Mistaken Identity: Conceptual Change, Pragmatism, and the Truth About Gender

by

Kevin Craven

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2023

Doctoral Committee:

Professor Ishani Maitra, Chair
Professor Elizabeth Anderson
Professor Susan Gelman
Professor Eric Swanson

Kevin Craven

kcraven@umich.edu

ORCID ID: 0009-0005-6653-7139

Preface

> "*To be specific, it [skepticism about cognition of things as they are] takes for granted certain ideas about cognition as an* instrument *and as a* medium, *and assumes that there is a* difference between ourselves and this cognition*."*
>
> - G.F.W. Hegel, *The Phenomenology of Spirit*, 47

I began thinking about gender in the fall of 2007 when, during a late-night dormitory conversation at Brandeis University, a friend asked if I'd heard the recent news of a pregnant man. I was shocked to hear that scientists had already worked out how to implant a working uterus in someone born without one. When it was clarified that this was a trans man who had been born with a uterus, I rolled my eyes. My friend was playing word games, I thought. If 'man' refers to anyone who wishes to be counted as a man, then there's nothing surprising about a pregnant 'man.' But gender-terms *don't* in fact refer to people on that basis — they classify people on the basis of biology. While my liberal sensibilities allowed me to see that anyone should be able to adopt whatever symbolically gendered behaviors they wish, I didn't see why any of that should lead us to change our views on what a man is or what 'man' means.

Things began to change when, later in my college career, I became friends with a trans man. I couldn't be so blithe about meanings when confronted by a flesh-and-blood person for whom I cared and for whom these meanings could be a matter of life and death. My viewpoint shifted: while the term 'man' in English conventionally refers adult human males, respect for my

friend demanded that I refer to him in the way he wished. I decided that I cared about people more than I cared about words.

This position, while not as obviously transphobic as my previous one, still felt wrong. While the obligation to respect my friend's identity was obvious, I didn't feel like I was fully seeing that obligation through. I may have scrupulously *said* the right things, but I knew that I wasn't *seeing* my friend as he saw himself. He knew himself to be a man. Unless and until I could join him in that knowledge, my utterances would be hollow.

This dissertation began as an effort to make good on my commitments to that friend, and to all the trans people in my life. In 2012 and 2013, while completing an MA in philosophy at Brandeis, I dove deep in the works of both Sally Haslanger and Eli Hirsch. I was fascinated by the possibility of radically different conceptual schemes, and the questions surrounding whether and in what ways one scheme could be better than another. While these questions are as old as philosophy itself, my interest in them was driven especially by my desire to make sense of how respect or disrespect for trans identities could be manifest in our concepts, and how these concepts could change to better realize our commitments.

I've arrived at a view that I think of as expressivist, though it's explicitly in the vein of Brandom's inferentialism rather than the attitudinal quasi-realism of Gibbard. I think such a theoretical paradigm allows us to make sense of the idea that we can have an obligation to think of others in a certain way — to adopt certain concepts or reject others. The reasons for this are many and will, I hope, become clear in the course of the dissertation. Here at the outset, I'll highlight something that only became clear to me as the project neared its end.

The activity of critiquing our own concepts raises a paradox. On the one hand, we don't want to simply take our concepts for granted — that's the whole point of critique, after all. On

the other hand, any such critique must be made *in terms of our own concepts*. We cannot 'step outside' our own conceptual schemes to evaluate our own thought from a God's-eye perspective; and even if we could, such evaluations would be irrelevant for us as soon as we returned from the perspective of God to our own everyday lives.

This paradox has been noted many times in the history of philosophy. It was central to Hegel's rejection of Kant's transcendental idealism: the idea of an uncognizable thing-in-itself is self-defeating. And it arises in different guises in each of my three chapters. When thinking of 'conceptual engineering' in general, we run into the problem that deep conceptual changes cannot be undertaken deliberately because doing so would require a conceptual mastery that we, by hypothesis, don't have. It arises for feminist theories of gender as a tension for immanent critique — a tension between uncritically reinforcing existing gendered practices and adopting an idealist view that utterly loses contact with them. And when we start thinking about the value of autonomy, it arises again as a tension between our desire to define ourselves on our own terms and the fact that this self-definition is meaningful only against a background of social meanings.

Like Hegel, I end up with a picture on which the transformation of our viewpoint is dialectical. That is, it's something that happens *from within* our existing viewpoints and practices. Our existing viewpoints and practices are unstable — they generate contradictions and problems for us that we must try to resolve. These resolutions require transformations of our perspectives and, indeed, of ourselves. None of this requires us to take up a viewpoint utterly outside our existing concepts; rather, it involves listening to one another and coming to see — to *feel* — how those concepts fail us.

A final note. I am, at least for now, a cis man. It might seem suspicious for me to even write a dissertation focusing on transgender issues, and doubly so to begin with an introduction

about how transphobic I was in college. Who am I to claim any kind of epistemic or practical authority in this domain? My only answer is that I have at least the minimal standing that any person living in a gender-structured society has to question the gendered practices that shape all our lives. I focus on transgender identities because coming to know and to learn from trans people has been so crucial to my understanding of gender in general and of myself as a gendered being. It's not that I have any special authority here, nor is it that trans identities are in special need of either justification or explanation. It's that everyone's self-understanding is constitutively tied to their understanding of everyone else. So, I hope that any readers for whom this project feels presumptuous will do me the favor of bearing in mind that it's just as much an effort to make sense of myself as it is to make sense of anyone else. And, as I'll emphasize in chapter three, my efforts to make sense of my trans friends on their own terms have led to a transformation in my understanding of not only myself but of the entire social world.

In that spirit, I'd like to acknowledge a few of the countless people who've taught me over the years. I thank my many academic mentors including Eli Hirsch, Marion Smiley, Brett Sherman, Berislav Marusic, Ishani Maitra, Elizabeth Anderson, Eric Swanson, Susan Gelman, and David Manley. I thank my fellow student philosophers including Tanya Kostochka, Mac Mackinnon, Rebecca Harrison, Emma Hardy, Filipa Melo Lopez, Kevin Blackwell, Margot Witte, Jason Lee Byas, Mercedes Corredor, Gillian Gray, Josh Hunt, Alice Kelley, Cameron McCulloch, Sumeet Patwardhan, Eduardo Martinez, Caroline Perry, Ariana Peruzzi, Joe Shin, Alvaro Sottil de Aguinaga, Angela Sun, and Elise Woodard. Finally, I thank other professionals who took the time to discuss these issues with me despite having established careers and no obligation to spend time mentoring me, including Maegan Fairchild, Sally Haslanger, and Katherine Jenkins.

I thank my mother for always being willing to fight McDonald's employees over the fact that her son wanted a Barbie rather than Hot Wheels for his Happy Meal toy. I thank my father for giving so much of himself, loving unconditionally, and keeping an open heart. I thank every friend and acquaintance with whom I've spoken, often inarticulately, on these issues. I thank them for both their patience and their impatience — and I think my sister for both at turns.

I thank my partner Aesha Mustafa, whom I met shortly after moving to Ann Arbor in 2014 and without whose loving care I certainly wouldn't have finished this project. I thank our beloved dog Bailey, who kept me going outside regularly during the pandemic and whose fur remains soft however hard the rest of the world may get.

Finally, I thank all the women and LGBT people who have, bravely and brilliantly, staked their claims to a place in this world. Even if I don't know you, I wouldn't be who I am without you.

Table of Contents

Abstract

This dissertation aims to contribute to two recently burgeoning literatures in philosophy: that surrounding *conceptual engineering* and that surrounding the metaphysics of gender. It begins with a criticism of the recent conceptual engineering literature, arguing that the idea of rationally warranted conceptual change raises irresolvable puzzles as long as the kind of rationality at work is assumed to be the familiar type of instrumental rationality. I then argue that another kind of rationally warranted conceptual change — one already investigated empirically by developmental psychologists such as Susan Carey — can provide a model for conceptual changes involving non-instrumental modes of reasons-responsiveness.

Chapter 2 argues that feminist approaches to the metaphysics of gender and the ethics of gender-ascription would be well served by exploring an anti-representationalist, pragmatist theory of linguistic meaning. The argument here is that such a pragmatist approach posits a constitutive relationship between the truth of a gender-ascription and the normative commitments it expresses. Theory that thinks of itself in this way can regard itself as a kind of *immanent critique*.

Finally, chapter 3 sketches how we might think about the ethics of gender-ascription and — if chapter 2 is correct about the constitutive relationship between these — the metaphysics of gender. It's argued that the fundamental values of freedom and equality, and the nature of social identity as a call for recognition, militate in favor of respecting transgender identities and disavowing hegemonic conceptions of gender.

Chapter 1 On The Very Idea of Conceptual Engineering

> "*Concepts are therefore grounded in the spontaneity of thinking, as sensible intuitions are grounded in the receptivity of impressions. Now the understanding can make no other use of these concepts than that of judging by means of them.*"
>
> - Immanuel Kant, *Critique of Pure Reason*, 205

> "*When I first started cutting up oxen, all I saw for three years was oxen, and yet still I was unable to see all there was in an ox. But now I encounter it with the imponderable spirit in me rather than scrutinizing it with the eyes. For when the faculties of officiating understanding come to rest, imponderable spiritlike impulses begin to stir, relying on the unwrought perforations. Striking into the enormous gaps, they are guided by those huge hollows, going along in accord with what is already there and how it already is*."
>
> - Zhuangzi, 29-30

Donald Davidson famously argued that no sense can be made of the idea of a conceptual scheme. This argument was made in response to a cluster of views like those of Thomas Kuhn according to which different individuals or communities can have sets of concepts which differ so radically from one another as to be tantamount to inhabiting "different worlds."[1] Kuhn expressed this by saying that the ideologies of competing scientific paradigms are *incommensurable*: there is no way to adjudicate the dispute by appealing to a vocabulary or to a set of facts or phenomena to which all parties can agree.

Davidson, in response, offered a kind of transcendental *reductio*. To establish that two conceptual schemes are incommensurable with one another, we first would have to specify *in*

[1] Kuhn (1996), 150

terms of our own concepts what the two schemes amount to, or what concepts they contain.[2] But this would only prove that the two schemes have a common measure in our own concepts, in which case they are not incommensurable. "Different points of view make sense, but only if there is a common co-ordinate system of which to plot them; yet the existence of a common system belies the claim of dramatic incomparability."[3]

I argue that something like this problem arises for the possibility of rationally warranted conceptual change, and therefore for the recently burgeoning literature on conceptual engineering. Fundamental to the idea of conceptual engineering is the claim that we can be rationally warranted in changing our concepts for a variety of reasons. The problem arises when we ask what the relationship between pre-engineering concepts and post-engineering ones is supposed to be. If the new conceptual scheme could have been articulated in terms of the old one, then the change from one to the other should not produce any philosophically interesting benefits. And while a transition to a more radically new conceptual scheme may well have the practical effects the conceptual engineer wants, its very incommensurability with our old scheme means that the transition cannot occur as a deliberate act of conceptual engineering. Or so I will argue.

Call this problem for conceptual engineering *The Davidsonian Dilemma.* In what follows, I begin by developing a partial taxonomy of conceptual engineering projects, distinguishing in particular between projects that aim at what I'll call *shallow* and *deep* conceptual change. I argue

[2] Of course, the way in which I've put things here is too quick. Translation is not obviously the only way in which to compare conceptual schemes. For my purposes, and for reasons that I hope to make clear, I think Davidson's point applies to conceptual engineering — arguably it applies more clearly here than to the cases to which Davidson himself applied it.
[3] Davidson (2001), 184

that each of these types of conceptual change runs into a horn of the dilemma: shallow change is insignificant, while deep change cannot be deliberately engineered.

The argument of the paper runs as follows. I begin with a discussion of the basic idea of conceptual engineering as rationally warranted conceptual change. I then identify some core aspects of a certain familiar type of rationality: means-ends or instrumental rationality. My main negative argument aims to show that all these core aspects of instrumental rationality generate problems for conceptual engineering. After establishing that we cannot — at least in many cases — view conceptual change as a kind of instrumentally rational choice, I begin to sketch an alternative picture on which it instead involves a kind of non-instrumental reasons-responsiveness. Here I draw on the work of psychologist Susan Carey, whose research on conceptual development in childhood and adolescence provides an example of the kind of non-instrumental conceptual learning that I think is philosophically interesting. Finally, I gesture toward how something like Carey's picture of conceptual change could be extended to other cases in philosophy, including ethically and politically fraught concepts like *terrorism* and *chastity*.

1.1 Conceptual Engineering: Intensions as Instruments

Part of what is at issue here is what conceptual change amounts to in the first place. And that, of course, raises the question of what concepts are in the first place. While there is no clear consensus answer to either of these questions, by far the most common approach in the literature to date is to view concepts as (or as corresponding to) *intensions.*[4] Conceptual change is a kind of

[4] Note exceptions to this e.g. in Eklund (2021). I personally don't believe in the intensions view - indeed part of the point of the dissertation is to argue against it.

change in what intensions serve as the contents of our thoughts and utterances. What kind of change? Here things begin to diverge quite a bit. I'll consider a few different examples.

While I'll mostly be using the term 'conceptual engineering' for the target of my arguments here, it also goes under the names 'conceptual ethics' and 'amelioration.' A brief discussion of each of these and how they relate to each other. I'll start with amelioration, since it's the earliest reference I'll discuss here and also what brought me to this topic in the first place.

Haslanger (2000) developed the idea of an *ameliorative project*. Haslanger's aims in developing this idea were varied and complex, but among them was to push back on the idea that the only thing the philosopher can do with concepts is to analyze them and leave them as they're found.[5] She argued that, besides purely descriptive projects in conceptual analysis or lexical semantics, we can also view our concepts through a critical lens:

> *On this [ameliorative] approach the task is not to explicate our ordinary concepts; nor is it to investigate the kind that we may or may not be tracking with our everyday conceptual apparatus; instead we begin by considering more fully the pragmatics of our talk employing the terms in question. What is the point of having these concepts? What cognitive or practical task do they (or should they) enable us to accomplish? Are they effective tools to accomplish our (legitimate) purposes; if not, what concepts would serve these purposes better?*[6]

While the meanings our words *actually* have may be determined by our linguistic conventions, there is a further question of what meanings they *ought* to have.She famously applied this to concepts of gender and race, urging that — in one way or another — adopting social-constructionist concepts of these human kinds would conduce to the realization of social justice. For present purposes, though, the important thing to focus on is that Haslanger

[5] Cf. Haslanger (2020)
[6] Haslanger (2012), 224

characterized concepts as *tools* that we apply for *purposes*, and urged that we can legitimately aim to improve those tools. Others in this ameliorative tradition include Katharine Jenkins and Kate Manne.[7]

The idea of concepts as tools lends itself naturally to the phrase I'll use: *conceptual engineering*. Cappelen and Plunkett (2020) gives a pat definition of this term:

> ***Conceptual engineering*** *= (i) The assessment of representational devices, (ii) reflections on and proposal for how to improve representational devices, and (iii) efforts to implement the proposed improvements.*[8]

As the above definition suggests, 'conceptual engineering' is a term with potentially very broad application. Actual projects that have been described with this term range from Haslanger-esque ones concerned with definitions of human kind terms[9] to revisionary analyses of the concept of truth.[10] As we'll see later in this chapter, it can include some almost head-scratchingly trivial instances of representational improvement as well.

Finally, we have *conceptual ethics*. This term is meant to encompass "a range of normative and evaluative issues about thought, talk, and representation." The term was popularized by Burgess and Plunkett (2013) and has mostly been championed by them since. How exactly it relates to the other labels, and to conceptual engineering in particular, has been a topic of some disagreement. While some explicitly use the terms 'conceptual ethics' and 'conceptual engineering' — and even 'amelioration' — interchangeably, Burgess and Plunkett themselves have argued that there's a meaningful distinction to be drawn between the three

[7] Cf. Jenkins (2016) and Manne (2018)
[8] Cappelen and Plunkett (2020), 3
[9] Cf. Dembroff (2016)
[10] Scharp (2013)

notions. They view conceptual ethics as mapping roughly onto (i) in the tripartite definition of engineering quoted above.[11]

While much could be said about the subtleties of the relations between these terms and the kinds of inquiry they pick out, I'll mostly abstract away from these. I want to speak about a particular problem or cluster of problems having to do with the idea of rationally warranted conceptual change. This problem will arise in different ways for different ameliorative/engineering/ethics projects, and may not arise for some of them at all. But I think it at least presents a puzzle for any philosopher who sees themself as engaging in a project of reflectively improving our conceptual schemes.

But what kinds of improvements are we talking about here? Just as diverse as the ways of talking about these projects are the goals at which they're aimed. As already mentioned, some of them aim at the realization of social justice. Others are aimed at making inquiry go better, either by better realizing some epistemic function or by solving a problem that our current concepts cannot.[12] Finally, concepts could be constitutively linked to things we care about in such a way that conceptual change takes on a transformative quality. This is hinted at in Burgess & Plunkett (2013): "our conceptual repertoire determines not only what we can think and say but also, as a result, what we can do and who we can be."[13]

Hence, conceptual engineering is both very broad and potentially very ambitious. How we think and speak influences everything from inquiry, to social structures, to our very sense of self. Underexamined, though, is *how* conceptual engineering is supposed to achieve its

[11] Burgess and Plunkett (2020)
[12] Cf. Brigandt (2010) for an articulation of the idea of an epistemic function.
[13] Burgess and Plunkett (2013a), 1091

ambitions. The next section is devoted to examining a specific way in which this may be understood: on the traditional model of instrumental rationality.

1.1.1 What is Instrumental Rationality?

I'm arguing that the literature on conceptual engineering hasn't adequately addressed the question of how conceptual change can be regarded as a rationally warranted activity. This is mostly because few in the literature have seriously grappled with the question or noticed the puzzles it raises (Eklund is at least a partial exception to this).

Much of the literature tacitly assumes that these conceptual choices will conform to the model of *instrumental rationality* — i.e. of the kind of rational choice that involves the bringing about of desired or otherwise worthy outcomes. In this section I'll say what I hope are some uncontroversial things about instrumental rationality. The aim is to highlight some relevant features that I think become problematic when we try to apply this model of rational choice to conceptual change.

There are multiple overlapping but distinct traditions when it comes to how to characterize instrumental rationality. A familiar one characterizes it in terms of the formal (but usually non-mathematical) relations between means and ends, e.g. "If you will an end you are rationally required to will the necessary means." This tradition stretches back at least to Aristotle, though contemporary disputes about the relations between means and ends draw more on Hume and Kant. I'm going to cut through all of these controversies and highlight this: a presupposition of all these disputes is that we characterize agents as adopting *means* to their *ends*. The end is an outcome or state of affairs that the agent has reason to bring about; the means is an action that

the agent has reason to perform in virtue of its bearing some appropriate (necessitating, probabilizing, etc.) relation to the end.

In this respect, traditional discussions of means-ends rationality differ little from more recent developments in mathematical decision theory. While decision theorists disagree with one another on a number of details, all approaches have one thing in common: what an agent ought to do in a given situation is a function of *preferences* over *prospects*.[14] The agent has a utility function defined over a set of possible world-states; this supplies the agent with their ends. We then imagine that the agent has a range of options, or possible actions, available to them; these take the place of means. Finally, the agent has beliefs (or belief-like attitudes) about what world-states are likely to follow (and how likely they are to follow) from each of the various possible actions. In classical decision theory, these latter attitudes are characterized using conditional probabilities. Ultimately, the agent is supposed to choose the action with the highest expected utility, where this is the sum of the products of each outcome's value and its probability conditional on the action.

