Sayre-McCord (2012)

[224] Alexander (2007), Bicchieri (2005)

[225] van Roojen (2018)

[226] Moore (1903)

as part of larger societal changes; at other times they could happen in a top-down manner by explicit argumentation, as has been the case with the influential contemporary ethicist Peter Singer. These changes could be evidence for the moral cognitivist for there being moral facts, which would drive these changes — our moral judgements *improve* over time, approaching some fixed ideal. For a non-cognitivist or a moral relativist, the changes constitute the evidence that there are no objective moral facts, and to say that our moral judgements 'improve' is only to beg the question of objectivity. There has been naturalistic attempts to ground such changes as due to our refined understanding of human welfare, as our scientific knowledge expands. However, it appears far-fetched to assume that a better understanding of the physical by itself translates into a better understanding of what aggregations of actions would provide better welfare. This is the case even if we do have natural explanation of certain moral norms, as is with the case of employing game-theoretic explanations.[227] This is because naturalistic explanations require extra components in order to dissolve the open question argument. The failure to reconcile our moral intuitions has made it difficult to settle on a semantics for moral claims.

My account above argued that the target of moral claims are domain dependent socio-linguistic realities. This is inadvertently a combination of treating moral claims as referring to conventions, while treating those conventions as *objective* target of description. I shall explain the source of their objectivity below.[228] Furthermore, this treatment allows for a game-theoretic explanation of moral intuitions. The difference between my picture and that of the naturalist is that with machine philosophy, we apply game-theoretic models to community-dependent intuitions rather than to a community-*independent* physical/economical reality. As for the open question argument. There is no question of why 'pleasure is good', since it is constitutive of the topic of normative ethics, as all intuitions are. That is distinct from the fact that we

[227] Skyrm (2003); Bicchieri (2005); Alexander (2007)

[228] This is in some sense Lewisian (Lewis 1969). However, my grounding for the objectivity of conventions is on statistical learning norms, rather than game-theoretic norms that deal with social systems. The difference is that the statical learning norms are common to all descriptive theorising, and governs objectivity at the very base level. In particular, *we don't need agents* for the objectivity of conventions (other than for a trivially enabling role – we need people to have communities).

can provide *explanations* of our moral intuitions. The real question is how we should weigh the different parameters of moral action, and how we should do feature selection.

4.4.2 How to Answer: 'What is a Moral Action?

I shall begin by setting out the general methodological framework in accord with what I have introduced in Chapter 2. I shall convey this in both standard philosophical language and in the language of machine philosophy, so as to clearly showcase the methodological mechanism.

First, I want to sketch the broader moral landscape I have in mind. The idea is akin to Carnapian linguistic frameworks.[229] In the case of moral values, I advocate what I call 'value frameworks'. This value framework fixes nE: the domain of normative ethics. In machine philosophy terms, we say that this value framework is the set of morally relevant intuitions in nE. In this case, the intuitions consist of value judgements. A value judgement is a value function over an action. Each action would span several parameters. For example, the action of murder would span the parameters of 'producing pain', 'vanquishing life' etc. These parameters are common to also other actions. We should note that 'producing joy' is the same parameter/on the same dimension as 'producing pain', they are simply on the opposite sides of the same scale. Furthermore, note that the parameters themselves could come in degrees: one could certainly produce different degrees of pain, and that would affect the evaluation of the action. Considering only the parameter of producing pain: Thrusting a sword through a pigeon is very bad; jabbing my friend with a pin is less bad. The claim that 'murder is bad' would be a negative evaluation of murder. It is therefore also a negative evaluation of 'producing pain', 'destroying life' etc. We can set the range of the function to the real interval between 0 and 1, where 0 indicates the worst possible moral action and 1 indicating the best.[230] We can judge different actions as

[229] Carnap (1928, 1934)

[230] Or from -1 to 1 if you prefer. The technicalities of this would depend on specific situations that arise in moral theorising

having degrees of ‘goodness’. For example, the action of ‘helping others’ might be assigned a value of 0.8 whereas the action of ‘murdering a pigeon’ might be assigned a 0. If we map our intuitions onto an *n*-dimensional space, where *n* is the number of morally laden parameters for actions, then we would have our raw data, which we would need to clean up with feature selection. See FIGURE 1 for the illustration of a toy value framework $V_1 = I_{nE}(X)$.[231] We have a diminishing return for pleasure, but an opposite evaluation scheme for harmlessness – it’s very bad even if you produce a medium amount of harm. Fairness is depicted on a linear scale. Of course these are just toy constructs. The actual distribution and the resulting function would depend on a study of our moral intuitions. In actual theorising, we need to keep each of the three parameters on separate dimensions. Of course we cannot visualise a multi-dimensional plane here.