Like the traditional approach, decision theory assumes a domain of possible outcomes over which the agent has preferences (the preferred outcomes are ends) and says that rationality consists in adopting the course of action (means) that bears the appropriate relation to those outcomes (probabilifying, maximizing expected utility, etc).

Both of these traditions imply that instrumental rationality is *voluntary*, *informed*, and *teleological*. I'll explain each of these ideas briefly.

Instrumentally rational choice is voluntary in the sense that the agent has a range of options that are in some sense open to them — various courses of action from which they can

[14] This way of putting things is borrowed from Steele (2020)

choose. Even if a specific action is rationally required, (e.g., there's only one way to bring about an outcome that I have decisive reason to bring about), there is a sense in which other courses of action are open to the agent. How exactly to characterize this openness, and indeed whether this openness is in fact instantiated in any of our real lives, is of course one of the oldest questions of philosophy. I'll set that all aside here, because my point is going to be that deep conceptual changes will tend to be obviously non-voluntary even given a commonsense understanding of what it is for an agent to have choices.

Instrumentally rational choice is informed in the sense that the agent undertakes the action *because* they understand the choice-relevant features of their decision-situation. Some caveats are necessary here: Obviously agents are neither factually nor logically omniscient. But in principle the picture is one of an agent who recognizes that they ought to bring some result about and for that reason takes steps to do so.

Finally, instrumentally rational choices are made for a specific kind of reason: teleological reasons. That is, the reasons for an action are all related to the action's promoting some desirable state of affairs. This is different from the last point: it's conceptually possible for an action to be informed in the sense that the choice-relevant features of the situation are known to the agent, but for some of these features not to reduce to desirable states of affairs.

I'm ultimately going to argue that all of these aspects of instrumental rationality create problems for a theory of rationally warranted conceptual change. And then I'm going to argue that we can have a picture of a kind of non-instrumental reasons-responsiveness that doesn't have these features.

1.1.2 The Davidsonian Dilemma

An old anecdote of uncertain origin poses the question how many legs a horse would have if tails were called legs. The answer, as any student of the use-mention distinction will tell you, is four: "calling a tail a leg does not make it so, you know."[15]

This story captures the first challenge for the would-be conceptual engineer: how do you expect changing our language, or our concepts, to change the world? At first blush, it seems like expanding the extension of "leg" to include tails is a mere shuffling of linguistic labels. Of course, conceptual engineers do not go in for the sort of crude linguistic idealism according to which simply calling tails "legs" would increase the number of legs in the world. Their claim, rather, is that what we call things can influence how we behave. But this, too, comes with its share of puzzles.

I think that those not schooled in the use-mention distinction often answer the tail question incorrectly. I think this is because they imagine a shift in usage and then answer the question as though that shift had occurred. But suppose we instead asked: If tails were called "legs," how many horseshoes would a horse need? I think people will have much less trouble with this question. Asking about relabeling is one thing, but mere relabeling does not, by itself, upset the networks of practical and inferential commitments people bring into the conversation. So, if we're aiming to increase horseshoe sales, it doesn't seem like getting people to refer to tails as legs is the way to go about it. Yet there's an interpretation of the conceptual engineering project on which that's roughly what it's trying to do.

We could diagnose the fallacy here as follows: Something's status as a leg carries with it a bunch of inferential and practical upshots, including that it's an apt target for being shoed. One possible upshot of adding something new to the ranks of the categorized-as-legs is to catch them

[15] "Suppose You Call a Sheep's Tail a Leg, How Many Legs Will the Sheep Have?"

up in this network, e.g., to get them categorized as to-be-shoed. But another possibility — one that's more likely when the thing very obviously doesn't fit into that network — is that the meaning of being categorized as a leg changes, e.g., by a weakening of the connection between something's being a leg and its being appropriately shoed.

In short, it seems like a relatively *superficial* conceptual or linguistic change — one that consists just in making the minimal changes required by changing what something is called — is uninteresting from the point of view of changing the world. Getting people to call tails "legs" wouldn't get them to change their behavior in any meaningful way. This is the first horn of the Davidsonian Dilemma.

It might be urged that there probably are many cases in which relabeling *does* get people to act differently. But, in many if not all of these cases, these behavioral changes will look like "tricking" the people in question, e.g., by exploiting what Cappelen has called "lexical effects."[16] Of course, there might be good instrumental reasons to trick people. But I think that the conceptual engineer aims for more than that: they want conceptual change as a process of rational thinking, learning, etc. They don't want to be — or don't want to *only* be — propagandists.

We might think, then, that what we need is a deeper change in our representational schemes and practices. There are many shapes this might take, including changes in beliefs, practical stances, and inference-dispositions. At the far end is what we might call *deep conceptual change* — change in what thoughts we're able to entertain, or in Kuhnian terms, a shift between incommensurable conceptual schemes. Burgess and Plunkett seemed to have

[16] See Chapter 11 of Cappelen (2018)

something like this in mind when they described changes in what we can do and who we can be.[17]

It is a matter of some controversy whether truly deep conceptual change is possible. Davidson thought it not to be. But what I want to show here is that, if deep conceptual change *is* possible, it cannot be undertaken as an act of 'conceptual engineering,' where this is understood as an act exhibiting instrumental rationality. To put the point schematically, suppose we are considering whether to transition from conceptual scheme C_1 to conceptual scheme C_2. If the choice whether to transition is instrumentally rational, then it should be voluntary, informed, and made for teleological reasons. Here we run into the second horn of the Davidsonian Dilemma: if C_1 and C_2 are deeply different from one another — if I, from C_1, cannot entertain C_2 thoughts — then how am I to choose between the two? The prospect raises puzzles for all three aspects of instrumental rationality.

Voluntariness is perhaps the most obvious hurdle here. If C_2 is unintelligible to me, what cognitive process allows me to choose it? I'm not going to be able to adopt it by flipping a switch in my head. Indeed, this is a problem for the idea of concept-acquisition in developmental psychology as well as in philosophy; more on this later.

Suppose I *could* flip a switch to adopt C_2. It's still hard to see how such a choice could possibly be an informed one. Given that C_2 is unintelligible to me, it's plausible that whatever reasons I have for adopting C_2 cannot be articulated using only the conceptual resources of C_1. If our conceptual repertoire determines what thoughts we can entertain, it will for that reason

[17] Of course, this is going to look more like a sliding scale than an on/off thing. Each of the changes I'll discuss might be thought of as approaching the kind of 'deep change' I'm interested in, and there's probably no bright line between being unable to think something and its being very difficult to do so. So, if I'm being careful, I'm going to say not that all cases fall neatly into one or the other horn of the dilemma. Rather, I'm going to say that each case will face a tension: it runs into the first horn to the extent that it's shallow and into the second to the extent that it's deep.

determine what reasons we can appreciate — and it may not be possible to appreciate the reasons for deploying a concept except by the deployment of that very concept. Echoing Cappelen, the principal reason to possess a concept of "rabbit" or "rational number" is to think and talk about rabbits and rational numbers.[18] I can't see the point of these concepts unless I already possess them.

This thought also puts us in a position to see one of the problems with teleology: just as teleological reasons are the wrong kind of reason for adopting a belief, they are in at least some cases the wrong kind of reason for adopting a concept. Plausible examples are provided by characteristically deontological concepts like *moral responsibility* and *obligation*. It is in the nature of these concepts that, in deploying them, we recognize *non*-teleological reasons. Any attempt to evaluate these concepts on the basis of their instrumental value would amount to a disavowal of them.[19]

Having put the Davidsonian Dilemma schematically, the remainder of my negative argument will consist in a consideration of cases. I'll start with some obvious examples of superficial conceptual changes. These cases will be meant to drive home that, even in cases where there are some practical upshots to superficial representational changes, they intuitively do not involve philosophically interesting *conceptual* changes of the kind the engineer might aspire

[18] Cappelen in this context was expressing skepticism about the notion of concepts having functions. I don't share his skepticism here, but I think the point he makes is useful in this context.

[19] I'll later discuss this idea as it relates to the concept of *terrorism*, but we can also turn to e.g., Anderson (1993) and Darwall (2006) for clear statements of the problem. Darwall draws from Strawson (1962) the thought that practices of holding-responsible cannot be vindicated by considerations of social desirability: "Desirability is a reason of the wrong kind to warrant the attitudes and actions in which holding someone responsible consists in their own terms" (Darwall 15). Anderson uses an example from *Star Trek: The Next Generation* in which the android Data starts an argument with a friend because he's learned that reconciliation tends to strengthen bonds of friendship. But authentic friendship cannot be conducted in this way: "If his anger isn't proper an sincere, and if their reconciliation is not based on a mutual agreement about what behaviors warrant anger, on sincere apologies by whoever misbehaved or misjudged the other, and on resolutions to act and feel appropriately as judged in terms of expressive logic, it won't be an authentic reconciliation on the basis of which the relationship can coherently continue" (Anderson 41).

to. Then I'll move on to an extended treatment of a more interesting case: the concept of terrorism. This concept is subject to very real contestation. Given that it's evidently practically important, we might think it can plausibly be interpreted as involving more deeply different conceptual schemes. So, I'm going to go through different ways of interpreting the question whether something should be regarded as terrorism. As before, I'm going to argue that shallow interpretations are either uninteresting or in some sense illicit. And, as we get to deeper representational changes, we're going to see that these changes cannot conform to the canons of instrumental rationality.

1.1.3 Preliminary Superficial Examples

Consider the difference between two ways of referring to the letters of the alphabet, one being the standard way ('A' is called *ay*, 'B' is called *bee*...) and the other being the NATO phonetic alphabet ('A' is called *alfa*, 'B' is called *bravo*…).[20] The latter way of talking has important advantages over the former, in which numerous letter-names sound sufficiently similar to cause confusion over a phone or radio. Were the community of English-speakers as a whole to transition to the NATO nomenclature, this would count as an act of conceptual engineering in some broad sense of the term. It would amount to a change, undertaken for practical reasons, in which linguistic expressions are associated with which intensions. For example, the intension usually expressed by "My name starts with *kay*" is now expressed by "My name starts with *kilo.*"

The alphabet case is a particularly stark example of superficial representational change. The transition to the NATO phonetic alphabet does not make any difference to what thoughts are available to us, but only provides a more convenient way of expressing the intensions we already

[20] This 'standard way' is of course standard only for speakers of English — that's part of where the need for the NATO version comes from. And even among speakers of English there is disagreement over what to call 'Z'.

were getting at. A slightly less trivial example is found in Cappelen (2018), in which he describes a 20th-century shift in the use of the word 'salad.'

> *At some point, not too long ago, a dish had to be served cold and have a high preponderance of green leaves in order to be a salad. At that time a concoction of cold cut fruit wouldn't be a salad. That has changed. Today, fruit salads are salads.*[21]

This example is slightly less trivial because the change in the meaning of 'salad' presumably did change what intensions people tended to express. Rather than giving people a more convenient way of expressing something that they were already expressing anyway, it brought something to people's attention that they likely weren't thinking of before. Indeed, it probably brought about real social changes by bringing fruit salads into public consciousness and thereby increasing the frequency with which fruit salads were prepared and consumed. Yet, as the passage from Cappelen illustrates, the change was still a superficial one. Prior to the change in the meaning of 'salad,' we could already talk about fruit salads as 'concoctions of cold cut fruit'. And now, post-change, we can get at the intension once associated with 'salad' by using the expression 'green salad.'

1.2 Terrorism

While Cappelen himself notes that 'salad' is not a particularly interesting example — most people don't care very much whether fruit salads count as salads — we could apply a similar analysis to at least some more practically significant cases. Consider a case from Chalmers (2011), which says:

[21] Cappelen (2018), 31

> *What counts as 'torture' or as 'terrorism' might be at one level a verbal issue that a philosopher can resolve by distinguishing senses. But in a rhetorical or political context, words have power that transcend these distinctions. If the community counts an act as falling into the extension of 'torture' or 'terrorism', this may make a grave difference to our attitudes toward that act. As such, there may be a serious practical question about what we ought to count as falling into the extension of these terms.*[22]

Chalmers here gestures toward a case of conceptual engineering as it tends to be understood in contemporary literature. To fill out the example a bit, suppose we start out speaking a version of English in which some type of action is not in the extension of 'terrorism,' and are deciding whether to transition to a version in which it does, making the minimal changes required by the relabeling. Such a change, were it to occur, would count as superficial in my sense.

What sort of practical upshot might such a change have? As the cases of the NATO alphabet and 'salad' illustrate, superficial conceptual changes can have practical consequences: they can make certain propositions more easily thought or communicated. There can of course be powerful instrumental — including moral or political — reasons to make such changes in how we speak. But I think that Chalmers is suggesting something quite different here, namely that including some action in the extension of 'terrorism' would get people to *think of* or *treat* it as terrorism. But it isn't clear that the envisioned semantic change would result in that kind of behavioral change, and there is a clear sense in which it *shouldn't* do so.

Let *A* be an action that contemporary English (E_1) doesn't classify as terrorism. E_2 is an alternative English that differs from E_1 only in that it classifies *A* as terrorism. These languages are intertranslatable, which implies that any consideration bearing on how we ought to act vis-a-vis *A* can be articulated in both languages if it can be articulated in either of them. So,

[22] Chalmers (2011), 2

transitioning to E_2 wouldn't supply me with any reasons for action that weren't already accessible to me from E_1.

Yet the question of what counts as terrorism very clearly *does* make a practical difference. We can see this in the term's contestation in the political sphere. Something's being categorized as terrorism has both cultural and institutional implications: terrorism is prosecuted differently in courts, and popular sentiment tends to be more in favor of harsh punishments and suspensions of civil liberties to combat actions categorized as terrorism. This category is so powerful that the question whether a given action fits its descriptive criteria sometimes feels like an afterthought. Consider the fact that it's often difficult to tell whether something counts as terrorism even once we've settled on a standard definition. Terrorism is typically defined by appeal to its motives — specifically by the motive to achieve some political or ideological aim by terrorizing a population — and motives are often unclear.[23] In light of this, it seems like the best thing to do when we're faced with something that looks like it might be terrorism is to suspend judgment unless and until we have more information about why it was done. But what actually happens is that people can't stand leaving such a powerful rhetorical weapon on the table. This leads not only to speculation about the causes of any one incident, but also to contestation over what the criteria should be. Do these contestations ultimately amount to rational discourse, or are they just power struggles?

On the one hand, we have what are pretty nakedly cynical attempts to bring state power to bear on one's political enemies. Egregious examples here include efforts by state lawmakers

[23] Cf. US Criminal Code, which requires that acts defined as terrorism "appear to be intended (i) to intimidate or coerce a civilian population; (ii) to influence the policy of a government by intimidation or coercion; or (iii) to affect the conduct of a government by mass destruction, assassination, or kidnapping…" (18 USC 2331)

to criminalize Black Lives Matter protests by categorizing them as a kind of terrorism;[24] the State of Georgia charging protesters with terrorism for sitting in trees;[25] and white supremacist Nick Fuentes asserting that Darrell Brooks driving into a Christmas parade in Waukesha, Wisconsin was "retaliation for the Rittenhouse verdict."[26] I think most of us can identify these as illegitimate uses of the concept. First, because they're being used to advance reactionary causes. But second, and more fundamentally, they're deeply incongruous with a commonsense understanding of what terrorism is and why it matters. Regardless of the ends toward which it's deployed and whether we agree with those ends, nonviolent resistance simply is not terrorism.

This latter point is illustrated further by cases in which people attempt to engineer the concept of terrorism toward *good* ends. Take Chris Hayes suggesting that terrorizing people should be sufficient to count as terrorism.[27] I'd suggest it's because people (rightly) observe that the "terrorist" label is used for white supremacist purposes and try to counter that by counting more white people as terrorists. Even if this were successful, it's equally perverse from the point of view of trying to figure out what's really terrorism and what ought to be treated as terrorism.[28]

The concept of terrorism thus provides a fertile, and potentially vexing, case study for the conceptual engineer. On the one hand, it seems clear that disputes over its extension occur and that these disputes have practical upshots. But it's equally clear that many of these disputes are in

[24] These attempts largely arose in 2016 and 2017 in Republican state legislatures as responses to Black Lives Matter protests that began in 2014 and 2015 after the killing of Michael Brown. Examples include bills in Arizona ("Arizona Senate votes to seize assets of those who plan, participate in protests that turn violent"), Washington ("Trump supporter in state Senate says some protests are 'economic terrorism,' should be felonies"), and North Carolina ("NC bill has tough penalties for disruptive protesters, 'economic terrorists'"). Fortunately, none of these laws passed.

[25] "Domestic terrorism charges in Georgia are prompting concern over political repression"

[26] "Waukesha: Tragedy Exploited by White Supremacists"

[27] "Chris Hayes: If This Isn't Terrorism, What Is?" 3:24

[28] My point here isn't that considerations of impartiality *per se* are illegitimate. As I'll discuss later, I think it's perfectly legitimate to change a concept's accepted extension because we decide its previous one made arbitrary distinctions. But Hayes – on this interpretation of what he's saying - is throwing out the baby with the bathwater, casting aside a core aspect of the concept in an attempt to equalize its application by fiat.

some sense perverse, involving shuffling labels as a kind of rhetorical sleight-of-hand similar to what we saw in the parable about tails and legs. The challenge for the conceptual engineer — insofar as they're interested in conceptual change as rational activity — is to come up with an interpretation of these engineering projects that *doesn't* look like mere propaganda or sleight-of-hand. That is, they must come up with an interpretation on which getting people to revise their concept of terrorism actually gives them *good reasons* to act differently.

1.2.1 Eccentric Demands

As I've indicated, I'll be arguing that any attempt to interpret changes in the concept of terrorism as instrumentally rational choice will run into the Davidsonian Dilemma. To drive this point home, I'm going to discuss a silly case. The point here is to give an obviously bad example of instrumentally motivated change; the aim then will be to show that the problems with the obviously bad, silly case carry over to any other possible interpretation of the change.

Suppose an eccentric billionaire offers to give me one million dollars for altering my concept of 'terrorism' to include obnoxiously long questions during Q&A sessions. There are at least three things this billionaire could be asking me to do: *call* questions "terrorism;" *treat* questions as though they were terrorism; or *believe* questions to be terrorism.[29]

1.2.2 Mere labeling and treating-as

Not much needs to be said about this one. It's easy enough for me to start calling questions "terrorism" - it's a relatively small ask in large part because it involves no major epistemic or practical upshots. This corresponds to points I've already made about how simply

[29] Of course, these are not fully separable in practice. But they're analytically distinct and addressing them separately will make an important point.

calling something “terrorism” doesn’t really constitute a conceptual change and doesn’t give us any significant practical reasons.