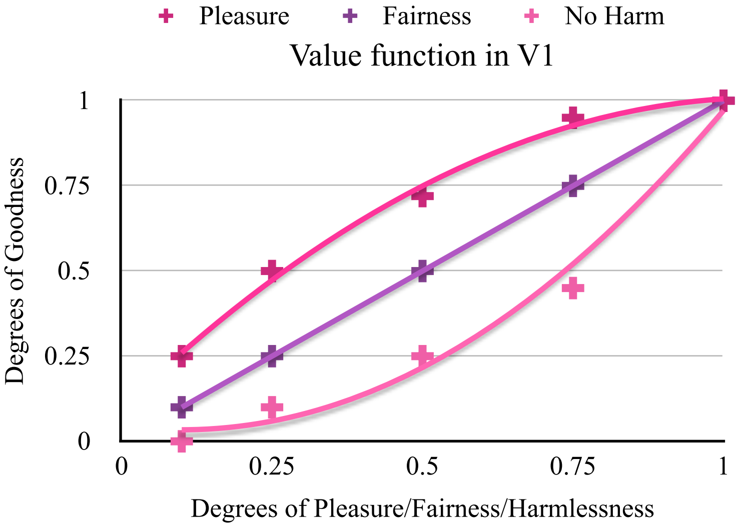

FIGURE 1

Now I can provide an account of how to justify a moral claim. Following our framework, and in line with Carnap, I claim that the truth and epistemic grounding of a moral theory $T_{V1}(X)$ is to be evaluated in relation to the framework V_1 in which it is

[231] V_1 is a shorthand for $I_{nE}(X)$, the latter follows our previous terminology of denoting the set of proper moral intuitions about X in nE.

asserted. Moreover, a moral claim *p* is to be evaluated against the best $T_V(X)$ we have such that V is the most coherent value framework we have.

In machine philosophy terms: Given a dataset X_{V1} of moral actions, which satisfy our moral intuitions V_1, we cannot evaluate $T_{V1}(X)$ other than against V_1 itself. The evaluation is achieved via measuring the bias and variance of $T_{V1}(X)$ against X_{V1}.

This is an instance of the general procedure of theorising I have introduced in Chapter 2. As a consequence, it makes no sense to compare moral claims *across* different value frameworks/domains. This is analogous to the claim that 'fish' in biology is a distinct topic from 'fish' in culinary art. However, value frameworks could be compared, just as the biologist could revise the ordinary language dataset of 'fish'. This is where I deviate from Carnap's attitude towards linguistic frameworks. For value frameworks, the codomain [0, 1] of the value function is shared across different frameworks. The differences between the frameworks reside in how much value they assign to each action. For example, in V_1, 0.5 degrees of 'producing joy' might be assigned a value of '0.5' degrees of goodness; whereas in V_2, 0.5 degrees of 'producing joy' might be assigned a value of 1. Moreover, the functions could differ: in V_1, we could have a logarithmic value function over 'producing joy', like in FIGURE 1, and in V_2, we might have a linear one. The evaluation of a value framework is via an assessment of statistical robustness, in the same spirit of why biologists had chosen to revise the dataset of 'fish', or why Clark and Chalmers had chosen to revise the notion of 'belief'.[232] In this capacity, we can say that some value frameworks are more 'coherent' than others. Here, 'coherence' is to be read as statistical uniformity, which allows for more robust and adequate models.[233] A value framework V_1 is better than a value framework V_2 just in case the most statistically adequate model for V_1 is more robust than that for V_2. When we revise our moral beliefs, we sometimes engage in this process of revising our value frameworks. For example, by abolishing slavery, we have modified, even if ever so slightly, our

[232] Clark & Chalmers (1998)

[233] We can think of this as a more uniform and robust account of 'consistent concepts' from conceptual engineering a la Scharp (2007)

intuitions on human rights and equality etc. i.e Permitting slavery produced a dataset that is less uniform than one that denounces slavery.

Finally we have the machinery for epistemically grounding a moral claim. Suppose we have a moral claim F(X), a value framework V such that V is the most up to date one ethicists have, and a $T_V(X)$ describing X_V. We say that the claim of F(X) is proper just in case F(X) satisfies $T_V(X)$ in a context C, where for C, F(X) functions just like $T_V(X)$. Alternatively, we can say that for an actual action *q*, we say that *q* is a moral action just in case that *q* is close enough or exemplifies[234] $T_V(X)$ for all the relevant practical ends of C.

For example, suppose that Mr. Mask is driving down a road, and in front of him lay five healthy young folks who happened to get stuck on the road. Alas, his brakes are bust, as he notices that on the sidewalk, there is one old grumpy granny drudging along. He could swerve onto he pavement, killing one but saving five. Suppose in addition that, unbeknownst to Mr. Mask, the car has a hidden function of lifting off on command. Now, suppose that the best theory of moral action we have is a version of act utilitarianism (AU) where the goal is to always maximise joy. If Mr. Mask ends up swerving and killing the one, but sparing the five, would he be acting morally in accord with AU? If we consider that it would have been possible for him to lift off, then no, he has failed to satisfy AU. However, he was ignorant of that option, so the moral context did not include that option. So for all intents and purposes, Mr. Mask did satisfy AU.

4.4.3 Epistemic Challenges and Descriptive Questions

Finally, I would like to provide an explanation of certain puzzling meta-ethical intuitions. I begin by addressing the question of where our moral intuitions come from.

[234] Depending on whether you use approximation or changing resolution for this situation.

Why do we hold moral beliefs? Where do the normative force of moral beliefs come from? Why do we have the intuition that morality is an objective matter? Why do different societies uphold different moral beliefs?

First, I want to address the question of 'from where do the moral judgements in my value frameworks arise'. I have claimed that our moral intuitions reflect our socio-linguistic reality with regard to moral parameters. However, this does not itself tell us why these socio-linguist realities are the way they are. There is no explanation. This enquiry is a purely empirical one. It is in the spirit of providing explanations for our observations beyond merely saying that our observations reflect physical reality. This task of providing explanations is orthogonal to the act of theorising. When I provide a theory $T_D(X)$, the target of observation or intuition would be X in D-specific terms. Following the same terminological practice, let's denote the set of intuitions about X in D as $I_D(X)$ (we can do the same for observations). Now, we can ask for an explanation of $I_D(X)$. This is a distinct project from providing and validating a $T_D(X)$, which aims to accurately describe $I_D(X)$. For example, the question of why (we have the intuition that) 'fairness is good' is a distinct question from trying to produce a theory in normative ethics that captures the intuition that 'fairness is good', which is again distinct from a theory in descriptive ethics that describes how we actually act. Moreover, an explanation for why 'fairness is good' cannot ground the statement that 'fairness is good'. The providing of an explanation for $I_D(X)$ lie squarely within meta-ethics, and we cannot have epistemic justification *across* domains. For example, Alexander has constructed game-theoretic ABMs on the parameter of fairness for actions to explain how a norm of fairness could arise in a social setting.[235] Such explanations seem to suggest a naturalised ethics. However, one must make a careful distinction here. The 'naturalised' entities here are not of moral claims – that is within our socio-linguistic reality, not physical reality. Rather, it is our moral intuitions themselves that are accounted for in naturalistic terms. This alone does not give us theories in normative ethics. After all, we could revise over our intuitions in the

[235] Alexander (2007)

process of theorising. The ultimate arbiter again is statistical norms for curve fitting. Now I can address the conflicting phenomena of moral relativity and moral normativity.

A.J. Ayer and the Vienna Circle claimed that moral statements have no cognitive content. What this means is that moral claims are meaningless, on the ground that any meaningful assertion must either refer to empirical facts, or logical facts.[236] Moral claims such as ‘murder is wrong’ seem to be in neither category. There also seem to be no referent for normative terms such as ‘wrong’. Hence non-cognitivists argue that moral claims do not refer to facts, but express preferences, emotions etc.[237] Hence for the non-cognitivists, moral relativity is not an issue.

It should be fairly easy to see now that moral relativity is an illusion caused by the fact that we do not know how to revise our value frameworks. i.e. We have a bunch of moral intuitions that can delineate a bunch of different value frameworks. Moreover, as machine learning has taught us, small individual differences in our value functions on actions can cause huge differences in the resulting datasets. That looks really bad. However, we shouldn’t let appearances daunt us. In a temporarily resurrected Strawsonian spirit: if a person holds a different set of moral beliefs from me, then that person would either be wrong (if her beliefs conflict with her moral intuitions), or that her *topic* of morality is simply a distinct one from mine. This also explains the meta-ethical intuition why it may sometimes seem that conflicting moral claims could not be reconciled, yet each party could provide good reasons. Indeed they *could be* good reasons, for in their respective framework, the claims could well fit their models that are statistically adequate *to the best of their abilities*.[238] This leads to the second phenomenon of why we seem to also think that moral claims have some objective grounding. The objective grounding is the epistemic requirement of coherence. In machine philosophy terms: statistical robustness and adequacy, which comes down to

[236] Ayer (1935)

[237] Stevenson (1937). See Nobis (2004) for commentary on Ayer and Stevenson

[238] In general, we haven’t been exactly sharp at recognising statistically bad models, since they never cared to map their moral intuitions onto an n-dimensional plane with parameters clearly laid out.

ensuring the descriptive accuracy of a theory in normative ethics. In line with my argument from 4.2.2, this descriptive adequacy contributes directly to the normative force of moral claims/theories. Moral claims have normative force crucially because moral claims are epistemically coherent. In machine philosophy terms, we say that a description $T_{nE}(X)$ of X_{nE}, where 'X' denotes moral actions has normative force *for moral human actions* partly because $T_{nE}(X)$ is descriptively adequate of X_{nE}.