What if the billionaire is asking me to *treat* questions as terrorism? Of course, much of the practical significance of the terrorist label comes from its implication in legal and other institutional machinery, and I’m not in any position to unilaterally alter those. But besides the institutional significance of the label, there’s a folk understanding that terrorism is wrong, that terrorists should be stopped by any means necessary and punished harshly. This latter point is especially important given the above-mentioned role of the concept of terrorism in the political sphere. This is what makes terrorism as a concept so powerful: not just its legal status, but its cultural significance as a category that justifies state violence and repression.

So, imagine I change my concept of “terrorism” not only in the sense that I start to call questions “terrorism,” but also in the sense that I start to treat them as terrorism. I’ll now be willing to stop lengthy questions by force, to prosecute people for asking them, and to endorse the state doing similarly.

There’s nothing terribly puzzling about this, but there’s clearly something perverse about it. It’s not appropriate to treat questions as terrorism. Why not? Because they *aren’t* terrorism or aren’t normatively similar to it in a way that warrants similar treatment. Notice that this latter point - that questions don’t, by their nature, warrant treatment as terrorism - holds true even if I’m supplied with compelling instrumental reasons to treat them as such. Plausibly it’s rational for me to take the money and start prosecuting people for questions. But here, again, it doesn’t seem like I’ve acquired any interesting reasons to change my concepts: I’m effectively taking a bribe to pretend that questions are terrorism.

This diagnosis extends to some of the cases discussed above. Clearly the point of legislators trying to categorize peaceful protest as terrorism was just to have it punished harshly. This attempted (sometimes successful) change in the legal meaning of a term is unjustified not only because it's used to defend an unjust status quo, but also because it involves a kind of moral perversity: blocking traffic simply isn't terrorism, so all else equal we tend to think it shouldn't be treated as such in legal contexts.

This latter point extends even to cases where we think that the reasons for altering the concept are better ones. Maybe Hayes' suggestion that we include all terrorizing public violence as terrorism would result in more even (e.g., less racialized) application of the concept and its legal consequences. But doing so would involve glossing over an important distinction between different types of violence warranting different types of responses. And this, again, continues to be the case even if we think we have sufficient reason to do it. Maybe society would be better off if we changed our legal definition of terrorism to exclude a requirement of political motive. But this is best understood as an instrumentally justified treatment of non-terrorism as though it were terrorism.

To sum up: Merely labeling something as "terrorism" does not, by itself, give us any reason to treat it as terrorism. And, conversely, the fact that we have decisive instrumental reason to treat something as though it were terrorism does not necessarily say anything about what our concept of terrorism should be.

1.2.3 Belief

There's a third thing we could be talking about when we talk about categorizing questions as terrorism: *believing* that questions are terrorism. But this doesn't look good for the would-be conceptual engineer, either. Now we're running into a familiar sort of reasons-for-

belief puzzle. It may actually be *impossible* to change my beliefs voluntarily, and in any case it's probably irrational to change my beliefs for instrumental reasons. The fact that I'd get a billion dollars for believing questions to be terrorism is a powerful instrumental reason, but it's a reason of the wrong kind. Belief as an attitude (or as a discursive practice) is constituted by certain norms, and believing for instrumental reasons conflicts with those norms.

Looking through this lens at our real-life examples, they again don't look great. Obviously, people like Fuentes are making judgments here for ideological reasons - those judgments are bad not only because the ideologies motivating them are bad, but also because they're unjustified on more straightforward evidential grounds. And if people are ideologically motivated to judge the shooters in Las Vegas or East Lansing to be terrorists - for either racist or antiracist reasons - then this seems like a perverse sort of concept-creep.[30] (It's perverse for the - in my view related - reason that desiring a particular extension for the term is a reason of the wrong kind and because using it in this way obscures something morally significant.)[31]

So far, I've explored characterizations of conceptual change that can be articulated in terms of our current concept of terrorism: metalinguistic changes, behavioral changes, and belief-changes intuitively do not involve any deep conceptual change, as is reflected in the fact that I can describe all these processes - including the attitudes they'd involve adopting - in terms of my own concepts. And this fact in each case works to undermine the engineering project. In the relabeling case, it saps conceptual engineering of its practical significance. In the treating-as

[30] While information suggesting an ideological motive for the Las Vegas shooter did eventually come to light, the conceptual point stands: because terrorism is political violence, we should at no point in time be more confident that something was an act of terrorism than that it was an act of political violence.

[31] What counts as a reason of the right kind depends on the constitutive norms that come with the concept. As I'll try to clarify later, the aim of the concept of terrorism is to organize our experience in a way that identifies a practically significant unity between a bunch of different acts of violence. The fact that it would be nice for two acts to be treated as sharing that unity is not by itself significant from the concept's point of view.

case, we saw that having instrumental reason to treat something as though it fell under some concept doesn't give us reasons to regard it as falling under the concept. And for belief, we ran up against old puzzles about the possibility and permissibility of instrumental reasons for belief. Yet this last observation - that concepts might be in some way analogous to beliefs - points toward a way out. Maybe we can model rationally warranted conceptual change as in some respects similar to rationally warranted change of belief.

1.3 Belief-change and gesturing toward the new approach

What we need is a kind of non-instrumental rationality. Belief provides a good starting point. Rational belief-change is *non-voluntary* and undertaken for *non-instrumental* reasons. I'll address these two points in turn.

There are multiple ways to cash out this notion of non-voluntariness as it relates to belief. Modal glosses are common, e.g., "you can't do otherwise." Perhaps the intuition that you can't flip a switch and believe something will be enough for some. Hieronymi has offered a different, more robust account on which a voluntary activity is one that you do by forming and executing an intention.[32] Intentions are non-voluntary just like beliefs are. What separates beliefs from intentions is that they're answerable to different sorts of considerations: intentions are answerable to any sort of considerations that bear on the question whether to do the activities at which they aim, while beliefs are answerable only to considerations bearing on truth. This is why you cannot believe something by executing an intention to believe it, i.e., why belief is non-voluntary.

[32] Hieronymi (2008), 366

It's also a commonplace that considerations of utility are reasons of the wrong kind for belief. Again, this point will strike many as intuitively obvious. Hieronymi, as we saw, ties this point together with the previous one: while intentions are answerable to any consideration that makes the targeted action a good thing to do (or targeted state of affairs a good one to bring about), beliefs are strictly answerable only to considerations bearing on truth.[33]

These are two respects in which the adoption or rejection of a belief is a kind of rational activity that breaks with the model of instrumental rationality. What about the third aspect, informedness? I think this is where we can find the difference between belief-change and conceptual change.

Suppose I'm playing poker, my opponent has just gone all-in, and I'm deciding whether to call. I'm holding a king-high flush, which given the cards on the table is the second-best possible hand. My opponent beats me only if they have the ace of the relevant suit. Do they have it? While it's impossible for me to fully know the answer to this question, whatever doxastic steps I take here will (or at least can be) informed in the sense discussed above. I understand all the possible combinations of cards my opponent could be holding. For each combination, I know what will happen if I call and my opponent is holding that combination. I also know how my opponent has behaved so far: how they've bet in this hand and perhaps in previous ones, what their general playing tendencies are, etc.

Beliefs in general have this feature. When I'm deciding what to believe (or, more suited to the poker example, what credences to assign), I can be modeled as surveying a range of possibilities all of which are intelligible to me. I'm deciding which of these possibilities is true

[33] Hieronymi (2008), 369

and which false (or which likely and which unlikely). And this is how belief-change differs from deep conceptual change.

The mysteries of deep conceptual change arise from the fact that, from the point of view of the starting conceptual scheme, the new conceptual scheme enables claims that the starting one cannot assess as either true or false (or to which it cannot assign a probability). So, while the non-voluntariness and non-teleological nature of conceptual change is no more mysterious than those same features of belief-change, conceptual change does have the distinct mystery of being something that seemingly has to happen 'blindly,' i.e. without possessing all the relevant information. This kind of incompleteness is different from that involved in typical cases of decision-making under uncertainty in that, if I *were* to possess all the relevant information, that in itself would imply that I had (indeed, would constitute my having) undergone the relevant conceptual change.

Given this, it might look like deep conceptual change necessarily involves a kind of non-rational leap of faith. Perhaps it sometimes does. But, I think, not always. Despite being non-voluntary and non-informed, I think these changes can be interpreted as cognitive successes. Not just because they result in a better understanding of the world, but also because the process whereby we come to them involves responding to (truth-relevant, non-teleological) reasons.

1.4 Carey on Conceptual Change

Susan Carey's *The Origin of Concepts* offers a rich and detailed account of conceptual development from early childhood through adolescence. It offers an account of how we go from a relatively sparse repertoire of innate representational devices to the robust system of learned theories we deploy as adults. Central to this account — and what interests me here — is the

problem of *discontinuity*, or of how humans "create new representational resources that are qualitatively different from the representations they are built from."[34]

Of course, the extent of conceptual discontinuity in actual human development is a matter of longstanding controversy. Jerry Fodor, whom Carey acknowledges as a foil, famously argued that all lexical concepts are innate.[35] My aim here is not to wade into this dispute, but to borrow from Carey to sketch a picture of what deep conceptual change might look like.

Carey's understanding of discontinuity explicitly draws on Kuhn's idea of incommensurability.[36] On Carey's view, two conceptual schemes are incommensurable when "one contains concepts that are not merely absent from the other but are actually incoherent from the point of view of the other."[37] This is roughly what we're looking for in seeking deep conceptual change.

Carey argues that humans undergo learning processes whereby they pass from innate representations to intuitive theories the concepts of which are not reducible to the innate representations themselves. And the education we receive in childhood and adolescence, if all goes well, involves transitioning from one intuitive theory to another, incommensurable one. If this is true, it follows that deep conceptual change can occur as a process of learning. Since learning is a paradigm case of rationally warranted attitude-change (compare learning a fact based on evidence that warrants belief-change), it follows that rationally warranted deep conceptual change is possible. My hope is that we can adapt the picture of how this happens in childhood to the cases conceptual engineers want to talk about. The upshot will be that, while it's

[34] Carey (2009), 18
[35] Fodor (1975), 80
[36] Carey (2009), 367
[37] Carey (2009), 359

possible to undergo a deep conceptual change for good practical reasons, the model of how this happens and the reasons at issue will be different from how it's usually understood.

1.4.1 Rational Number

There is a stage of development at which children possess a concept of natural number but not of rational number.[38] They've learned to count up to the limits of what their memory and possession of numerical vocabulary allow, and they understand that there is no highest number. They understand addition and subtraction in terms of putting things into or taking them out of a set, and also in terms of counting up and down. While multiplication is learned later, it's also learned as just an extension of addition. But they deny that there is a number between 0 and 1. While they no doubt understand that at least some objects can be divided in half, they do not recognize ½ as a number.

How do we explain rational numbers to them? Our first instinct might be to explain them in terms of division, e.g., ½ is just the result of dividing 1 by 2. But they lack a concept of division — that's bound up in the concept of fractions or rationals and must be learned alongside them. And the strategy we used with multiplication — analyzing it as a kind of addition — won't work here. Division-talk can't be translated into subtraction-talk like multiplication can into addition. In short, a conceptual scheme that recognizes rational numbers is incommensurable with one that only recognizes natural numbers. They are incommensurable, with one being strictly more expressively powerful than the other.

1.4.2 Matter, Weight, Density

[38] See also the fascinating discussion of how children go from rudimentary quantitative concepts — which include the ability to track sets of observables, to compare magnitudes, and to understand basic natural-language quantifiers — to the system of counting and the natural numbers.

Children possess a concept of *object* which is closer to our concept of *physically real* than to our concept of *physical object*. This case is much harder for us adults to get a grip on. Unlike the previous case, where the child's numbers can be seen as a special case of ours, the child's concept of object is as unintelligible to us as ours is to them. One challenge for Carey's theory is to characterize how the child represents objects. Only with such a characterization in hand can she go on to explain how the child's theory of objects changes in the course of their education.

They distinguish objects from abstract entities like ideas, but they place physical objects into the same category as shadows and light. This has wide-ranging consequences for their theory of what properties objects have and how they interact. For example, they reject the principle that all objects have weight and that two objects cannot be colocated. Shadows lack weight and can overlap each other.

Children understand that some objects have weight, but the role of weight in their theory is much less central than it is in ours. As we saw with the shadow example, weight is an incidental property of certain objects rather than a defining characteristic of objects as such. More puzzlingly, weight is not distinct from density. Rather than separate concepts of weight and density, children have an undifferentiated concept of *weight/density* — I'll call it *heaviness* — that does double duty.

This last point is important. When Carey says that children have an undifferentiated concept heaviness covering both weight and density, the philosopher may be tempted to hear this as meaning that the child's concept is a mereological sum of ours — that heaviness is ambiguous between weight and density or that children believe that weight and density are the same thing.

On this picture, what the child learns is just to better distinguish between things that were already picked out by their concept of heaviness.

But this picture is wrong. Heaviness cannot be translated into the language of weight and density. This can be seen by considering the inferential role of the concept of heaviness: it's a monstrous hybrid of the two that lacks the explanatory import of either.

Heaviness, like density, is non-additive. Children accept that a pile of 50 lentils has weight while holding that a single lentil weighs nothing. If you take a lump of clay that they say has weight and add to it a tiny amount of clay — an amount small enough that it lacks *felt* weight — they'll say the resulting lump weighs the same as it did before. So it seems like the concept of heaviness is used to make sense of the everyday phenomenon of felt weight, but it's not conceptualized as anchored in an additive concept like *mass* (i.e. amount of matter). Heavy objects are hard to pick up and move around; a heavy object exerts more force on the things around it (e.g., collapsing sponge bridges or rolling into things and knocking them around); heavy objects tend to sink when placed in water.

The chimeric nature of the concept of heaviness is driven home by an experiment in Carey (1991).[39] Children were shown a smaller block of steel and a larger block of aluminum and were shown that the two blocks weighed the same. When asked how they can weigh the same when they're of such different sizes, children had what sounds like a good answer: "Because steel is heavier."[40] It's tempting here to interpret the child as using 'heavier' to mean 'denser.' But this interpretation is undercut by the next stage of the experiment. Children were shown a second pair of blocks, one steel and one aluminum, now of equal size. When asked about the relative weights of the blocks, children said they would weigh the same. "Because the

[39] Carey (2009), 393
[40] Carey (2009), 392

steel and aluminum weighed the same before."[41] If these children had a stable concept of density — even if it were one that they picked out with a term ambiguous between density and weight — you wouldn't expect them to make this mistake.

The lesson of Carey's experiment is that children do not distinguish between weight (a property of objects) and density (a property of substances). This is despite the fact that they *do* distinguish between objects and the substances of which they're made![42]

Another interesting feature of Carey's experiment is that the children themselves were aware of the contradiction their concept of heaviness led them into. They could see their prediction that the equal-sized cubes would weigh the same falsified by watching them placed on a scale. They were bothered by this, but didn't know how to resolve it. This leads us into Carey's discussion of how conceptual change occurs.

1.4.3 Quinean Bootstrapping

Carey proposes a process she calls "Quinean Bootstrapping." She does not claim this is the only process by which discontinuous conceptual change can occur, but she thinks there's good evidence it does occur in the cases she discusses as well as in the history of science. She divides it into six parts.

Symbols are learned and the relations between them articulated.[43] These symbols are at first only partly interpreted. Your understanding of their meaning may be exhausted by your knowledge of their inferential relations, or you might have a partial and imperfect mapping from your existing representations onto the new ones. Then comes the interpretation stage in which modeling processes are used to provide an interpretation of the hitherto uninterpreted (or partly

[41] Carey (2009), 392
[42] Carey (2009), 384
[43] Carey (2009), 307

interpreted) symbols. These mechanisms include analogy, thought experimentation, limiting case analysis, etc. While they "combine and integrate representations from distinct domain-specific conceptual systems," the mechanisms themselves are domain-general and are applied to various problems across contexts and lifespan.[44]

These mechanisms are *problem-solving* techniques. The learner is equipped with an uninterpreted theory and faced with the problem of trying to provide an interpretation of it. Simultaneously, the learner faces problems with their existing theory that the new one, once interpreted, can solve. The solutions to these problems are intertwined. This is illustrated by the development of the concepts of physical object and rational number, which develop alongside and support one another.

Children limited to the natural numbers deny that there is a number between zero and one, observe that things can be divided. They can see that a line segment exists between 0 and 1 on a number line, and that it's possible to carve up this line segment. This observation can be bolstered by analogy of the divisibility of physical objects or quantities, e.g. teachers can use a quantity of flour to support a mapping between divisibility of number and divisibility of matter.

But divisibility of matter is *also* something the child has to learn. While children understand that a pizza can be cut into slices, they don't view matter as infinitely divisible in the way we know numbers to be — they think that dividing an object in half a certain number of times will result in its disappearing altogether.[45] So, they don't have a preexisting idea of continuous matter that they can reach for in understanding the new idea of continuous number.

[44] Carey (2009), 418
[45] Carey (2009), 406

The concepts of infinitely divisible number and infinitely divisible matter develop in tandem with one another.[46]

The lesson here is that learning the concepts of weight, density, and rational number is not a stepwise process. The child does not build these new concepts out of existing ones, nor do they master one new concept before moving on to the next. Rather, they are faced with a cluster of problems with interlocking solutions. They're taught a formal system of division and fractions that at first is unintelligible to them, and they have to make sense of it. They're presented situations that raise puzzles they can't solve with their existing physical concepts — puzzles that require a distinction between weight and density and the observation that weight is additive. Eventually, there emerges a concept of matter as continuous, taking up space, and having weight. The weight of an object is determined by the amount of matter in it. Density is the ratio between weight and size. The analogy between dividing numbers and dividing objects supports the development of both concepts: mathematical division is made intuitive by analogy with splitting objects while the principles of mathematical division are carried back over to objects to support an understanding of objects as continuous.

1.5 A Lesson for Rationally Warranted Conceptual Change

The transition from natural numbers and heaviness to rational numbers, weight, and density is a learning event. It's not just that the latter conceptual scheme is superior to the former — though it is. The transition shares the two features of belief-change that distinguish it from instrumental rationality:

[46] Carey (2009), 434

First, it is neither voluntary nor involuntary. It requires the learner to exercise will and intelligence to solve the problems with which they're faced. But the learner does not *intend* prior to possessing a concept of rational number, to form a concept of rational number. Or, perhaps better, learning the new concept is not something they can do merely by executing an intention.

Second, the reasons to which the change is a response are not exclusively teleological. While children are better off after the change, they're motivated by a desire to make sense of the world and considerations to which they're directly responding are, in a broad sense, evidential (e.g. resolving contradiction). They face a problem and they solve it, but this solution constitutes a transformation of viewpoint that doesn't fit into the framework of instrumental rationality. They're motivated by observations, by inconsistencies, etc. rather than by considerations of what the world will be like after they've changed their concepts.

It also has the feature that distinguishes deep conceptual change from belief change, namely that it is non-informed. Learners observe the problems their current concepts raise, but don't see a solution. To see the solution is to have already solved the problem. There is a 'leap of faith' here in the sense that you have to experiment, 'get-a-feel,' do some trial-and-error. But that's a leap that all theorizing involves.

In short, the transition from childhood concepts to adult ones involves non-instrumental reasons-responsiveness of the sort I've said should interest the would-be conceptual engineer. Now let's try and analogize it to other cases.

1.5.1 Good Conceptual Change

One thing you might notice is that, arguably, the history of philosophy is chock full of *good* proposals for conceptual change. I'm going to argue that these changes are possible, but that they will typically (or paradigmatically) be cases involving a kind of non-instrumental, non-

deliberate reasons-responsiveness. To make things concrete, I'll be discussing examples related to the concept of terrorism.

Jonathan Glover's 1991 paper "State Terrorism" argues - what I think is today pretty uncontroversial in at least some political circles - that the concept of terrorism should be extended to include some actions by state actors. At this point, the concept of 'state terror' is pretty familiar. Yet our most influential lawmaking and enforcement institutions almost always define the term 'terrorism' such that 'state terrorism' is nearly a contradiction in terms.

The Patriot Act, as well as various state legislatures, have defined 'terrorism' in such a way that only actions outside of state legitimacy can count. So, while it's technically possible for politicians, law enforcement, etc. to commit acts of terror, the state as a whole always has the ability to absolve itself by declaring its actions legal. And while terroristic state actions may count as war crimes according to the ICC, a state's ability to terrorize its own population is less restricted. Glover points out that:

> *Conventionally [war and revolution] are not thought of as terrorism. To include them would be to expand the concept in a perhaps perverse way. But to exclude them may be to sustain a conventional blindness to important similarities to the standard cases of terrorism.*[47]

Claudia Card's paper in the same volume develops a more radical proposal. In "Rape as a Terrorist Institution," Card argues that sexual violence as a kind of informally conventionalized social practice - i.e. as a kind of institution - has the function of shoring up male supremacy by terrorizing women. Like Glover, she acknowledges that this is likely to appear to readers as an awkward changing of definitions:

> *What is philosophically interesting is that without disputing the facts many do not yet apply the concept of "terrorism" to rape. Recognizing that the concept applies is yet another*

[47] Glover (1991), 257

step in clarifying what is wrong with rape and how bad it is in relation to other abuses.[48]

Also like Glover, Card thinks that our tendency to exclude her target phenomenon from the extension of 'terrorism' masks important similarities between it and the standard cases of terrorism. It "ignores the terrorism of *sexual* politics. Ethically, that exclusion is arbitrary and irresponsible. It maintains an invisibility of routine violence against women, underlying visible sexist stereotypes."

The basic point, for both Glover and Card, is that there's a kind of *normative unity* between the standard cases of terrorism and the cases they offer. We think that terrorism is a special kind of wrong because it involves manipulating a group of people through fear of (relatively) indiscriminate violence.[49] None of us wants to be terrorized, and terrorizing people is a way of coercively imposing your will on others. It's also different from more straightforward forms of coercion in that it relies on a kind of anxiety: you don't know what exactly will keep you safe, if you'll be next, etc. From this point of view, it makes perfect sense to include state terror and systemic sexual violence in the category. The tendency for people to suddenly disappear creates the sense of an unaccountable, inescapable tormentor with whom one must comply. And widespread sexual violence creates in women the sense that they are never safe without the protection of a man. Both of these induce compliance by creating a generalized sense of unsafety — a threat that's all the more omnipresent for the fact that it's never fully articulated.

For Glover and Card, this normative unity warrants a change in how we classify things. The notion of "warrant" at work here needs to be elucidated. While Glover and Card don't explicitly commit to this claim, it's natural to read them as saying that state violence and sexual

[48] Card (1991), 297
[49] "Indiscriminate" here is compatible with the violence being targeted at members of a specific group. What I mean is that it makes *any* member of that group a possible target.

violence warrant the relevant attitudes and actions *in virtue of the kinds of phenomena they are*. Put another way, the features these phenomena share with standard terrorism are features that inherently warrant a certain kind of response from us. This is very different from an instrumental rationale. The idea here isn't that it's a *good idea* to adopt the same stance toward state violence and sexual violence that we adopt toward IRA bombings - it may or may not be - but that their having such-and-such features is, in itself, a good reason for adopting that stance toward them.

With this picture in hand, let's return to our list of approximations of conceptual change from before. Suppose you read Glover and Card and you revise your concept of terrorism to include state actions and sexual violence. What will you actually be doing?

First, you'll relabel and start *calling* these actions terrorism. As before, there's not much interesting to say here. Relabeling by itself isn't very interesting - it's interesting in this case only because it's attended by the other kinds of change we've been discussing.

Second, you'll start treating these actions as terrorism. Here things start to get interesting. Part of the point of Glover's and Card's essays is to convince readers that their target phenomena ought to be treated as terrorism because they're relevantly similar to things we already treat that way. But, if their case is effective, we'll have avoided the bribe-taking problem from earlier. They're not trying to convince you that things would turn out well if you treated these phenomena as terrorism. Rather, they're trying to convince you that they inherently warrant such treatment - that failing to treat them this way would violate a constitutive norm of moral judgment and practice which says we should treat relevantly similar things in relevantly similar ways.

Third, I think you'll come to *believe* that their target phenomena are acts of terrorism. This one is likely to be much more controversial. Does the fact that something is similar to

terrorism - even similar in a way that warrants treatment as terrorism - count as a reason (of the right kind) to believe that it is terrorism? It's hard to answer this question while staying neutral on thorny issues surrounding what concepts are in general, the nature of the concept 'terrorism' in particular, and what a conceptual change (either in general or in this one case) amounts to.

The first thing to say in favor of believing Glover and Card's cases are terrorism is that, while our earlier worries had to do specifically with *instrumental* reasons for belief being perverse, we don't have that problem here. We could consistently - and, I think, plausibly - claim that instrumental reasons are reasons of the wrong kind for judging something to be terrorism without making the stronger claim that moral considerations in general are impermissible. If we think that 'terrorism' is a thick moral concept, then it seems obvious that moral considerations will bear on whether judgments couched in terms of it are correct. Why wouldn't noticing that things match the descriptive criteria for the application of a thick moral concept count as a good reason (by the lights of the concept itself) to apply the concept?

Finally, we might interpret the change of attitudes Glover and Card call for as a bona fide conceptual change. Following Carey's model, we could say that a reader's idiolect prior to encountering these papers and their idiolect afterward are locally incommensurable. This interpretation of the representational change is supported by Glover's and Card's quotes above about how they aren't trying to convince you of new facts so much as to get you to organize those facts differently in your cognitive landscape.

Assuming this shift does involve transitioning between locally incommensurable conceptual schemes, can we model the shift as instantiating the kinds of learning process Carey describes? I think so. Recall that the conceptual learning process is motivated in part by the learner encountering problems their current conceptual scheme cannot resolve. The failure to

capture a normative unity is arguably one such problem. This is a different kind of problem from the ones Carey talks about — it involves a violation of norms against non-arbitrariness rather than an empirical inadequacy or outright contradiction — but it's a problem nonetheless. The problem is both highlighted and resolved through analogical reasoning: why are these cases, which are so similar to standard cases of terrorism, not classified as terrorism? Glover's and Card's favored solution is to carry that analogy through and expand the extension of the term. But there are two complications of this problem-situation that Carey's framework illuminates.

First, Glover's and Card's proposed solution isn't the only one available to the reader. Having noticed the problem, the reader also has the option to retool their concept of terrorism in ways that continue to exclude Glover's and Card's target phenomena. They could, for example, highlight that paradigm cases of terrorism involve *explicit* ideological motive (i.e. motive that's known to the perpetrator) of a kind that's typically (though not always) absent in cases of sexual violence. Or they could form views of political legitimacy that grant state violence a special status that sets it apart from non-state violence, even in cases where the two adopt similar methods. My point isn't that these are equally good ways of handling the problem. It's that appreciation of the problem doesn't force any particular solution; that the formulation of (and adjudication between) solutions is non-deductive in roughly the way Carey describes.

Second, if Glover's and Card's arguments cause the reader to shift to a conceptual scheme locally incommensurable from the one they started with, this will imply a change in the inferential role of the concept of terrorism as well as related concepts. And I think that such a change does occur. Accepting Card's argument, for example, requires us to see individual actions as caught up in social institutions in a way that we may not have before. So while we might previously have assumed that terrorism must be ideologically motivated and ideological

motivations are always explicit, Card might convince us that individuals' actions can be norm-guided and indirectly ideological even if the agents involved don't see them as such. In this way, the inferential connections between 'terrorism,' 'ideology,' 'institution,' and 'motive' are transformed.

1.5.2 Wholesale Rejection - Chastity

While I think it's plausible to talk about Glover's and Card's arguments about terrorism as inducing conceptual changes, I'm sure there will be lots of disagreement about that. It's certainly possible to interpret them as doing something less ambitious than inducing a truly deep conceptual change. But I still want to offer an example of a deep conceptual change that's undertaken for practical, but non-teleological, reasons. This is important because, while Carey provided a detailed account of how deep conceptual change can occur and how it can be interpreted as a learning process, her focus was on scientific and mathematical concepts rather than moral or political ones. Given that many conceptual engineers are mostly interested in these latter concerns, it would help to have a clear case involving them. So instead of talking about the extension of a given concept changing, let's talk about a concept being rejected entirely.

Take the concept of sexual chastity. To indulge in a bit of autobiography, I was raised in a somewhat sexually conservative religious culture in which the concept of sexual chastity had quite a bit of purchase. The idea was that sexual behavior of the wrong kind (with too many partners, outside of marriage, with the wrong kind of person) made people - especially but not exclusively women - symbolically dirty. This symbolic idea of sullying was bolstered by various elements of moral and religious doctrine as well as a suite of more descriptive claims (some true and some false) surrounding the physical effects of sex. Premarital and homosexual sex (and, for many whites in my community, interracial sex) were sinful and immoral. They permanently

harmed one's moral character. This moral damage was also thought to be marked on the body. Teenage pregnancy - a very real problem in my community which had the highest rate in the state at the time - was a physical manifestation of the moral danger of impermissible sex.[50] So were STIs, the danger of which was emphasized by abstinence-only sex education classes which also deemphasized the availability of methods for reducing risk. Finally, there was the widespread and entirely mythical idea that repeated insertive sex (again, only if it was of the wrong kind) would result in deformation of the vagina - a physical marker of impurity that also had the effect of making the woman undesirable to future partners.

Of course, much of what I'm saying here was either masked entirely from me or at best only obscurely tracked. Over time, one of the effects of my education - especially what I learned from those in feminist and queer liberation movements - was that the ideological scaffolding which made intelligible the concept of chastity was false. Contact with another person's genitals doesn't alter a person morally or, in most cases, physically. A crucial aspect of this education was the realization that the concept of sexual chastity plays a role in a system of patriarchal control. This system conflicted with broader commitments to freedom and equality that I regarded as more fundamental than my commitment to the ideology of sexual purity.

The effect of all this is that I simply stopped making judgments in terms of chastity. It's not that I *believe* no one is chaste, or that everyone is. Nor is it that I think making judgments in terms of chastity would produce bad results. Rather, it's more like the concept is no longer intelligible to me. Of course, I'm able to track (at least roughly) what sort of judgments and treatments would be appropriate by the lights of the concept. But if I look at a sex worker, for example, and acknowledge that she would count as 'impure' from the standpoint of that concept,

[50] See data for Vance County in 2005, accessible via "Archived State Statistics."

it feels like a kind of anthropological ‘inverted commas’ claim rather than a full-throated judgment that deploys the concept of impurity.

This case is interestingly different from the terrorism case. Whereas that could plausibly be interpreted as just a change in belief about what does and doesn’t count as terrorism, this one can’t. It seems I’ve become indisposed to even make the distinction that the concept of sexual chastity makes. Why? Because the concept encodes a kind of normative commitment: that someone has engaged in certain kinds of sexual contact is a reason to treat them in such-and-such a way. But it simply *isn’t* a reason to treat them in that way, and you’d only think it to be if you’re bound up in male supremacist ideology. These aspects of people’s sexual history either don’t naturally group together or, if they do, don’t warrant the kind of normative stance that the concept of sexual chastity recommends.

Returning to Carey’s idea of problems and the domains of problems concepts try to solve, there are a couple of things we can say about the concept of chastity. We already saw what problems the concept of chastity raises, but which ones does it try to solve? There should be some answer to this; just as the child’s concept of heaviness helped them to navigate a world of mid-sized solid objects, there must be some cognitive task to which the concept of chastity is suited. This task, I would suggest, is that of navigating a male-supremacist world and, relatedly, that of maintaining the system of male supremacy. One way of making sense of why rejecting the concept of chastity was the correct move was that the ‘problems’ it sets out to solve are illegitimate from the start.

1.5.3 Conclusion

I’ve tried here to lay out a problem for conceptual engineering as rationally warranted conceptual change, where the model of rational warrant is that provided by the notion of

instrumental rationality. I argued that this picture of conceptual engineering faces a dilemma: if the conceptual changes it involves are shallow, then they won't (or shouldn't) have interesting payoffs; if they're deep, then they can't be done instrumentally. I then laid out a model, informed by Carey's work in developmental psychology, of how humans in fact learn to master conceptual schemes that are deeply different from those they start out with. This learning process involves a kind of non-instrumental reasons-responsiveness, and I think that something like Carey's picture of this process could be carried over to cases of interest to conceptual engineers.

Chapter two articulates a structurally similar problem for ameliorative projects in feminist theory of gender. I argue that such theories also face a dilemma: one between the epistemic and practical aims of feminist theorizing. But while the Davidsonian Dilemma of this chapter has to do with how deeply different various conceptual schemes are from one another, I think the feminist theorist's dilemma traces all the way to how we think of representation and truth in general.

Chapter 2 From *Resisting Reality* to *Redefining Realness*

> "*From these exchanges I have learned that feminism – in the form of a tacit belief that women are human beings in truth but not in social reality – has gone deep into women and some younger men…*"
>
> - Catharine MacKinnon, *Feminism Unmodified*, 216

> "*If someone asks you to call them "they," you call them "they." But, if you're like me, you don't want to just do a thing because you dogmatically believe it's the woke thing to do. You want to understand why you're doing it. I don't just want to tolerate nonbinary people. I want to be a convert, I want to believe about them what they believe about themselves…"*
>
> - Natalie Wynn, "Pronouns" 25:46

Trans women are women.[51]

Therefore: It is true that trans women are women. And 'Trans women are women' is true. And 'Trans women are not women' is false.

Janet Mock is a woman.

Therefore: It is true that Janet Mock is a woman. And 'Janet Mock is a woman' is true. And 'Janet Mock is not a woman' is false.

[51] I'm going to treat this sentence for now as equivalent to 'All trans women are women." This of course isn't how bare plurals typically function, but I think it's a commitment that most people who use this sentence take it to stand for.

A feminist theory of gender, or of gender-ascription, should be able to honor these claims and inferences. I'm going to argue that the best way to secure this result while also getting at several parallel desiderata for a theory is to adopt an pragmatist account of the meanings of our gendered vocabulary.

I'm particularly going to be in conversation with a tradition stemming mostly from the work of Sally Haslanger on which feminist theorists conceive of themselves as engaged in *ameliorative projects*. While the notion of an ameliorative project has shifted in Haslanger's own work, there's a particularly influential conception on which the aim of these projects is to revise our concepts or word-meanings to align with our values or practical goals. Hence, for Haslanger, feminist theorizing can be judged not only on the traditional theoretical criteria of empirical adequacy and explanatory power, but on the further practical criterion of advancing our legitimate ends. As we'll see, a second criterion — one I'll call *expressive adequacy* — later emerged, most clearly in Jenkins (2016). I argue that the problems Jenkins points out for Haslanger are in fact ones that confront Jenkins' view as well. I'll then diagnose the problem: the whole literature has presupposed that the *truth* of a gender-ascription is separable from the *values* it expresses. I'll then propose a pragmatist alternative on which these two aren't separable, and I'll argue that such an alternative fares better on all our desiderata.

2.1 What Should a Theory of Gender Do?[52]

The ameliorative tradition is a pragmatist one. Central to pragmatism (in the Haslangerian tradition as well as in classical American pragmatism) is that theorizing is beholden not only to what are traditionally regarded as theoretical concerns, but also to *practical*

[52] Three of these four criteria are explicitly mentioned on p. 23 of Haslanger (2012)

concerns, e.g., our values and goals.[53] In this spirit, I want to lay out and motivate four desiderata for a theory of gender-ascription, two theoretical and two practical.

I believe that the chief epistemic virtue of a theory of gender, as of any other theory, is truth. 'True according to whom?' or 'True in what language?' we might ask. At risk of sounding flippant: True according to me — and, if I can persuade you, according to us. True in the language I'm speaking, i.e., in English.

This way of putting things is too simple, of course. For one thing, we can't assume that there's such a thing as being 'true in English' *simpliciter*. As Saul, Bettcher, and Dembroff point out, gender-terms may be polysemous or contextually variant.[54] So, there may not be context-independent facts about who is and isn't a woman, man, nonbinary, etc.[55] But even with these caveats in mind, we can say that theory itself should be true in whatever the relevant contexts are for theorizing, and that it should generate the right predictions about which particular gender-ascriptions will be true in various other contexts. As we'll see, amelioration's relationship to truth has been a source of angst in the literature. Instead of trying to lay this out at the beginning, I think the best move now is just to say that truth is *prima facie* desirable and to see in later sections how problems with truth have emerged as the literature has developed.

Second, a theory should be *illuminating* in some respect. Truth is cheap; understanding is hard to come by. Feminist theorists typically proffer theories as illuminating how gendered social practices function. There are of course many ways of getting at this — we could look at gender and law, gender and education, etc. My focus here will be on what I'll call *the structuring*

[53] Of course, there is a sense in which any theory of the social must be 'responsive to our goals and values' insofar as these goals and values are themselves among the objects of our inquiry. The pragmatist makes the further claim that values serve not only as *objects* of inquiry, but as *inputs* in the same sense as theoretical virtues like parsimony and empirical adequacy. Cf. Legg (2021) for a general discussion and Anderson (2004) for a discussion of how this manifests for specific research programs in social science.

[54] See Saul (2006), Bettcher (2013), and Dembroff (2018)

[55] Or, more carefully, about which of the relevant sentences are true.

function of gender discourse. Why is it that someone's being categorized as a woman or man has such far-reaching implications for their life chances? What is the connection between gender-ascription as a speech act and all the physical and institutional apparatus we have built up around gender?

(Why the focus on discourse, and on ascription specifically, when there are all these other less-discursive elements of gendered social structures? Two reasons: First, there is a sense in which the discursive act of categorizing someone as belonging to a gender is prior to these other elements of the structures, and always lies at their center. E.g., an infant's subjection to the system of gender as it appears in medical contexts begins with their being categorized by a doctor, and how they're positioned subsequently will depend on subsequent classificatory judgments made by other participants in these institutions. Second, ascription is important for Haslanger's project because she's chosen to approach it as one of specifying extensions for gender-terms. These definitions are themselves just very general gender-ascriptions, and their contents are designed in part to make explicit (what I'll later call) the interpellative function of gender-ascription. So, by focusing on ascription, I'm in contact with how Haslanger has framed these issues.) [56]

Now for the practical goals. First, we would like a feminist theory of gender to in some way contribute to the fight for social justice. How can a theory do this? There may be multiple ways. For one thing, as Barnes points out, understanding how gender (or gender discourse) works can be a step toward combating gendered oppression. It's hard to solve a problem that you don't understand; insofar as gender-ascription plays a role in the problem of gendered

[56] See also Haslanger (2012), 241: "Typically the act of classifying someone as a member of a social group invokes a set of "appropriate" (contextually specific) norms and expectations. It positions her in a social framework and makes available certain kinds of evaluation; in short, carries prescriptive force."

oppression, it seems like understanding gender-ascription could point toward levers for social change.[57]

Last is what I've called the criterion of *expressive adequacy*: we would like our theory to express respect — or at least not to express disrespect — for the identities of trans people. The motivation here is straightforwardly ethical: disrespecting trans identities, treating them as illegitimate, is wrong. So, while it's not obvious that a feminist theory of gender and gender discourse needs to express respect or disrespect for anyone at all, it should at least be compatible with our ethical obligations in this domain. What exactly those obligations are is itself a matter of some controversy, even among direct stakeholders. As I hinted at above, I think it involves believing that trans people are who they say they are, i.e., that their self-ascriptions are true and that incompatible ones are false.[58]

To repeat: our four criteria for a feminist theory of gender are (1) That it be true, (2) That it illuminate the structuring function of gender discourse, (3) That it be capable of contributing to fights against oppression, and (4) That it be expressively adequate, especially vis-a-vis the identities of trans people.