Furthermore, moral revolutions or change, and the intuition that we have higher moral standards than our ancestors could be explained by the objective criteria we can use to compare two value frameworks a la 4.4.2. Moral revolutions are thus simply epistemic improvements. Hence moral progress is really no different to the revisions made in factual disciplines such as in the natural sciences, for moral progress is no more than the improvement on the coherence of our value systems, via better analyses of our moral data.

4.5 Conclusion

In this chapter, I have argued that there is no distinctive normative project. The normality of a theory is a relation between the theory and the target domain for which the theory is normative. In particular, the normative force of the theory depends on the descriptive accuracy of the theory in its own domain. Crucially, all theorising follow the norms of statistical learning.

To conclude, here is a summary of the moral landscape I have provided.

Let us denote a domain of normative ethics as nE; the set of intuitions/value framework in nE for X as $I_{nE}(X)$, abbreviated as V_1; the dataset of X in nE, which satisfies $I_{nE}(X)$ as X_{nE}; a theory of X in nE as $T_{nE}(X)$:

We say that $T_{nE}(X)$ is justified just in case it is descriptively adequate over X_{nE}.

Let another value framework be denoted by $I_{nE2}(X)$, abbreviated as V_2. We say that V_1 is more coherent than V_2 just in case V_1 enables a more robust description of X than V_2. In this case, we revise V_2 to V_1.

We say that $I_{nE}(X)$ is a reflection of the socio-linguistic reality for ethicists, in morally relevant terms. This has no bearing on whether $T_{nE}(X)$ is normative for a distinct domain Q. In particular, the normativity of $T_{nE}(X)$ for Q depends jointly on two propositions: 1. that $T_{nE}(X)$ is descriptively adequate of X_{nE}, and 2. there is a normative claim *N* such that $T_{nE}(X)$ contributes to the satisfaction of *N* in Q. *N* is itself a socio-linguistic fact in Q.

Therefore the project of explaining $I_{nE}(X)$ is itself a descriptive project on a certain aspect of our socio-linguistic reality.

And finally, we say that a moral claim F(X) in nE is proper or that an action *q* is moral just in case it exemplifies $T_{nE}(X)$ in a context C, the degree of approximation or resolution required determined by C. Voila!

Bibliography

1. Alexander, J. McKenzie (2007). The Structural Evolution of Morality. Cambridge University Press.
2. Allen, Donale (2003). The History of Infinity. *Texas A&M Mathematics.*
3. Audi, Robert (1983). Foundationalism, epistemic dependence, and defeasibility. *Synthese* 55 (1):119 - 139.
4. Austin, John L. (1956). A plea for excuses. *Proceedings of the Aristotelian Society* 57:1–30.
5. Axelrod, Robert (1980a), "Effective choice in the Prisoner's Dilemma", *Journal of Conflict Resolution* 24, 3–25.
6. —— (1984), *The evolution of cooperation*. New York: Basic Books.
7. Ayer, A. J. (1935) "The Elimination of Metaphysics" in *Language, Truth and Logic*. Penguin Classics; Edition: 26. April 2001.
8. Azzouni, Jody (2004). Theory, observation and scientific realism. *British Journal for the Philosophy of Science* 55 (3):371-392.
9. Baker, J. (1980). Scriabin's Implicit Tonality. *Music Theory Spectrum, 2*, 1-18.
10. Basheer, I. & Hajmeer M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods* 43, 3–31.
11. Bealer, George (1998). Intuition and the Autonomy of Philosophy, in *DePaul and Ramsey 1998*: 201–240.
12. Bicchieri, Cristina (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.
13. Black, Max. (Ed.). 1950. Philosophical Analysis. Englewood Cliffs: Prentice-Hall.
14. Blome-Tillmann, Michael (2015). Sensitivity, Causality, and Statistical Evidence in Courts of Law. *Thought: A Journal of Philosophy* 4 (2):102-112.
15. Boghossian, Paul (2009). Virtuous intuitions: comments on Lecture 3 of Ernest Sosa's A Virtue Epistemology. Philosophical Studies 144 (1):111-119.

16. BonJour, Laurence (1985). *The Structure of Empirical Knowledge*. Harvard University Press.

17. —— (2010). The Myth of Knowledge. *Philosophical Perspectives*, 24: 57–83.