[57] Another possibility is that the theory could have what we might call *hermeneutical* value: it can help people to make sense of the social world and their experiences in it. This kind of project is especially important for people living at the margins, for whom the default social meanings are likely to be felt as inadequate. Historical examples of this abound: the development of the idea of sexual harassment; the very idea of the sex/gender distinction; intersectionality. In all cases, something that's articulated in a largely theoretical context makes its way into the wider culture because it helps people to interpret their experiences and themselves. This is related to the earlier theoretical point about illumination, and plausibly represents another respect in which the theoretical and practical dimensions of theorizing are inseparable.

[58] How pressing this criterion appears will depend partly on how you conceive of your theoretical project. Insofar as a theory of gender or gender-discourse remains neutral on the question of who counts as what gender — e.g., because it aims to characterize gendered practices without making any particular ascriptive claims — the criterion of expressive adequacy perhaps seems less pressing. But to what extent can a theory remain neutral? Even if your theory doesn't specify the exact extension of its terms, it doesn't follow that *no* commitments in this domain are taken up. There's always at least an implicit understanding of what your subject matter is. So, I don't think that approaches like those taken in Barnes (2017) or Dembroff (2018), which try to separate the metaphysics of gender from the ethics of gender-ascription, are viable.

These goals are interrelated. As I'll argue, the criterion of expressive adequacy puts pressure on us to have a theory that allows for trans-inclusive gender-ascriptions to be true, so truth and expressive adequacy are related to at least that extent. We also think that a theory's ability to illuminate a phenomenon is tied both to its (approximate?) truth and its utility in changing the world. Compare this to scientific theories: Einsteinian physics makes better sense of the world than Newtonian mechanics; it does this in part by making predictions that are truer to our observations; it also enables us to make GPS satellites and such. Returning to the gender cases, promotion of justice is tied to expressive adequacy insofar as expressive disrespect for trans identities is part of the social process by which trans people are oppressed.

All this has been pretty schematic so far. To see how these issues have arisen concretely in the ameliorative tradition, it's best to start with this tradition's roots in Haslanger's earlier work on the subject.[59]

2.2 Haslanger's Pragmatism

The idea of an ameliorative project was originally developed by Haslanger as a way of framing and defending her positional accounts of human kinds such as gender and race — accounts on which to be a woman, or a man, or white, or Black, is to be positioned as these things within a gendered or racialized social practice. This has the consequence that changes in our practices can induce changes in what race or gender someone is, up to and including eliminating race and gender altogether. And, since specific racial and gender categories are defined in terms of their being positioned hierarchically, Haslanger's definitions (supplemented

[59] While I'm calling these 'roots,' it's important to understand that Haslanger's own work was in part an effort to synthesize various traditions and thus was always deeply rooted in various traditions in both philosophy and feminist theory.

by the claim that racial and gender hierarchies shouldn't exist) imply that a just world would be one in which no one is a woman, man, white, or Black.

There's an obvious objection to these definitions. Almost no one — or almost no one without a PhD — would define 'man' in terms of positioning within a hierarchical social practice. And even highly attuned feminists won't tend to think that a world without gendered oppression would be a world in which, *by definition*, women and men do not exist. So, in what sense can Haslanger claim to be talking about gender at all?

Worries about how a metaphysical theory can be informative without changing the subject are familiar to analytic philosophers. But similar worries can also be found in an older philosophical tradition that Haslanger herself draws on when she characterizes herself as aiming to engage in an "immanent critique."[60] As critical theorists going back to Marx have often observed, a legitimate goal of a theory is to *critique* our ordinary concepts rather than simply taking them for granted. So, while this subject-changing worry can't be dismissed out-of-hand, it also can't be taken as conclusive. The challenge is to make sense of how ideological critique is even possible. On the one hand, we can't 'stand outside' the system of meanings history has left to us; if doing so were even possible, it would render what we say irrelevant and hence not a critique. But we also can't see ourselves as simply taking those meanings for granted. We need a sense in which we're tackling those meanings, so to speak, from the inside.

Enter the ameliorative project. Haslanger's views on what these projects amount to have changed over time, but perhaps the most concise development — and the earliest, and most influential — is this:

[60] Cf. Haslanger (2014), 24. Haslanger most commonly talks about "critical theory" and "ideology critique" rather than "immanent critique;" I choose the latter term because it emphasizes the point, as Stahl (2014) puts it, that the critique "must accomplish the difficult task of taking up a stance that is both appropriately critical of, and sympathetic to, the self-understanding of those whom it addresses" (Stahl (2014), 5).

> *On this [ameliorative] approach the task is not to explicate our ordinary concepts; nor is it to investigate the kind that we may or may not be tracking with our everyday conceptual apparatus; instead we begin by considering more fully the pragmatics of our talk employing the terms in question. What is the point of having these concepts? What cognitive or practical task do they (or should they) enable us to accomplish? Are they effective tools to accomplish our (legitimate) purposes; if not, what concepts would serve these purposes better?*[61]

Ameliorative approaches to meanings differ from those in conceptual analysis or lexical semantics in that they are explicitly about *improving* our representational devices to better suit our needs.Part of what makes social phenomena theoretically and practically interesting is that, as objectively real as they may be, they're always in some sense *up to us*. They are not forced upon us by God or nature; we can change them. This theme of the mutability of the social arises in multiple ways across Haslanger's writings. First, there's the fact that the social phenomena she's describing can (and in many cases should) change. This, in fact, is the point of her book's title: social groups like genders and races are socially constructed and real, but should be resisted, i.e., dismantled. Second, and more pertinent to this paper, the meanings of our words or concepts are up to us — they reflect our values and can do a better or worse job of serving our interests.

Haslanger's earliest move, then, was to allow that her definitions may not be literally true in English, but that they *ought* to be true — that we ought to adopt definitions of 'man' and 'woman' like those she offered. Here the response to the subject-changing worry is that it is, in an important respect, beside the point.[62] While Haslanger's positional definitions may not be *true*

[61] Haslanger (2012), 224

[62] For the sake of making my argument more clearly, I'm glossing over important details of Haslanger's early views. While early Haslanger was open to the thought that her views were revisionary (or what we might pejoratively call "subject-changing"), she was always clear that the practical aims of her theory could not be

in ordinary English, they possess the theoretical and practical virtues of illuminating how gender structures our lives and hopefully contribute to combating gendered injustice. Even if this counts as changing the subject, it's changing it from a subject that harms us and clouds our understanding to one that helps and clarifies.[63] But how exactly does it help and clarify?

As for clarification, Haslanger claimed that most people's intuitive ways of thinking about gender were *ideological*, masking gendered privilege and subordination by allowing people to refer to biology as the basis for social relations.[64] This is, of course, a pervasive theme in the history of feminist theory that predates even Beauvoir's inauguration of its second wave. Haslanger hoped that her positional definitions could destabilize sexist ideologies by making salient the ways in which gendered discourse helps to produce social structures that are both contingent and unjust. Even if her audience's initial response to her definitions was to reject them as obviously out of sync with what the relevant words mean, they might notice — despite themselves — that the definitions tracked an important dimension of how their gender-ascriptions function socially.

Haslanger was also optimistic that centering genders as social structures could help us, as theorists, to make sense of various gendered phenomena. To foreshadow, she thought that the important notion of gender identity could be analyzed in terms of gender as a social structure:

---

achieved unless the theory overlapped considerably with existing discursive practices. For example, it mostly preserves the actual extension of the term 'woman' despite revising the intension quite a bit.

[63] A second interpretation holds that Haslanger's ameliorative definitions are not revisionary, but in fact capture what we've been talking about all along. The idea here is that linguistic meanings are in an important sense determined by normative considerations, such that the fact that we *should* be talking about something is reason to think we *already are* talking about it. This reading is similar to the one developed in Barnes (2017) according to which Haslanger is a kind of robust (meta-)ontological realist. On this reading, Haslanger isn't saying that there are no privileged ways of categorizing and that we should just pick the ones that suit our purposes; she's saying that a definition's suiting our purposes is *evidence* that the category it picks out is objectively privileged. A version of this view was developed by the publication of Haslanger (2012) and has continued to be developed in more recent works, e.g., her (2020a) and (2020b).

[64] Cf. Haslanger (2012), chapter 3

identity consists in a certain psychological relation one might bear to the norms determined by the structures in which one lives. The idea here is not that structures are more fundamental than these other phenomena in some robust, ontological way. Rather, it was that centering structures — in what she calls a "focal analysis" — is helpful for organizing and making sense of the relevant phenomena in a systematic way.[65]

I want to note two ways in which Haslanger, at least at this point in her career, differed from the way in which I'm construing the whole project of feminist theorizing about gender. First, while I've said I wanted to focus on the structuring function of gender-discourse, Haslanger's focus was on gendered social structures. These are different things. Second, Haslanger had not yet identified expressive adequacy as an important criterion. This second point is the one I'll address first, since it's the one that's been addressed most clearly and has gone on to shape subsequent literature.

2.3 Jenkins and Expressive Adequacy

One famous objection to Haslanger's social position definitions is that they are exclusionary to trans people. In making one's status as a woman or man depend upon how one is perceived and treated by others, Haslanger's definitions implied that some trans people who do not 'pass' do not belong to the gender with which they identify.[66] Katherine Jenkins tried to fix this while accepting Haslanger's modest pragmatism by offering an alternative, identity-based ameliorative definition of 'woman.'

Jenkins' view is pluralist. It acknowledges that the social-structural phenomena Haslanger highlights are real and important, both from the viewpoint of social science and from

[65] Haslanger (2012), 8
[66] See Jenkins (2016), 398 for a detailed explanation of why Haslanger's view entails this.

that of feminist theory and practice. But Jenkins argues that, given the historical marginalization of trans women both in society at large and within feminist movements, practical considerations tilt decisively in favor of using 'woman' to refer to those with a female gender identity rather than those socially positioned as women.

Jenkins' theory of gender identity — which draws on Haslanger's theory of racial identity — is a rich one that deftly navigates the issue of how we can be in some sense *responsive to* or *evaluable under* norms without having internalized them, or indeed having any very specific kind of relationship with them. What's relevant here, though, is how Jenkins' approach resembles Haslanger's methodologically while differing from it in substance. As we've seen, Jenkins accepts the legitimacy of an ameliorative project as "arriving at a concept … that a particular group should aim to get people to use, given a particular set of goals that the group holds."[67] This project is "revisionary"[68] in the sense that it involves positing meanings for terms that may be out of sync with how those terms are used in most contexts. Jenkins differs from Haslanger in giving center-stage to a previously overlooked set of stakeholders — trans women — and identifying respect for those stakeholders as a key practical desideratum for an ameliorative inquiry. Thus we end up with a different output in the form of an identity-based, rather than a positional, account of what it is to be a woman, though the ameliorative character of the process remains roughly as Haslanger envisioned it. And while Jenkins' identity-based account improves on Haslanger's in important ways, it also inherits one of its central problems: its difficulty with truth. If it presents itself as a proposed revision to the truth-conditions of our words or concepts, it cannot present itself as saying something that's already true in our shared

[67] Jenkins (2016), 395
[68] Jenkins (2016), 395

language. Indeed, this shortcoming of the revisionary ameliorative project comes into even sharper relief once we've identified expressive adequacy as a criterion for feminist theory.

While Jenkins' work has become a canonical recent articulation of the need for expressive adequacy, Jenkins is not the first person to notice this need. And, as others have noticed, it's not clear that we can get expressive adequacy without also regarding our gender-ascriptions as true and trans-exclusive ones as false. Perhaps the clearest case of this is in Saul (2006). There Saul suggests a contextualist reading of Haslanger's project on which her ameliorative definitions are "not proposals about what these should always mean, but instead about what these terms should *sometimes* mean."[69] This would allow us to get around the subject-changing worry by acknowledging that we're stipulating a non-standard meaning for use in constrained, theoretical contexts. But as Saul points out, by conceding that ordinary speakers' claims might be mostly true, we commit ourselves to *agreeing* with the very claims that Haslanger wants to critique as ideological. Suppose, for example, that Bob utters 'Carol is not a woman,' where Carol is a trans woman. The ameliorative approach is compatible with — indeed is tailor-made to be compatible with — Bob's utterance expressing a true proposition. But if we acknowledge Bob's utterance as true, then (assuming that we and Bob share a language) there's no good way for us to avoid committing ourselves to the claim that Carol is not a woman. The ameliorative approach threatens to doom us to joining Bob in his transphobia.[70] This problem persists even if we abandon the 'contextualist' understanding of the ameliorative project and insist, with Jenkins, that trans-inclusive meanings are the only ones we should use. So long as the ameliorative project is understood as revisionary — as aiming to induce a change in the truth-

[69] Saul (2006), 140

[70] This worry is familiar from the literature on slurs: any view on which a sentence containing a conventionalized slur is *true* commits us, via disquotation, to the slurring sentence.

conditions of gender-ascriptions — we lack the resources to say that Bob's transphobic utterances are false.

If what I've said about Bob's case is right, then a commitment to expressive adequacy vis-a-vis trans identities entails engaging with dominant, transphobic discursive practices on something like their own terms. Trans women are women. Trivially, that trans women are women entails that 'trans women are women' is true. Respect for the identities of trans women requires believing that they are who they say they are — that they are women. So, it requires believing that it's true that they are women, that 'trans women are women' is true, etc. Ameliorative approaches — at least on the revisionary interpretation — don't secure this result. The best they can allow is that, while 'trans women are women' isn't true, and therefore trans women *aren't* women, they *ought* to be women.[71] Some people — including some trans people — find this to be satisfactory. I do not.

2.3.1 A Puzzle about Structuring Function

One might be tempted to think that, because Jenkins' account de-centers structure in the way it does, it would therefore sacrifice some of the explanatory power of Haslanger's account. But it doesn't, and I think the fact that it doesn't ends up being instructive.

Jenkins characterizes her view as *pluralist* in the sense that it allows for the existence of both positional and identity-based gender concepts or gender kinds. She acknowledges that the phenomena Haslanger identifies are important, and early on in her (2016) presents this as good

[71] Two complications. First, for the conceptual engineer to summarize their position by saying that trans women ought to be women may seem to involve a use-mention fallacy. But it needn't do so: on some constructionist understandings of gender, the conceptual change advocated by the conceptual engineer would in fact change who is a woman (Cf. Dembroff (2016) (MS)). Second, we might think that 'woman' has different meanings to different groups of people, and this could allow us to say that many trans people's self-ascriptions are true (Cf. Bettcher (2013)). This point is well taken, but it still permits trans-exclusionary gender-ascriptions to be true as uttered by members of dominant linguistic communities.

reason for feminist theorists to hold onto a concept of what she calls "gender as class," complete with definitions of what it is to be classed as a woman or man that closely resemble Haslanger's definitions.[72] The claim, then, appears to be that there's room for both types of account.

Despite her pluralism, Jenkins argues that, in light of practical considerations, the word 'woman' should be reserved exclusively for the concept of gender as identity.[73] That is, while it's important for us to understand the social dynamics that undergird the concept of being *classed as* a woman, we shouldn't build these dynamics into the definition of what it is to *be* as woman.

One of the key virtues of Haslanger's account is supposed to be its unmasking potential: it illuminates the ways in which being classified as a woman positions one in a hierarchical set of social practices. But why think that the way to do this is to *define* 'woman' as one who is so positioned? What is gained by adopting this definition rather than just describing the relevant social structures while remaining noncommittal about who is and isn't a woman? If Jenkins is right, then the answer is: nothing, really. Nothing is lost, explanatorily speaking, by describing some people as being 'gendered as women' or 'positioned as women' while remaining agnostic about whether they're women. And something is gained expressively, because we've left ourselves room to reserve the title of 'woman' for all and only those who want it.

Recall, though, that one of the things we might want a feminist theory of gender to do is to explain how gender-ascription functions pragmatically, e.g. what its illocutionary effects are. I think attending to this fact helps us to understand why Haslanger made the move she did. The idea is that, at some level of granularity, there are sets of normative consequences that come with gender-ascription as a matter of these ascriptions' conventional meaning. These conventions,

[72] Jenkins (2016), 408
[73] Jenkins (2016), 419

though, are at the level of pragmatics: they're part of what we *do* with gender-ascription rather than what we *say* with it. Because this pragmatic force is left implicit, though, it's easy for speakers to overlook it and see themselves as merely stating facts.

On this reading, the point of Haslanger's definitions is to make the pragmatic force of gender-ascription explicit as part of its content. If you can make people conscious of how their discourse functions to subordinate, it's harder for them to justify it. But the tension between a theory reflecting a practice and its critiquing the practice reemerges here. Making the structuring (i.e. subordinating) function of gender-ascription explicit in its content will tend to disrupt that function. But this means that an interpretation of our speech on which we're *describing* people as subordinated cannot explain how this speech *functions* to subordinate.

This might sound like an odd complaint. After all, the ameliorative project — at least on this interpretation of it — was never intended to capture what we already meant by our terms. But that's exactly the problem. Amelioration, understood as an instrumentally motivated revision in our concepts or our language, effectively gives up on the project of immanent critique. An immanent critique, recall, engages with dominant ideologies on their own terms without accepting those terms uncritically and as they're found. If a definition of 'woman' is going to work as part of an immanent critique of our gendered practices, it must be recognizable to us as correctly characterizing those practices as we find them. And the ameliorative definitions — Jenkins' as much as Haslanger's — fall short in this respect. It's true that gender-ascription functions to situate people in gendered social practices, and both accounts try to capture this fact. But Jenkins places this structuring function outside of the meanings of the gendered terms, while Haslanger locates the function in the wrong aspect of the terms' meanings — in their truth-

conditional contents. The lesson here is that a successful theory of structuring function will have to account for how this function outstrips the descriptive or assertional content of an ascription.[74]

2.4 Diagnosing the Problem

So far I have argued that ameliorative projects, as understood by Jenkins and (at least in her earlier works) Haslanger, fall short of the requirements of an immanent critique. Such a critique must present our practices to us in a way that keeps them recognizably *ours* while also revealing their shortcomings. Because amelioration aims to *revise* our meanings, it cannot critique these meanings from the inside. This difficulty emerges in multiple guises.

First is amelioration's relationship with truth. Here we see the ameliorative project as insufficiently immanent and — as a result! — insufficiently critical. If a definition of 'woman' is revisionary in the sense that it proposes a discontinuous break from our current ascriptive practices, then it follows that the definition can't be recognizable to us as characterizing those practices. And for that very reason it cannot show those practices to be wrong on their own terms — we saw this with the uncomfortable possibility that, even if we can establish contexts in which trans-inclusive gender-ascriptions are true, we may not be able to secure the result that trans-exclusionary ones are false. Thus the failure of our ameliorative definition to be true in the language of our interlocutors also jeopardizes its expressive adequacy.

Second, the ameliorator's desire to posit definitions which break with or disrupt the oppressive functions of gender-ascription turns out to be in tension with the desire to make sense

[74] This problem also emerges for Ásta's conferralism. She distinguishes between conferred properties and base properties, the latter being 'what speakers are trying to track.' Given that speakers are explicitly trying to track something *other than* the conferred property, it seems odd to say that the conferred property is what they manage to talk about when they use a given term. How is it that they 'miss' the base property and refer to the conferred one? I don't think there's a convincing story here. I think the problem Ásta's running into is that she wants to make sense of how ascriptions function socially, but her representationalist commitments lead her to trying to build that function into truth-conditions in a perverse way.

of these functions as part of a gender-ascription's meaning. This tension appears for Haslanger as a definition which tries to make explicit gender-ascription's social function, but does so by positing meanings that can't be regarded as ours. For Jenkins, it appears in the form of a pluralist gender-ontology that tries to make sense of structures while giving pride of place to identities — but again ultimately posits a definition that's radically out-of-touch with dominant understandings and hence cannot effectively critique them.

One possible move here is just to give up on immanent critique as I've characterized it. Maybe we don't need a single theoretical apparatus to give us a metaphysics of gender, a semantics of gender-terms as they're used in dominant (i.e. trans-exclusionary) contexts, *and* a critique of those very structures and meanings. This bifurcation strategy has arguably become a dominant thread in recent literature, being endorsed in various forms by Barnes, Jenkins, and Dembroff. Jenkins and Dembroff are particularly clear about this, claiming that we can adopt a metaphysical account on which some trans women are socially constituted as men while objecting to that constitution on political grounds and refusing to participate in it in our everyday gender-ascriptions. The trouble with this, in my view, is that a metaphysical account on which some trans women are constituted as men *just is* a very abstract way of misgendering these trans women. Because our metaphysical theorizing is part of the very practices it seeks to interpret; we cannot simply leave it in the ontology room. This issue is brought into especially sharp relief in Dembroff (2018) where they entertain the possibility that we may be obligated to believe that trans people's self-ascriptions are true even if the evidence points to their being false.[75] I can only speak for myself here: I literally cannot believe that trans people are who they say they are while accepting a metaphysics on which they're not. One or the other has to give. Fortunately, I

[75] Dembroff (2018), 45

think this bifurcation between theory and practice can be avoided — and the project of immanent critique saved — by shifting to a more thoroughgoing pragmatism.

2.5 Pittsburgh Pragmatism[76]

The type of pragmatism I'm advocating for is inspired by what's sometimes called 'The Pittsburgh School,' and especially by the work of Robert Brandom, Quill Kukla, and Mark Lance.[77] On this approach, the meaning of a linguistic expression is a matter of the expression's functional role in a norm-governed discursive practice. Rather than starting with the idea of bits of language attaching to objects in the world, we start with normative statuses — sets of commitments and entitlements — attaching to participants in a discursive practice. These normative statuses can be altered by discursive moves, paradigmatically by the utterance of a sentence. The meaning of a subsentential expression is then a matter of how its embedding in various sentences works to determine the discursive moves a speaker makes in uttering those sentences.

The advantage of this approach, from the viewpoint of immanent critique, is that it is self-consciously an attempt to make sense of our discursive practices from within those very practices. It acknowledges that linguistic meaning becomes intelligible as such only in the context of, and through our participation in, a discursive practice which is irretrievably social. The thought then is that talk of what our words mean, however abstract it may get, is always the activity of socially embedded beings trying to make explicit for themselves what was already implicit in their practices. If in this process the we encounter something in our practices that we

[76] Henceforth I'll use the term "pragmatism" as a shorthand for "Pittsburgh pragmatism." While, e.g., Haslanger is a pragmatist in her own way, I'm talking here about pragmatist theories of truth and meaning which she explicitly rejects.
[77] Specifically, Brandom (1994) and Kukla and Lance (2009)

cannot accept, then we'll have to revise those practices to resolve the contradiction. This is immanent critique.[78] There are two main theoretical tools I want to lay out: *deontic scorekeeping* and *interpellation*.

2.5.1 Deontic Scorekeeping & The Two-Sided View of Concepts

Brandom's notion of deontic scorekeeping is influenced by Lewis' "Scorekeeping in a Language Game."[79] Where it differs is in what it takes interlocutors to track. While Lewis saw us as tracking things like presuppositions, contextual relevances, etc. Brandom sees us in the first instance as tracking *normative statuses*. His clearest (relatively brief) articulation of this ideas is as follows:

> *Deontic scores consist in constellations of commitments and entitlements on the part of various interlocutors. So understanding or grasping the significance of a speech act requires being able to tell in terms of such scores when it would be appropriate (circumstances of application) and how it would transform the score characterizing the stage at which it is performed into the score obtaining at the next stage of the conversation of which it is a part (consequences of application).*[80]

Central to this picture is what Brandom, following Dummett, calls a "two-aspect model" of concepts. The idea is that a concept will have both inputs and outputs, corresponding to circumstances under which its application is appropriate and consequences of its application. While thinking of concepts as functions is commonplace, this picture differs from the standard representationalist one in that the inputs and outputs capture more than just the putative

---

[78] Brandom's conception of this explicitating project is inspired by Hegel's view of the development of consciousness. And like Hegel, Brandom sees this process as one of the agent reconciling itself to the community by taking on an understanding of its practices on which the practices are rational. Hence *Making It Explicit* begins with an excerpt from T.S. Eliot's "Four Quartets" which reads: "And the end of all our exploring / Will be to arrive where we started / And know the place for the first time." The idea here is that, in making sense of our practices, we can transform our understanding of them without ever abandoning them. But, as Marx showed, we can take on this explicitating project without assuming that its final end is to reconcile us to our starting-points.

[79] Lewis (1979)

[80] Brandom (1994), 183

extension of the concept. On this picture, the successful application of a concept to an object returns as its output not (just?) a truth value, but a cluster of normative statuses capturing the discursive and practical commitments one takes up for oneself and attributes to others in applying the concept.

Here's an example. Suppose I assert that Spot is a dog by uttering the sentence 'Spot is a dog.' The kind of pragmatism under discussion here will want to understand the meaning of this sentence or assertion by asking two questions: What are the normative grounds for my being committed or entitled to assert what I do? And what further commitments and entitlements follow from my asserting it? The first question is best gotten at via the second. One of the central commitments one takes up in making an assertion is a commitment to demonstrating one's entitlement to the assertion if challenged. If someone challenges my assertion that Spot is a dog, I can demonstrate my entitlement to the assertion by, for example, providing evidence in the form of a photograph of Spot. And if things are happy — if my interlocutors are satisfied with my photograph or if they simply believe me without asking for proof — then they too will come to be entitled to make that claim in the future. The normative consequences of my assertion therefore extend beyond me as an individual; an entitled assertion on my part can ground someone else's entitlement to an assertion with the same content.[81]

A crucial aspect of all this is that the inputs and outputs of a speech act can involve non-linguistic phenomena in important ways. One of the things that can entitle a speaker to claim that something is red is its being (visibly?) red. And the commitments we take up in making claims are not just evidential but also practical. If Spot is running amok and I say to horrified onlookers

[81] Of course, Brandom cannot here help himself to a standard notion of content-as-intension. The second chapter of *Making It Explicit* is devoted to developing an idea of semantic content grounded in the idea of inference as a discursive activity.

"That's my dog," I commit not only to proving that Spot is in fact my dog but also to cleaning up after him.

While Brandom acknowledges these observational and practical dimensions of meaning, he always gives pride of place to the act of assertion. And while I don't deny that gender-ascriptions are assertions, I want to focus on a different (though perhaps related) aspect of their meaning: the ways in which they position people in gendered practices. For this it will be helpful to turn to Kukla and Lance.

2.5.2 Interpellation

While Brandom gives the act of assertion a central place in his account of discursive practice, Kukla and Lance emphasize the variety of different normative act-types available to us. The normative grounds and consequences of promising or consenting will differ systematically from those of an asserting. Perhaps chief among these is that promising and consenting generate commitments and entitlements to *actions* rather than to claims or to bits of evidence.

In Chapter 8 of their book, Kukla and Lance discuss a kind of pragmatic function which, following Louis Althusser, they call *interpellation*.[82] To interpellate someone is to place them into "a particular, concrete location in normative space."[83] An (ultimately too simple) example of this can be found in the children's game of tag. Whoever is 'it' is uniquely entitled to tag other players by touching them. This act of tagging shifts the tagged player's position in the system of norms constitutive of tag. They become it, which is to say they become committed to chasing the other players and entitled to interpellating them as it by tagging them.

[82] Kukla and Lance (2009), 134
[83] Kukla and Lance (2009), 180

But the tag example is too simple because interpellation, as Kukla and Lance characterize it, has both alethic and constative elements: "it outstrips its own recognitive content, recognizing someone as *already being* a particular person with a normatively defined identity, but at the same time helping to constitute and solidify this identity."[84] This point is crucial for present purposes, because it's here that the normative commitments voiced by a gender-ascription will make contact with our assertional commitment to the ascription being true. By interpellating someone in a practice, I hold them to the norms constituting the practice. But I don't see myself as having arbitrary authority to decide how people are rightly placed. Unlike the child who can make someone it simply by tagging them, an interpellative act purports to be warranted by something outside the act itself.

Take the old example of an umpire calling balls and strikes. In one way, an umpire who shouts "Strike one!" constitutes the pitch as a strike — the umpire's call is what ultimately determines how the pitch will affect the numbers on the scoreboard. But, as any engaged spectator will attest, the umpire's authority here is not arbitrary. They can *get it wrong*, e.g. by calling a pitch below the knees a strike. Indeed, it's perfectly natural to respond to such a call by saying 'That wasn't a strike,' though we all know it will be recorded as one. So what gives? Does the umpire constitute pitches as strikes or merely observe that they are? The answer is that they do both. The umpire's call normatively positions each pitch as a ball or strike; but the difference between a good umpire and a bad one comes down to their skilled judgment in recognizing which pitches *really are* strikes.[85]

---

[84] Kukla and Lance (2009), 183

[85] Cf. Brandom (1994), 184: "But though the attitude of the umpire does determine the status of a throw as a strike for official scorekeeping purposes … the use of nonscorekeeping vocabulary in stating the rules … establishes a perspective from which the judgment of the umpire can nonetheless be understood to be *mistaken*." That these calls have alethic import is further reinforced by the phenomenon of 'robo-umps,' or automatic ball-strike systems. While

2.5.3 Applying to Gender

Here, then, is the basic picture of how this framework would apply to gender-ascription. The meaning of 'woman' is understood in terms of the normative statuses that attach to someone when they're positioned as a woman in a gendered social practice. But rather than building this social positioning into the descriptive content of a gender-ascription, we see gender-ascriptions as inducing shifts in deontic score — i.e., shifts in the normative statuses interlocutors claim for themselves and attribute to others. By saying 'Sam is a man,' I attribute to Sam whatever commitments and entitlements come with the status of being a man in the relevant context. And, just as importantly, take up a commitment of my own to practically recognize Sam — that is, to *treat* him — as possessing those commitments and entitlements. In the context of sex-segregated housing or hygiene facilities, for example, declaring that Sam is a man will involve recognizing him as entitled to enter men's spaces but not women's.[86]

This is not to deny that, in uttering 'Sam is a man,' I normally am also making an assertion and taking up ordinary assertional commitments (to providing evidence, for example). Indeed, this assertional dimension is crucial to the interpellative function of my ascription. The concept *man* — like the concepts *dog*, *strike*, and *President* — is two-sided: it has both consequences of application and conditions for its application being appropriate and hence entitled. Just as with any other assertion, my assertion that Sam is a man commits me to defending the assertion if challenged — i.e. showing that it was entitled by Sam's meeting the

some decry the loss of magic in removing a human element from the game, the obvious appeal of such systems is that they're *more accurate* — a notion that wouldn't make sense unless we understood the human umpire's job as consisting in tracking some state of affairs that obtains independently of their calls.

[86] Things are more complicated than this, of course. Men might be permitted to enter women's restrooms in an emergency, for example.

concept's application-conditions. Challenges of this sort amount to demands that I produce some reason to think my assertion (in this case, my gender-ascription) was true.

2.5.4 Truth, Extension, Definition

One last piece of the puzzle needs to be put in place here. As I argued earlier in the paper, feminist metaphysicians cannot meet the practical criterion of expressive adequacy while disregarding the epistemic criterion of truth. This is where I think the advantage of expressivism lies. While early forms of expressivism distinguished themselves by denying that expressive language is truth-apt, it has more recently become commonplace for pragmatists to extend their expressivism to truth-talk itself.[87] The details vary from one account to another, but the basic idea is that we can understand truth talk in terms of the expressive meanings of the sentences to which the truth predicate is applied. Consider, for example, the following sentences:

1. Janet is a woman.
2. 'Janet is a woman' is true.
3. It's true that Janet is a woman.

(2) and (3) follow from (1) via trivial disquotational inferences. The pragmatist can then maintain that inferences of this sort exhaust the meaning of 'true,' and hence that claims made using the truth predicate have no substantive meaning that is not inherited from the sentences to which the truth predicate is applied.[88] Once we have in hand the idea of 'truth' as a way of

[87] The stock example of early expressivism in metaethics is Ayer's emotivism, while the truthy kind of expressivism is famously exemplified by Gibbard and Blackburn.

[88] It's possible to get into the weeds regarding how a redundancy or other deflationary theory of truth can handle things like indirect discourse, embeddings, quantification over propositions, etc. I will not articulate, much less

endorsing sentences (or sentence-tokenings, or the discursive moves made by those tokenings), we can use this to make sense of talk about reference and extension, as well as indirect discourse. For example,

4. Janet is in the extension of 'woman.'
5. 'Woman' refers to all and only people with a female gender identity.
6. What Janet says is true.

(4) is just a way of saying that the result of applying the predicate 'woman' to Janet is truth — i.e., that Janet is a woman. (5), analogously, says the same thing of all and only those with a female gender identity. (6) is a way of echoing some claim of Janet's. If the relevant claim of Janet's is a claim to be a woman, then an utterance of (6) will amount to a claim that Janet is a woman.

Perhaps this will be felt as anticlimactic. Why make so much noise about truth if truth itself is, so to speak, nothing to write home about? I think the best answer to this question is to reverse it: if truth is some substantive relation that obtains between our judgments and the world — if it is in that way outside our epistemic and practical perspectives — then why should we care about *that*? Indeed, I suspect that this image of truth as something alien to our perspective is part of what's made philosophers both within and outside the amelioration literature question how important it is for their theories and their individual gender-ascriptions to be true.

defend, any particular technical theory here. If it turns out that no expressivist theory of truth is workable, then that would torpedo the argument of this paper. But I'm comfortable assuming that some such theory is workable.

Conversely, the overriding importance of truth becomes evident once we acknowledge that our view of what's true is our view *simpliciter*.

2.6 The Payoff

Now, with all that groundwork in place, I can say how expressivism solves the ameliorator's problems. First, and most crucially for my purposes, it gives us a way to satisfy both the practical criterion of expressive adequacy and the epistemic criterion of truth. It achieves this by essentially collapsing them into one and the same criterion. Feminist theorists can start with a practical commitment to respecting the identities of trans people. Following through on this practical commitment requires discursive recognition of trans women as women. That recognition, in turn, involves taking and treating trans women's identity claims as true and trans-exclusive claims as false. Once we have the values right, we get the facts for free.

Of course, getting the values right is hardly a trivial matter. While most contemporary feminist theorists will want to claim that transphobes' gender-ascriptions are false in virtue of voicing the wrong values, the transphobes could well say the same thing of the feminist theorists. Nothing I've said in this chapter settles that dispute. But I do think it lays the groundwork for progress by revealing its fundamentally normative character. While normative disagreements aren't exactly famous for being easily resolved, taking the fight onto explicitly normative ground prevents people from relying on the ideological view of gender as simply a natural fact.

As the earlier case of Bob illustrates, the pragmatist account can retain the debunking potential of Haslanger's approach. While expressivism allows us to regard our gender-ascriptions as true, it also makes explicit that gender-ascription is a way of taking on a practical commitment the content of which is defined by a social practice. In so doing, it invites all parties to a dispute over gender to take responsibility for articulating the normative contents of their claims and for

justifying those claims as social actions rather than as normatively inert statements of fact. The aim of Chapter 3 is to make some progress on this, sketching a picture of the normative contents of identity claims in general and identifying some principles we can appeal to in justifying ourselves to one another.

What's more, positing this constitutive connection between gender-ascription and gendered practices arguably does an even better job of illuminating the structuring function of gender discourse than Haslanger's view does. While Haslanger's definitions make explicit the fact that being categorized as a woman places one in a social hierarchy, building that hierarchy into the definition of the term does nothing to clarify why categorization should have the effects it does. After all, I can recognize someone as being positioned in a system of norms without seeing myself as having any reason to treat her as those norms prescribe. The two-aspect model, as articulated by Brandom, solves this problem. This model makes sense of how the structuring function of gender-ascription could be built into these ascriptions' meanings while also outstripping their merely assertional content. The assertional aspect comes in on the side of input, or the circumstances under which the concept is appropriately applied; the structuring function comes in on the side of the output, or the normative consequences of its application.

In short, I think a pragmatist approach to gender discourse has the potential to do it all. Not only does it tie expressive adequacy to truth, it also allows us to promote social justice via the routes proposed by both Haslanger and Jenkins. Disrespect for the identities of trans women is an important aspect of transphobic oppression; a pragmatist view allows us to repudiate this disrespect in a particularly full-throated way. Understanding the structuring function of gender discourse is an important step in dismantling oppressive social practices; a pragmatist view does more to clarify this function than a traditional representationalist view could.

2.6.1 Gender Disagreement as Normative Disagreement

We saw earlier that an ongoing worry for Haslanger's account is that it 'changes the subject,' or that adopting her proposed definitions will result in 'talking past' people who continue to use gender-terms in the standard ways. Put another way, this is the old problem of immanent critique: we want to critically engage with dominant ideologies while acknowledging that we cannot just stand outside of them; we want to critique them without accepting them. Expressivism, it turns out, has a familiar way of handling this sort of problem.