18. Bovens, Luc & Hartmann, Stephan (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

19. Braddon-Mitchell, David & Nola, Robert (2009). Introducing the Canberra Plan. In David Braddon-Mitchell & Robert Nola (eds.), *Conceptual Analysis and Philosophical Naturalism*. MIT Press. pp. 1--20.

20. Broome, John (1994). Discounting the Future. *Philosophy and Public Affairs* 23 (2):128-156.

21. Cappelen, Herman (2012). *Philosophy Without Intuitions.* New York: Oxford University Press.

22. —— (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

23. Carnap, Rudolf (1934a) [1937], *Logische Syntax der Sprache*, Vienna: Springer. Translated by Amethe Smeaton as *The Logical Syntax of Language*, London: Routledge, 1937.

24. —— (1950a) [1956], Empiricism, Semantics, and Ontology, *Revue Internationale de Philosophie*, 4(11): 20–40. Reprinted in Carnap 1956b: 205–221.

25. —— (1950b). “On Explication”, Chapter 1 of *Logical foundations of probability.* University of Chicago Press, Illinois.

26. —— (1963). P.F. Strawson on Linguistic Naturalism. In Paul Arthur Schilpp (ed.), pp. 933–940.

27. —— [*Aufbau*] (1928) [1967], *Der logische Aufbau der Welt*, Berlin: Weltkreis. Second edition, Hamburg: Meiner, 1961. Translated into English as *The Logical Structure of the World*, Rolf A. George (trans.), Berkeley, CA: University of California Press, 1967.

28. Carter, J. Adam (2013). A problem for Pritchard’s anti-luck virtue epistemology. *Erkenntnis* 78 (2):253-275.

29. Chalmers, David (2010). *The Character of Consciousness*. Oxford University Press.
30. —— (2012). Constructing the World. Oxford University Press.
31. Chandler, Jake (2017). Descriptive Decision Theory. *The Stanford Encyclopaedia of Philosophy.*
32. Chang, H., 2005. A Case for Old-fashioned Observability, and a Reconstructive Empiricism. *Philosophy of Science*, 72(5): 876–887.
33. Clark, Andy & Chalmers, David J. (1998). The extended mind. *Analysis* 58 (1):7-19.
34. Cummins, Robert (1998). "Reflections on Reflective Equilibrium", in DePaul and Ramsey 1998: 113–128.
35. Dancy, J.; Sosa, E. & Steup, M. (ed.) (2010). *A Companion to Epistemology: Second Edition*. Wiley Blackwell.
36. de Ridder, Jeroen (2014). Epistemic dependence and collective scientific knowledge. *Synthese* 191 (1):1-17.
37. DePaul, Michael Raymond and William M. Ramsey (eds.), (1998), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry,* Lanham, MD: Rowman and Littlefield.
38. DeRose, Keith (1992). Contextualism and knowledge attributions. *Philosophy and Phenomenological Research* 52 (4):913-929.
39. Diallo, S. Y., Padilla, J. J., Bozkurt, I., & Tolk, A. (2013). Modeling and simulation as a theory building paradigm. In *Ontology, Epistemology, and Teleology for Modeling and Simulation* (pp. 193-206). Springer, Berlin, Heidelberg.
40. Domingos, Pedro (2012). A few useful things to know about machine learning. *Communications of the ACM.* Vol. 55 Issue 10: 78-87
41. Elgin, Catherine (2007). Understanding and the facts. *Philosophical Studies* 132 (1):33 - 42.
42. —— (2017). *True Enough*. Cambridge: MIT Press.

43. Enoch, David; Spectre, Levi & Fisher, Talia (2012). Statistical Evidence, Sensitivity, and the Legal Value of Knowledge. *Philosophy and Public Affairs* 40 (3):197-224.
44. Field, Hartry H. (1994). Deflationist views of meaning and content. *Mind* 103 (411):249-285.
45. Field, Hartry H. (2008). *Saving Truth From Paradox*. Oxford University Press.
46. Frigg, Roman.(2009). Models in Physics. *In The Routledge Encyclopedia of Philosophy*. Taylor and Francis. Retrieved 2 Apr. 2021, from https://www.rep.routledge.com/articles/thematic/models-in-physics/v-1.
47. Gert, Bernard & Gert, Joshua (2016). The definition of morality. S*tanford Encyclopedia of Philosophy*.
48. Gettier, Edmund (1963). Is Justified True Belief Knowledge? *Analysis* 23 (6):121-123.
49. Goldman, Alvin I. (1999a). A Priori Warrant and Naturalistic Epistemology, in James E. Tomberlin (ed.), *Philosophical Perspectives*, 13: 1–28.
50. —— (1999b). *Knowledge in a Social World*. Oxford University Press.
51. —— (2007). Philosophical intuitions: Their target, their source, and their epistemic status. *Grazer Philosophische Studien* 74 (1):1-26.
52. —— (2010). Systems-oriented social epistemology. *Oxford Studies in Epistemology* 3:189-214.
53. Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press.
54. Günther, Mario (2021). Epistemic Sensitivity and Evidence, *Inquiry: An Interdisciplinary Journal of Philosophy*.
55. Guyon, I. & Elisseeff A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 1157–1182.
56. Hannon, Michael (2019). *What's the Point of Knowledge?* Oxford University Press.
57. Hardwig, John (1985). Epistemic dependence. *Journal of Philosophy* 82 (7):335-349.
58. Harman, Gilbert (1986). *Change in View*. MIT Press.

59. Haslanger, Sally (2000). Gender and race: (What) are they? (What) do we want them to be? *Noûs* 34 (1):31–55.

60. Hawke, Peter (2017). Theories of Aboutness, *Australasian Journal of Philosophy*

61. Hensen, B., Bernien, H., Dréau, A. *et al.* (2015) Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526,** 682–686. https://doi.org/10.1038/nature15759

62. Hetherington, Stephen (2005). Knowing (How It Is) That P: Degrees and Qualities of Knowledge. *Veritas: Revista de Filosofia da PUCRS* 50 (4):129-152.

63. —— (2011). How to Know: A Practicalist Conception of Knowledge. John Wiley & Sons.

64. Hudson, Robert (2014). Saving Pritchard's anti-luck virtue epistemology: the case of Temp. *Synthese* 191 (5):1-15.

65. Hume, David (1739). *A Treatise of Human Nature*.

66. Jackson, Frank (1986). What Mary Didn't Know. *Journal of Philosophy* 83 (5):291-295.

67. —— (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.

68. —— (2000). Reply to Yablo: What do we communicate when we use ethical terms. *Philosophical Books* 41:24-29.

69. Jesseph, D.M. (1998). Leibniz on the Foundations of the Calculus: The Question of the Reality of Infinitesimal Magnitudes. *Perspectives on Science 6*(1), 6-40.

70. Johnson, Robert and Adam Cureton, "Kant's Moral Philosophy", The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/kant-moral/>.

71. Kahneman, Daniel (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus & Giroux.

72. Kant, Immanuel (1785). *Groundwork of the Metaphysic of Morals*.

73. Kelp, Christoph (2013). Knowledge: The Safe-Apt View. *Australasian Journal of Philosophy 91 (2):265-278*.

74. Kim, Minsun & Yuan, Yuan (2015). No cross-cultural differences in the Gettier car case intuition: A replication study of Weinberg et al. 2001. *Episteme* 12 (3):355-361.

75. Koller, D. & Sahami M. (1996). Toward optimal feature selection. In L. Saitta (Ed.), *Proceedings of the Thirteenth International Conference on Machine Learning (ICML),* pp. 284–292.

76. Kripke, Saul A. (1975). Outline of a theory of truth. *Journal of Philosophy* 72 (19):690-716.

77. —— (1980). *Naming and Necessity*. Harvard University Press.

78. Kuhn, Thomas .S. (1962). *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press, reprinted,1996.

79. Leitgeb, Hannes (2013) Scientific Philosophy, Mathematical Philosophy, and All That, *Metaphilosophy*, 44(3): 267–275.

80. Leitgeb, Hannes & Carus, André (2020) “Rudolf Carnap", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/sum2021/entries/carnap/ >.

81. Lewis, David (1969). *Convention: A Philosophical Study*. Wiley-Blackwell.

82. —— (1983). *Philosophical Papers: Volume I*, New York: Oxford University Press.

83. —— (1988a). Relevant Implication, *Theoria* 54/3: 161–74.

84. —— (1988b) Statements Partly About Observation, *Philosophical Papers* 17/1: 1–31.

85. —— (1996). Elusive knowledge. *Australasian Journal of Philosophy* 74 (4):549 – 567.

86. Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2020). Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895.*