One familiar argument for expressivism in metaethics is that it can make sense of moral disagreement better than can traditional (i.e., representationalist) realist views. If the meaning of 'right,' is just its conventionally determined reference to a particular partition of logical space, then how can two thinkers or two communities substantively disagree with one another about what is right? For concreteness, imagine two isolated communities, one of which uses 'right' in a way that corresponds to a deontological ethics (e.g., torture never gets classified as 'right') and the other of which uses it in a utilitarian way (e.g., torture is 'right' if it maximizes utility). There's a clear sense in which these communities seem to disagree with each other: one *treats* utility-maximizing torture as obligatory while the other treats it as impermissible. The pragmatist explains this by saying that the meaning of 'right' is not (only or principally) its extension, but its role in practical deliberation or social practice. Because 'right' plays this same action-guiding role in both communities, it has the same meaning in both communities. Hence, they mean the same thing by 'torture is sometimes right' regardless of how robustly their dispositions to apply the term 'right' to acts of torture may differ. On the other hand, if the meaning of 'right' is just its extension, then it seems we'd have to interpret them as talking past one another.

If gender-ascriptions voice a kind of normative judgment, then we might expect this dynamic to reappear in the case of disagreements over gender. Returning to Bob: we can say that we and Bob have a shared understanding of the action-guiding significance of someone's being a woman, but disagree about who *is* a woman, i.e. about the extension of the term. This disagreement in extension is compatible with our meaning the same thing by the term, because there's a continuity of normative role. The pragmatist about gender thus avoids the subject-changing worry.

Translating this into the kind of expressivism I've endorsed here, we'll say that we share with Bob an understanding of the consequences of application for the term 'woman' — we agree, at least to some extent, on the commitments and entitlements that come with being a woman. Where we disagree is in the term's application-conditions, or in the grounds for entitlement to the relevant normative status. Bob thinks that entitlement to that status depends on reproductive anatomy, while we think that it depends on identity. It's possible to make this disagreement explicit using conditionals taking application-conditions as antecedent and normative outputs as consequent: "If you sincerely identify as a woman, then you are one."

But, just as metaethical expressivism does not by itself say whether the utilitarian or deontologist is correct, this pragmatist picture of our disagreement with Bob doesn't settle the dispute. For that, we'll need to say something about what might ground a person's legitimate entitlement to occupying a given normative status. Saying more about this is the goal of the next chapter. For now, I think we can say that reproductive anatomy *is not* a good basis on which to coercively assign gendered statuses which will in turn determine so much about people's life chances. Basic considerations of autonomy militate in favor of at least a *prima facie* presumption

that people should be able to occupy statuses of their choosing, especially when those statuses play such a central role in structuring our lives.

2.6.2 Redefining Realness

In the classic documentary *Paris is Burning*, interviewee Dorian Corey articulates a notion of 'realness' as it was understood by some in the Harlem ball subculture of the 1980s. Corey says that realness is "to be able to blend," to pass as "a real woman … a real man, a straight man." This concept became the namesake for trans activist Janet Mock's memoir *Redefining Realness*. As the title suggests, Mock believes that the notion of realness is flawed because it gives power to dominant, transphobic understandings of what it is to be a woman. Mock, as I read her, urges us to adopt an alternative notion of realness as something like authenticity or trueness to oneself. This authenticity is understood in the context of a lifelong project of "self-definition," which Mock characterizes as "a responsibility" with respect to "the many varied decisions that we make to compose and journey toward ourselves."[89] Mock's conception of realness clearly signals her rejection of dominant understandings of gender as a biological fact. But it also signals a rejection of any kind of deconstructive skepticism. Gender is neither a natural fact nor an ideological myth; it is something that we build up and make real through our actions.

I believe that something like the notion of realness proposed by Mock can be captured by moving toward a pragmatist account of gender-ascription. Dominant representationalist approaches to gender concepts have had trouble with expressive adequacy because they posit a gap between truth and judgment, fact and value. A pragmatist account could solve this problem

[89] Mock (2014), 172

by instead positing a constitutive connection between the truth of a gender-ascription and the normative commitments voiced by it.

I do think it's important here to acknowledge where I depart from Mock, and why I think those departures are both justified and still in the spirit of what she's saying. Mock's rejection of conventional understandings of gender leads her toward a very individualist line on which one's self-concept is essentially private and doesn't have to answer to others. This is fundamentally different from the picture I've suggested, which centers the communicative act of *ascribing* a gender, not only to oneself but to others. The third chapter of this dissertation will be devoted to exploring why I think our identities must be answerable to others. For now, I'll say that, to whatever extent one does or doesn't have to take others' perspectives into account when journeying toward oneself, it's definitely going to be important that we be recognized as being who we say we are.

Chapter 3 Identity, Autonomy, and Amelioration

> "*Thus, in the sciences of man in so far as they are hermeneutical there can be a valid response to 'I don't understand' which takes the form, not only 'develop your intuitions,' but more radically 'change yourself'*."
>
> - Charles Taylor, "Interpretation and the Sciences of Man," 54

> "*Speech and language, however ceremonious, complex, and convoluted, are a way of revealing one's nakedness; and this revelation is, really, our only human hope. But this hope is strangled if one, or both of us, is lying.*"
>
> - James Baldwin, *The Evidence of Things Not Seen,* 43

3.1 Recap

Chapter one laid out some problems for the idea of rationally warranted conceptual change. The crux of these problems was that deep conceptual change cannot be a product of instrumental reasoning, as this would require the reasoner already to have access to all the relevant concepts. I then proposed an alternative picture on which rationally warranted conceptual change consists in a kind of non-instrumental reasons-responsiveness. Rather than voluntarily selecting from a suite of equally intelligible options based on which option will bring about the best consequences, this kind of change involves encounters with the world and with the ways in which our existing concepts fail to make sense of it.

Chapter two turned to feminist theories of gender, the metaphysics of gender, and the ethics of gender-ascription. I argued that feminist theorists' epistemic and practical goals must be addressed together, and that this is best achieved by a radically pragmatist approach on which the

truth of a gender-ascription is bound up in its normative warrant. A gender-ascription is an interpellation; its output is a normative status; its claim to truth is captured by that fact that it is challengeable, i.e. that it carries with it a commitment to showing that the one to whom the ascription is applied genuinely meets the concept's application-conditions. I didn't say anything about what considerations can be brought to bear in determining whether someone *is* genuinely entitled to a given normative status. That's the question this chapter attempts to answer, but it's a question that can be asked at two levels.

First, how can we adjudicate between competing claims about what entitles the claiming of a given status? Once we've arrived at an understanding of gender as normative status, how do we decide what the application-conditions of our gender concepts are and should be? Chapter two ended on this note, with disputes about gender interpreted as disputes about the application-conditions for concepts whose application-consequences were held fixed.

But there's a second question about which concepts, or which normative positions, there should be. This is pressing because, as we'll see, disputes about the application-conditions for gender concepts inevitably bring in issues about what gendered normative statuses there are or should be.

Roughly, my answers to the two questions are as follows. First, *given* a shared understanding of which gendered normative statuses are to be available, considerations of autonomy point strongly toward a permissive attitude on which individuals may choose which gender category they're slotted into. Second, the question of which identities should be available at all can be answered only through complex negotiative processes. Core to these processes is an ideal of mutual accountability. Perhaps little can be said now about where these processes will, or ideally would, take us. But, once I've developed my picture of how this negotiative process

works, I'll be in position to make two claims. First, that hegemonic masculinity is *illegitimate*. Second, that queer identities can have positive value as critical responses to hegemonic masculinity.

The picture I offer here is one on which the reasons we have for recognizing an individual's identity are grounded in that individual's authority to make claims on us. Identifying is a kind of claim-making; whether a given token identifying is *true* is a matter of whether the claims it makes are legitimate — i.e. authoritative — ones. If someone makes a claim on me and I'm unsure whether they have the authority to make it, the situation I find myself in is very different from one in which I'm trying to decide which objectively-specified state of affairs is best. Rather than thinking exclusively in terms of *utility*, I must think in terms of the agent-relative normative relations I bear to my interlocutor. This section is devoted to spelling these thoughts out.

3.2 Cox and Williamson

In May 2014, the *National Review* published a column by writer Kevin Williamson entitled "Laverne Cox Is Not A Woman."[90] The piece — a response to Cox having been featured on the cover of *Time* some days earlier — was a kind of sequel to a 2013 piece on Chelsea Manning, the title of which differed from that of the 2014 publication only in that Williamson did not bother to use Manning's chosen name. One piece begins with the subheading "Facts are not subject to our feelings;" the other, "Pronouns and delusions do not trump biology." The overarching theme of both is that one's status as a woman or man is a plain fact of one's biology,

[90] https://www.nationalreview.com/2014/05/laverne-cox-not-woman/

and that recent moves toward legitimizing transgender identities represent an Orwellian attack on truth and reason.

While Williamson's positioning of his viewpoint as obvious fact and his opponents' as mere fancy exemplifies a familiar strategy of naturalizing existing social arrangements, he does at one point show some awareness that the issue is importantly political. Contrasting trans identities with homosexuality — the latter of which he considers to be essentially private — he says that "The mass delusion that we are inculcating on the question of transgendered people is a different sort of matter, to the extent that it would impose on society at large an obligation … to treat delusion as fact…" He also shows awareness that the matter is not a straightforward disagreement over the facts, but a war of conceptual regimes: "Every battle in the war on reality," he says, "begins with the opening of a new linguistic front."[91]

The sense that the dispute is in some way verbal or conceptual, as well as awareness of its political dimensions, is shared by those on the other side. Janet Mock begins Part One of her suggestively titled memoir, *Redefining Realness*, with a quote from James Baldwin: "One cannot allow oneself … to live according to the world's definitions: one must find a way, perpetually, to be stronger and better than that."[92] Chapter two could be read as an attempt to cash out the notion of living according to a definition. There the idea was that the meanings of a gender-ascription has two aspects: its application-conditions and the consequences of its application. To live according to a particular definition of gender, then, is to be embedded in a discursive community with gendered normative statuses and conditions under which one is positioned as having those statuses. Baldwin's admonition not to live according to the world's definitions, then, translates

[91] https://www.nationalreview.com/2013/08/bradley-manning-not-woman-kevin-d-williamson/
[92] Mock (2014), 12

into a call for resisting dominant discursive practices, rejecting either the application-conditions of our gender-concepts or even the gender-concepts themselves.

But how, exactly, are we to be "stronger and better?" What's wrong with the world's definitions, and what should we be aiming for when we resist them? That's the project of this chapter: to lay out concepts, understood on this model as social practices, that? can be evaluated from an ethical or political perspective — and how such evaluations could effect change.

This chapter brings together the previous two in a couple of ways. First, chapter two's pragmatist picture of gender-ascription (and of gender-concepts or definitions as generalized ascriptions) is assumed as background. Second, the process by which our gender-concepts can and should change will mirror the process of rationally warranted conceptual change laid out in chapter one. The problems with conceptual engineering identified in chapter one will emerge here, though in a slightly different guise. The process of resolving disputes about which identities should be available (i.e. should exist at all) can be understood as an interpersonal analogue of the process of rationally warranted conceptual change outlined in chapter one1. Social identity is a matter of how one relates to others; carving out an identity is a process of negotiation; this process is rational insofar as it resolves the contradictions in our relationships by articulating and responding to reasons that we can issue to one another by the making of claims.

3.3 What's wrong with misgendering?

In chapter two, I argued that (a certain interpretation of) Haslanger's ameliorative approach could not get all the results that feminist theorists want. In particular, I argued that it failed to adequately capture what's wrong with transphobic claims like those made by Williamson. By failing to secure the result that Williamson speaks falsely, it mystifies why we

need to address Williamson's claim at all — or at least the sense in which Williamson and we actually disagree.

I did not, however, discuss how the problems with conceptual engineering articulated in chapter one carry over to the normative evaluation of gender concepts. Since this chapter is about how my pragmatist approach does a better job here, it will help to spell out how the problems from chapter one arise in this context.

Recall the discussion in chapter one of the concept of terrorism. I argued that the question whether a given activity ought to be categorized as terrorism reduces to the question whether it *is* terrorism — that attempts to 'engineer' the concept by appealing to the instrumental value of various conceptual schemes were doomed to pointlessness, perversity, or downright impossibility. But the lesson was not that we can't have rationally warranted changes in our concept of terrorism. It was rather that such changes will involve efforts to make sense of the world around us, its descriptive and normative unities and disunities, etc.

I think the problems I identified with engineering the concept of terrorism actually arise even more sharply for the engineering of gender concepts. An instrumentalist approach to the dispute between Cox and Williamson will ask us to consider the downstream effects of their competing conceptual schemes. Feminist instrumentalists like Katherine? Jenkins will say that Williamson's concepts are objectionable because they cause harm, particularly to trans people. The inadequacy of this approach is brought into sharp relief by this example example from Talia Mae Bettcher:

> *Consider someone who lives as a woman, sees herself as a woman, and has been sustained in a subculture that respects her intimacy boundaries, only to find that she is subject to violence because she is "really male." She goes through mainstream institutions (hospitals, jails) where she is housed as male, searched as male, and turned away from shelter as male. This invalidation is not only of an individual's self-identity*

but also of an entire life that has been lived with dignity in a competing cultural world.[93]

The instrumentalist take on this case would be that classifying a trans woman as "really male" results in her being "housed as male … [and] turned away from shelter as male" which are serious harms. Hence, we ought not classify her as "really male."

There are two problems with this line of reasoning. First, why should such important choices about how to treat a person hinge on a classificatory choice? After all, if the way we treat someone vis-a-vis housing is causally downstream of the way we classify them, then it seems we ought at least in principle to be able to correct the housing-related harms without changing how the person is classified. And if the reasons for which we ought to treat someone a certain way are intelligible independently of their gender category membership — as they must be if the instrumental rationale is to gain a foothold — then it seems we ought to be able to appreciate and act on those reasons without settling the question of who is a woman. If trans women are more likely to face violence in men's prisons, then this would give us a reason not to place them in men's prisons, full stop.[94] This point will generalize: any consideration we might bring to bear that doesn't reference gender category membership seems *ipso facto* to be one that can't decide the question what gender category someone belongs (or ought to be counted as belonging) to.

Second, there is a sense in which the instrumental rationale gets things backward. Whether a given form of treatment counts as a harm will sometimes depend upon what gender category someone rightly belongs to. Categorizing a trans woman as "really male" harms her precisely because she *isn't* "really male," and our treating her as such encodes a contemptuous refusal to acknowledge her for who she is, or (to say the same thing, on my view)

[93] Bettcher (2013), 242

[94] I would suggest that we have very good reasons to avoid putting nearly anyone in prison, but that's neither here nor there.

who she has right to claim to be. Bettcher's talk of "invalidation" tracks this point. Moreover, while the causally downstream harms (e.g., physical violence) are distinct from the act of categorization itself, it's natural to view these consequences as flowing from the more fundamental fact of contempt whose basic expression is the category-choice. The material violence visited on trans people is intelligible as a socially meaningful and ideologically motivated act only by reference to the act of categorization.[95]

These problems exactly mirror those we encountered with the concept of terrorism: instrumental reasons are reasons of the wrong kind for conceptual change in both cases. As before, I argue that the correct lesson isn't that we shouldn't change our concepts, but rather that conceptual changes are warranted by non-instrumental rationality. In all the examples from chapter one — terrorism, number, and chastity — the thought was that we encounter problems as we try to make sense of the world around us. Our concepts lead us into contradictions; they make bad predictions; they make arbitrary distinctions, or they conflate things inappropriately. The process of changing our thought to solve these problems is not one of selecting from a menu of available conceptual schemes; it's one of creatively transforming our viewpoint. After the transformation, our old concepts will in some sense seem incoherent to us — yet we'll also be able to see our new way of thinking as providing us with a clearer understanding of what our old one grasped only dimly.

The case of gender is importantly different from those in chapter one. While those had to do with objects, or activities, or character traits, this case is about *identity*. Concepts connected to social identity have a special normative structure: they encode pictures of what *matters about* a person, what can rightly be expected of them, and hence what demands can be made by and of

[95] This same point is made in Barnes (2017)

them. Of special importance here is the act of *identifying* — of categorizing oneself or another as a member of a social identity group. Because identifyings — e.g., gender-ascriptions — have the interpellative function described in chapter two, they implicitly constitute demands. Identifying oneself as a member of a group is a way of laying claim to a position in social-normative space; ascribing group membership to another is a way of recognizing them as having the relevant normative status.

3.4 Identity & Autonomy

It is common to see disputes like the one between Cox and Williamson construed as being about *identity*. And, among those on Cox' side of the dispute, it is equally common to diagnose the wrongness of Williamson's position in terms of its invalidating or disrespecting Cox' identity. But what is identity in the first place, and why does it demand our respect?

Probably the most consistent and important thread running through folk discussions of gender identity is that it is a highly personal thing. The Human Rights Campaign characterizes it as an "innermost concept of self;"[96] other characterizations include "our internal experience and naming of our gender,"[97] an "internal sense,"[98] and "how you feel inside."[99] This commitment to the internality of gender identity is intertwined with an ethical and political commitment to recognizing individuals' authority over their gender identities. The importance of this authority is sometimes cashed out in terms of self-determination or, as I'll say, autonomy.

[96] https://www.hrc.org/resources/sexual-orientation-and-gender-identity-terminology-and-definitions
[97] https://www.genderspectrum.org/quick-links/understanding-gender/
[98] http://www.transstudent.org/definitions/
[99] https://www.plannedparenthood.org/learn/sexual-orientation-gender/gender-gender-identity.

What tends to emerge here is a kind of liberal approach on which gender identity is a private matter over which only the individual concerned ought to have any say. Cox, for example, had this to say in a 2014 interview:

> *I think transwomen, and transpeople in general, show everyone that you can define what it means to be a man or woman on your own terms. A lot of what feminism is about is moving outside of roles and moving outside of expectations of who and what you're supposed to be to live a more authentic life.*[100]

Cox' talk of "moving outside of roles," taken at face value, seems to voice an ideal that I'll call *Identity Individualism*.[101] The core commitment of Identity Individualism, as I understand it, is to the individual's absolute sovereignty over their own gender identity, or to their not needing to justify their identity to anyone else. One obvious upshot of this is that anyone can rightly lay claim to any gender identity and no one else will have the authority to question them. But, as Cox recognizes, the claim that anyone can claim any identity implies the further, less obvious claim that anyone ought to be able to *interpret* their own identity in whatever way they please. In other words, it's not just that anyone can be a woman; it's that anyone can decide — for herself and without having to justify it to anyone — what it means for her to be a woman.

I argue that Identity Individualism is untenable. The lifeblood of the dispute over gender is precisely the *publicity* of gender identity. Gender concepts could not have the significance for us that they do if there weren't some shared understanding of their meanings. And even if we *could* develop gender identities in the unaccountable way that Identity Individualism prescribes, it would follow that there's no real need for Cox and Williamson to argue with each other. If

[100] https://www.damemagazine.com/2014/06/01/laverne-cox-i-absolutely-consider-myself-feminist/
[101] This is not to say that Cox herself necessarily endorses Identity Individualism.

both are permitted to define for themselves what being a woman or man amounts to, there's no way for a real dispute to get off the ground.

Yet identity does seem to be deeply personal, even once we've acknowledged its social dimension. So we're left with a tension between two seemingly essential facts about identity. On the one hand, it feels about as personal as anything ever could, and this seems to speak in favor of respecting people's identities. On the other hand, that very respect can only be understood in interpersonal terms, as a matter of viewing and treating people in the ways they wish to be viewed and treated.