87. Ludlow, Peter (2005). Contextualism and the new linguistic turn in epistemology. In *Gerhard Preyer & Georg Peter (eds.), Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford University Press. pp. 11–51.
88. Ludwig, Kirk, 2007, "The Epistemology of Thought Experiments: First Person versus Third Person Approaches", *Midwest Studies in Philosophy*, 31: 128–159.
89. Machery, Edouard ; Stich, Stephen ; Rose, David ; Chatterjee, Amita ; Karasawa, Kaori ; Struchiner, Noel ; Sirker, Smita ; Usui, Naoki & Hashimoto, Takaaki (2015). Gettier Across Cultures. *Noûs*:645-664.
90. Maddy, Penelope (2007). *Second Philosophy: A Naturalistic Method*. Oxford University Press.
91. Maher, Patrick (2007). Explication Defended. Studia Logica 86 (2):331-341.
92. Malcolm, Norman (1949) Defending Common Sense. *Philosophical Review* 58, 201- 220.
93. Manley, David; Dunaway, Billy & Edmonds, Anna (2013). The Folk Probably do Think What you Think They Think. *Australasian Journal of Philosophy* 91 (3):421-441.
94. Mayo-Wilson, C. (2014). Reliability of Testimonial Norms in Scientific Communities. *Synthese* 191, 55–78.
95. Mayo-Wilson, C., K. Zollman, and D. Danks (2011). The Independence Thesis: When Individual and Social Epistemology Diverge. *Philosophy of Science* 78(4), 653–677.
96. Moore, G. E. (1903). *Principia Ethica*. Dover Publications.
97. —— (1925). A defence of common sense. In *J. H. Muirhead (ed.), Contemporary British Philosophy, Second Series.* George Allen and Unwin.
98. —— (1939). Proof of an external world. *Proceedings of the British Academy* 25 (5):273–300.
99. Muldoon, R., C. Lisciandra, and S. Hartmann (2014). Why are there descriptive norms? because we looked for them. *Synthese* 191, 4409–4429.

100.Nagel, Jennifer (2012). Intuitions and Experiments: A Defense of the Case Method in Epistemology. *Philosophy and Phenomenological Research* 85 (3):495-527.

101.Nobis, Nathan (2004). Ayer and Stevenson's Epistemological Emotivisms. *Croatian Journal of Philosophy* 4 (1):59-79.

102.Nozick, Robert (1974). *Anarchy, State, and Utopia.*

103. —— (1981). *Philosophical Explanations.* Harvard University Press.

104.Oakley, Tim (2006). A Problem About Epistemic Dependence. In *Stephen Hetherington (ed.), Aspects of Knowing. Elsevier Science.* pp. 17.

105.Olsson, E.J. (2015). Gettier and the method of explication: a 60 year old solution to a 50 year old problem. *Philos Stud* **172,** 57–72 (2015)

106.Pappas, George (2014). Internalist vs. externalist conceptions of epistemic justification. *Stanford Encyclopedia of Philosophy*.

107.Peirce, C. S. (1905). What Pragmatism is., *The Monist*, Volume 15, Issue 2

108.Pritchard, Duncan (2009). *Knowledge*. In John Shand (ed.), Central Issues of Philosophy. Wiley-Blackwell.

109. —— (2010). Cognitive ability and the extended cognition thesis. *Synthese 175 (1):133 - 151.*

110. —— (2012). Anti-Luck Virtue Epistemology. *Journal of Philosophy 109 (3):247-279.*

111. —— (2015). Epistemic dependence. *Philosophical Perspectives 29 (1):305-324.*

112.Pritchard, Duncan & Bernecker, Sven (2011). *The Routledge Companion to Epistemology.* Routledge.

113.Proietti, M., Pickston, A., Graffitti, F., Barrow, P., Kundys, D., Branciard, C., ... & Fedrizzi, A. (2019). Experimental test of local observer independence. *Science advances*, *5*(9), eaaw9832.

114.Pust, Joel (2001). Against Explanationist Skepticism Regarding Philosophical Intuitions, *Philosophical Studies*, 106(3): 227–258

115. —— (2017). Intuition. *Stanford Encyclopedia of Philosophy*.

116.Resnik, Michael (1987). *Choices: An Introduction to Decision Theory*. Univ of Minnesota Press.

117.Rottschaefer, William (1976). Observation: Theory-Laden, Theory-Neutral or Theory-Free? *Southern Journal of Philosophy* 14 (4):499-509.

118.Russell, Bertrand (1910). Knowledge by Acquaintance and Knowledge by Description. *Proceedings of the Aristotelian Society*, 11: 108–128.

119.Ryle, Gilbert. 1961. "Use, Usage and Meaning." *Proceedings of the Aristotelian Society, Supplementary Volumes* 35, 223-230.

120.Sayre-McCord, Geoff (2012). "Metaethics", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2014/entries/metaethics/>

121.Scharp, Kevin (2007). Replacing truth. *Inquiry: An Interdisciplinary Journal of Philosophy* 50 (6):606 – 621.

122.Schoeman, F. (1987). Statistical vs. Direct Evidence. *Noûs, 21*(2), 179-198. doi:10.2307/2214913

123.Seyedsayamdost, Hamid (2015). On Normativity and Epistemic Intuitions: Failure of Replication. *Episteme* 12 (1):95-116.

124.Shapiro, Stewart (2005). Logical consequence, proof theory, and model theory. In *The Oxford Handbook of Philosophy of Mathematics and Logic*. Oxford University Press. pp. 651–670.

125.Skyrms, Brian (2003). *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

126. —— (2010). *Signals: Evolution, learning, and information*. Oxford University Press

127.Sosa, Ernest (1998). Minimal Intuition, in *DePaul and Ramsey* 1998: 257–270.

128. —— (1999). How to Defeat Opposition to Moore. *Philosophical Perspectives* 13 (s13):137-49.

129.Stenning, Keith & van Lambalgen, Michiel (2008). *Human reasoning and cognitive science*. Boston, USA: MIT Press.

130.Sterelny, Kim. (2012). *The evolved apprentice*. MIT press

131.Steuer, R.H. (1985). Artificial Distintegration and the Cambridge-Vienna Controversy, in P. Achinstein and O. Hannaway (eds.), *Observation, Experiment, and Hypothesis in Modern Physical Science,* Cambridge: MIT Press, 1985, 239–307

132.Steup Matthias; Turri, John, & Sosa Ernest, (eds.) (2013). *Contemporary Debates in Epistemology*. Wiley-Blackwell; 2 edition.

133.Stevenson, Charles Leslie (1937). The Emotive Meaning of Ethical Terms. *Mind* 46 (181):14-31.

134.Stich, Stephen (2013). Do Different Groups Have Different Epistemic Intuitions? A Reply to Jennifer Nagel. *Philosophy and Phenomenological Research* 87 (1):151-178.

135.Strawson, Peter F. (1950). On Referring. *Mind* 59, 320-344.

136. —— (1959). *Individuals: An Essay in Descriptive Metaphysics.* London: Methuen.

137. —— (1963). Carnap's Views on Conceptual Systems versus Natural Languages in Analytic Philosophy. In Paul Arthur Schilpp (ed.), pp. 503–518.

138.Sucholutsky, I., & Schonlau, M. (2020). 'Less Than One'-Shot Learning: Learning N Classes From M< N Samples. *arXiv preprint arXiv:2009.08449.*

139.Tarski, Alfred (1933). The Concept of Truth in the Languages of the Deductive Sciences (Polish), *Prace Towarzystwa Naukowego Warszawskiego, Wydziall III Nauk Matematyczno-Fizycznych 34*, Warsaw. (Expanded English translation in Tarski 1983, 152-278).

140. —— (1944). The Semantic Conception of Truth, *Philosophy and Phenomenological Research* 4(3), 341-376.

141.Tversky, Amos & Kahneman, Daniel (1974). Judgment under Uncertainty: Heuristics and Biases. *Science* 185 (4157):1124-1131.

142.van Inwagen, Peter (1997). Materialism and the Psychological-Continuity Account of Personal Identity, *Philosophical Perspectives*, 11: 305–319.

143. van Roojen, Mark, "Moral Cognitivism vs. Non-Cognitivism", *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/moral-cognitivism/>.

144. Wason, P. C. (1968). "Reasoning about a rule". *Quarterly Journal of Experimental Psychology*. **20** (3): 273–281.

145. Wason, P. C. & J. Evans (1983). *Thinking and reasoning: Psychological approaches*. Ed. Evans, J. Routledge.

146. Weatherson, Brian (2003). What good are counterexamples? *Philosophical Studies* 115 (1):1-31.

147. Weinberg, Jonathan M; Nichols, Shaun & Stich, Stephen (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29 (1-2):429-460.

148. Weisberg, M. & R. Muldoon (2009). Epistemic Landscapes and the Division of Cognitive Labor. *Philosophy of Science* 76(2), 225–252.

149. Weisstein, Eric W. "Circle." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/Circle.html. Accessed 9 Oct 2018.

150. Westergaard, Peter (1968). Conversation with Walter Piston. *Perspectives of New Music* 7, no.1 (Fall-Winter) 3–17

151. Wheeler, Gregory (2016). Machine Epistemology and Big Data. In L. McIntyre and A. Rosenberg (Eds.), *The Routledge Companion to Philosophy of Social Science*. Routledge

152. Wigner, Eugene P. (1961), Remarks on the Mind-body Question, in: I. J. Good, *The Scientist Speculates*, London, Heinemann

153. Williamson, Timothy (2000). *Knowledge and Its Limits*. Oxford University Press.

154. —— (2004). Philosphical 'intuitions' and scepticism about judgement. *Dialectica* 58 (1):109–153.

155. —— (2007). *The Philosophy of Philosophy*. Blackwell.

156. Wittgenstein, Ludwig (1953). *Philosophical Investigations.* Wiley-Blackwell.

157. Zollman, K. (2010). The Communication Structure of Epistemic Communities. In *A. Goldman and D. Whitcomb (Eds.), Social Epistemology: Essential Readings.* Oxford University Press.