3.5 Identity as Law & Autonomy as Self-Legislation

In "Trans Identities and First-Person Authority," Bettcher defines a notion of "existential self-identity" as "an answer to the question "Who am I?" where this question is taken in a deep sense … What am I about? What moves me? What do I stand for? What do I care about the most?"[102] To adopt an existential self-identity is simultaneously to view oneself as a certain kind of person and to take up the commitments that come along with being that kind of person. Related to this is the notion of *avowal* — a kind of first-personal report of one's inner life that isn't merely factual, but rather involves *taking responsibility* for what one says.[103] Bettcher observes that avowal, as a way to take on an existential self-identity, is "obviously connected to issues of autonomy" and to the possibility of one's autonomy being disrespected or even curtailed.[104] In disregarding someone's avowal, where that avowal is connected to their identity, I treat myself as having a kind of dominion over that person. *I* say who they are and what matters to, for, and about them.

[102] Bettcher (2009), 110
[103] Bettcher (2009), 101
[104] Bettcher (2009), 103

All of this is strikingly like the account of practical identity developed by Christine Korsgaard in *The Sources of Normativity*. For Korsgaard, a practical identity is "a description under which you value yourself, a description under which you find your life to be worth living and your actions to be worth taking … Your reasons express your identity, your nature; your obligations spring from what that identity forbids."[105] For both Bettcher and Korsgaard, an identity is a kind of self-understanding that includes not only descriptive beliefs about oneself, but normative convictions in the form of standards for evaluating oneself. For Korsgaard in particular, one identifies as a certain *kind* of person, and views oneself as subject to norms in virtue of the kind of person one is.

Korsgaard would also agree that identity is importantly connected to autonomy, though she would understand this connection in a more Kantian way. A distinctive feature of Kantian views is their characterization of autonomy as *self-legislation*. There are two sides to this idea. First is that autonomy consists in acting under a norm that is, in some sense, one's own. Indeed, arguably the core commitment of Kantian constructivism in metaethics is that normativity must 'come from within' — that no norm can be genuinely binding for an agent unless the agent herself acknowledges it as binding in a first-personal way. Korsgaard's account of practical identity is one way of spelling out this old Kantian idea of normativity as coming from the agent's first-person point of view.

Second, though, Kantian autonomy is not a matter of acting arbitrarily or unaccountably. As the language of 'legislation' suggests, autonomy will consist not in an absence of normative bonds, but rather in being bound by norms that are self-imposed.[106] Of course, one of the central challenges for a Kantian conception of autonomy is to say something about this notion of self-

[105] Korsgaard (1996), 101
[106] As Brandom (1979) puts it, the Kantian understands freedom as constraint by norms.

imposition, or a norm's being 'one's own,' that doesn't collapse into wantonly acting in whatever way one pleases.

There is a parallel between the Kantian dilemma of autonomy as self-legislation and the paradoxes faced by Identity Individualism and Identity Voluntarism. Just as ownership over one's social identity is significant only against a background of social meanings which are necessarily both objects and products of public negotiation, freedom in general is — on the Kantian view — only intelligible as constraint by norms that one can see as enjoying an appropriate sort of objectivity or non-arbitrariness. I argue that this is in fact more than a parallel: the social meanings in which identity becomes possible *are* the norms constraint by which constitutes — or *may* constitute — autonomy. To get a clearer picture of how this might be so, we'll want to say something more about the two key ways in which we relate to norms: being bound by them and legislating them for ourselves.

3.6 Acting Under Norms

What is it to be 'bound by' or to 'act under' a norm? Given my stated affinities with Korsgaard, it's natural first to consider the idea of *reflective endorsement*. Reflective endorsement of a norm occurs when I ask whether I really have reason, all things considered, to act in the way a norm prescribes; if I answer *yes* to this question, then in so answering I reflectively endorse the norm. So, we might say, acting under a norm is acting from reflective endorsement of that norm — it's thinking of that norm as binding on one. But this is problematic for at least two reasons. First, we can (and arguably must) act under norms without ever reflecting on them. Second, insofar as we're interested in the imposition of norms as a threat to autonomy, we want to allow that it's possible to act under norms that *aren't* one's own.

On the first point, it's helpful to draw on the remarks in Railton (2006) and Brandom (1994). Railton points out that most of our norm-guided behavior manages to be so despite our lack of any conscious representation of the norm itself. And this seems perfectly in order for the most part; if a norm's bindingness required that we actually reflect on it, we'd be bound by hardly any norms at all. A similar point is especially theoretically pressing for Brandom, who views recognition of norms as binding on one as a precondition for having explicit, propositional beliefs in the first place.[107] Brandom's account here is less naturalistic than Railton's, but the point is largely the same: we can treat something as a reason in practice, and in so doing come to implicitly act under a norm, without ever making this implicit commitment explicit in speech or propositional belief.

As I read them, both Railton and Brandom see normative guidance as having two essential ingredients: (1) the agent or agents *respond to* the norm by trying to conform to it; (2) the agent or agents regard failure to conform to the norm as warranting sanctions. Now it might seem like (2) here is just a sneaky way of saying the agent must (at least implicitly) endorse the norm, in which case we again run into trouble in thinking about how we can make someone act under a norm that they don't endorse. But I think this appearance can be explained away if we notice that normative guidance happens at both the individual and social levels.

At the social level, it's more natural to talk about *norm-structured activity* rather than norm-guided action. Norm-structured activities adhere to (1) and (2) in a way that doesn't require any particular attitude to be adopted by each and every participant in the activity. Whereas individual agents respond to norms, a norm-structured activity *realizes* norms as behavioral regularities. Of course, this regularity will tend to be far from complete — norm-

[107] Cf. Brandom (1994), 20

violations are common — and this is where (2) comes in. Sanctions play a key role in constituting a cluster of behaviors as a norm-structured activity, and in determining what norms are structuring it. The first point is illustrated by behavioral regularities that aren't produced by sanctions: people regularly recoil when they hear the sound of nails on a chalkboard, but the minority who do not are not then sanctioned for having violated a norm. So there is no norm dictating recoiling and the activity of recoiling is not norm-structured. The second point is illustrated by prudential decisions made in the context of games. It is a good idea to kick the extra point rather than to go for two when the former is sufficient to put one in the lead late in the game. But going for two isn't penalized, so it's not among the rules of football that one must kick in these situations.

Once we have a norm-structured activity in place, we can understand individuals embedded within these activities as participating in them, where this participation need not imply endorsement. A gym teacher might exercise their authority to make children play a game of kickball. In so doing, the teacher compels the students to participate in a norm-structured activity. Insofar as the students do this, they act under the norms of kickball, taking on the normative statuses determined by the rules of the game. Students who violate those rules will be sanctioned by the other participants in the game.[108] All this is compatible with none of the students wanting to play, and even with all of the students thinking that kickball is boring and not worth playing. Notably, in such a case, even the students who wish not to play will have to participate in the sanctioning and therefore *treat as sanctionable* some behaviors that they might,

[108] That is, they will be sanctioned under the rules of kickball. They might also be sanctioned by the teacher for refusing to participate in the assigned activity, but these latter sanctions are connected to the educational practices in which the teacher has their authority, rather than to the practice of kickball.

in their heart of hearts, view as permissible or even as something they'd like to do themselves if only they could get away with it.

Gendered practices are importantly different from games vis-a-vis their ubiquity and mandatoriness. With games, one can simply refuse to play. One cannot opt out of gender. This is why gender — and other sorts of persistent social identities — present an especially crucial site for both the exercise and the curtailment of autonomy. One might wish to act under a different set of norms than those available in the communities to which one has access. In such cases, continuing to live as a social being will mean participating in norm-structured activities the constitutive norms of which one would like to reject. This in turn means not only having to try to follow the rejected norms lest one be sanctioned; it also involves sanctioning others for violating those norms. In this way, hegemonic social practices can produce a kind of fragmented agency: one is not merely a passive recipient of abuse, nor is one merely 'pretending' to act under a norm in order to avoid punishment. Rather, the norms are treated as binding in practice even while they're repudiated in the mind.

3.7 A Norm of One's Own

Now we must ask what it is for a norm to be "one's own." The end of the last section made it sound as though any consciously repudiated norm is *ipso facto* not one's own. Were we to accept this claim, we might also be tempted to accept its converse: any norm that one consciously endorses is one's own. I want to reject this picture from the start. It is a central commitment of Kantian accounts of normativity that, while normativity does in a sense come from within, it does not follow that we may value whatever we please. Indeed, Kant himself insisted that acting purely on one's inclinations is incompatible with autonomy. While we may not agree with Kant on the details here, I do want to retain the crucial point that autonomous

action as such is always answerable to something other than the momentary inclinations of the individual agent.

A typical way of getting at this is to posit some justificatory procedure. Korsgaardian reflective endorsement is one option here, as is the Rawlsian method of reflective equilibrium. But one might well ask — and many have asked — why the fact that a norm survives such a justificatory procedure should matter. As was emphasized before, we for the most part do not in fact reflect on the norms under which we act. So, unless we constantly act under norms that have no hold on us, reflection cannot be necessary for normativity. But if we account for this sort of thing by making the justificatory procedure merely hypothetical, we lose the sense that it's tracking what we care about. Why should I care what some 'idealized' (whose ideal?) counterfactual version of me would think? So, the worry presses as before: normativity seems to either come from outside the agent or to be just whatever the agent pleases in the moment.

Here I think the answer is to appeal to a notion of *interpersonal* justifiability. The justificatory procedure whereby a norm becomes one's own is the process of justifying the norm, and the conduct it recommends, to those of whom the norm makes or might make demands. At first blush, this might sound like a way of abandoning the project of locating normativity within the first-person viewpoint of the agent. But, for reasons I'll try to make clear, I don't think this is true.

3.8 Disputes Over (Only) Application-Conditions

It might be thought that my way of talking about concepts provides us an easy way of splitting the difference between Identity Individualism and its opposite (Identity Totalitarianism?). If the problem with a thoroughgoing Identity Individualism is its failure to recognize the importance of shared meanings, then we'll need to make some room for shared

meanings. And if the thing that's right about Identity Individualism is that it seems people should have considerable authority over who they are, then we'll need to work out a way for people to choose their own positions within the matrix of shared meanings. Perhaps we could achieve this within the two-sided view of concepts by suggesting the following: we have a fixed suite of normative standings constituted by the normative outputs of our concepts, but the *inputs* — the concepts' application conditions — are relaxed. Perhaps gender could be like membership in a political party: while one's gender or party membership influences what one may do (what restrooms one may enter or what primaries one may vote in) one has total, unaccountable authority over which group one joins. Call this view *Identity Voluntarism*.

There's a lot to like about Identity Voluntarism. It permits us to interpret the dispute between Cox and Williamson as a substantive dispute about the application-conditions of a normative concept, the application-consequences of which are an object of shared understanding. In fact, I think that to some extent this is exactly what's going on. While it's easy to be impressed by how different the gender ideologies of someone like Cox are from those of someone like Williamson, it's also easy to overlook the considerable overlap which, as we've seen, is what permits the dispute in the first place. While there's a lot of background theory that feminists and essentialists disagree on, they all understand that who counts as a woman will have immediate consequences for who will have access to certain physical and social spaces such restrooms, sports teams, and social services. So perhaps the dispute just comes down to this: Williamson thinks access to these spaces should be determined by reproductive biology, while Cox thinks it should be based on voluntary choice.[109]

[109] Cox needn't think that this voluntary principle is bedrock, and in fact there's evidence that she doesn't. In an interview, she says that "I am a woman and I deal with the realities of being a woman." There the idea is plausibly

If Identity Voluntarism is couched in terms of the deontic scorekeeping picture of meaning, then we can solve the problems faced by the instrumental account of amelioration. First, this captures the sense in which disputes over the legitimacy of trans identities are about who trans people *really are* (I covered this in the second chapter). Second, it reveals that misgendering involves a treatment of someone's claims as irrelevant. Misgendering thus expresses disrespect for the person directly, and this is (at least *prima facie*) intrinsically objectionable. Third, it makes the harms done by misgendering intelligible as downstream consequences or practical expressions of the contempt principally manifested by the misgendering itself. Here the reasons we're furnished with are not instrumental ones having to do with what the consequences of recognizing trans identities would be; they're deontic ones having to do with trans people's authority to demand our recognition.

But the paradoxes of autonomy still loom here. Identity Voluntarism, couched in terms of the deontic scorekeeping picture of gendered discursive practice, does make clear the way in which gender identity is both personal and social and it gives us a powerful way of connecting this up with hallowed traditions in ethical theory. But it also gives us a very sharp picture of why individualism can't work: no network of normative statuses can possess all possible statuses, so my adoption of a status — of an identity — constitutes a demand not only that I be treated in a certain way, but also that many other ways of being be foreclosed for those who share a social world with me. The next section will try to make this point clearer by connecting it both to the inferentialist tradition I'm drawing from and to some examples.

---

that the substantive criterion for being a woman is having certain experiences. This would still make for a relevant difference with Williamson.

3.9 Harmony & Incompatible Identities

The paradox of autonomy we've been circling around can be given clear expression within the deontic scorekeeping picture of meaning I'm proposing. On this picture, we say that claiming an identity implies a demand that others recognize one as possessing the relevant normative status. If autonomy consists in being able to live under norms that can be regarded as one's own, then autonomy for all cannot consist in everyone being able to choose whichever norms they please in a wholly unaccountable way. This is because the recognition one demands in claiming an identity directly requires that others be placed in complementary positions. This not only means that others have to act in a certain way; it means that other identities that others may want to claim may become unavailable to them. To see how this is so, we'll talk about Dummett and Brandom's appropriation of his views on what he calls *harmony*.

On Dummett's view, grasping a concept requires understanding both its application-conditions and the consequences of its application. However, while Dummett treats these two aspects of a concept — its inputs and outputs — as distinct, he also stresses that they cannot vary totally independently of one another. While Dummett criticizes accounts considering *only* application-conditions as naive, he dismisses as "almost equally naive" any view according to which "we have the right to attach whatever evaluative meaning we choose to a form of statement irrespective of its descriptive meaning."[110] There must, Dummett argues, be some sort of "harmony" between the two aspects of the concept.

These remarks of Dummett's were aimed primarily at logical vocabulary; the disharmony he had in mind is the kind exemplified by the 'tonk' operator which, by combining the introduction rules of disjunction with the elimination rules of conjunction, permits anything to be

[110] Dummett (1981), 455

inferred from anything. But the idea was also extended to non-logical concepts. Indeed, Dummett himself extended it to slurs, claiming that slurs are disharmonious in much the same way as 'tonk.' While I don't agree with Dummett's analysis of what disharmony amounts to in these cases, I nonetheless take his view as inspiration and here discuss two ways in which I believe concepts can be disharmonious.[111]

Suppose we adopt Dummett's "almost equally naive" view according to which any set of application-conditions can be matched with any normative role. On this view, we ought to be able to introduce a concept — call it *porture* — whose application-conditions are those of *purple* and whose practical consequences are those of *torture*. The concept itself seems to encode a category mistake. This concept is disharmonious in that the application-conditions and the normative role are mismatched in such a way that the concept literally cannot be realized in a discursive practice. How does one treat a violet as a war crime?[112]

The second sort of disharmony is less dramatic than that illustrated by *porture*. Suppose we tried to flip the inputs and outputs of *it* and *not-it* in the game of tag. Then the rules would be: all players but one start out as chasers; when a chaser tags the runner, the tagger then becomes the runner while the previous runner joins the chasers. What results is a perfectly playable,

[111] Dummett diagnosed 'tonk' as disharmonious because it involved a non-conservative extension of the language. The idea is that, while introducing new logical vocabulary is okay, it's okay only if the inferences that vocabulary licenses were already available in the relevant language. Extending this analysis to slurs, Dummett thinks the problem with *boche* is that it licenses an inference from someone's being German to their being brutish. I agree with Brandom that this analysis cannot work because some non-conservative changes are in fact called for and because *which* changes count as conservative is language-relative in a way that generates embarrassing results when thinking of slurs and other objectionable thick terms.

[112] This is different from Brandom's account of disharmony, which he developed as a replacement for Dummett's. For Brandom, the deployment of a concept involves an implicit commitment to an inference from the concept's application-conditions to its consequences. Such an inference can be made propositionally explicit by a conditional whose antecedent states the application-conditions and whose consequent states the consequences. On Brandom's view, the problem with *boche* isn't that it's non-conservative, but rather that it licenses a materially bad inference. The incompatibility manifested by *porture* is stronger than this; an inference from something's being purple to its being a crime isn't just materially bad — it's unintelligible.

though unfair, game. But the flipping of conditions in this case results in such fundamental changes to how the game is played that it's hard to see it as the same game at all. And, since *it* and *not-it* were defined by their place in the game of tag, it follows that the runner and chaser roles in this new game aren't the same as it and not-it. So, in attempting to change the application-conditions for *it*, we ended up transforming the game to such an extent that our purposes were defeated.

Both sorts of disharmony show that there are limits to what application-conditions can be paired with what normative roles. This is a problem for a view which, like Identity Voluntarism, hopes to find room for individual autonomy within a space of shared meanings by loosening application-conditions. We might find that the loosening that Identity Voluntarism calls for would result in a breakdown, or at least a significant transformation, in the space of shared meanings.

Gender concepts are a case in point. As radical feminists like Andrea Dworkin and Catharine MacKinnon have pointed out, popular understandings of gender are deeply interwoven with popular understandings of (hetero)sexuality, which in turn place great emphasis on bodily difference.[113] This bodily difference is made especially salient in penile-vaginal intercourse, which is symbolically understood as domination. Hence most people's understandings of themselves as gendered are inextricable from their understanding of their bodies as sexed, and of the kind of sexualities for which their bodies are suited. This network of meanings simply could not survive without the reference-point that it finds in bodies. From the point of view of someone immersed in this network, to say that anyone can be a woman or man is the same as saying that no one can be.

[113] Cf. Dworkin (1987) and MacKinnon (1987) and (1989).

I want to talk here about the trope on display in the show *It's Always Sunny In Philadelphia* in the relationship between Mac and Carmen. When the two meet, they're immediately attracted to one another. After a brief conversation, it's revealed to Mac that Carmen is trans. From this moment, the comedy centers on the relationship between two things: Mac's homophobia and Carmen's penis. They end up getting together, and in one scene Mac says to Carmen:

> *"Here's the deal. I feel like we make out, and it's great — I mean* really *great — but then things start to get hot and heavy … and sometimes — just sometimes — I bump up against it and I … I just can't handle that*."[114]

The show here participates in a long-running comedy trope centering on heterosexual men's discomfort with being attracted to trans women. This point is usually put in terms of the men's interpretation of their desires as homosexual. This is, of course, a consequence of their categorizing trans women as men. (Mac finds out that Carmen is trans when his friend Dennis says, "That's a dude."[115]) So, you might think the tension could just be resolved if Mac went ahead and regarded Carmen as a woman. So, in that sense, maybe Mac's identity as a straight man and Carmen's identity as a woman are compatible.

But Mac's conception of manhood doesn't allow for this. Carmen's conventional feminine attractiveness may allow Mac to momentarily forget that she's someone he'd normally categorize as a man, but that only lasts until he bumps up against her penis. This is what stops Mac from fully regarding Carmen as a woman, because Mac's understanding of manhood is so centrally anchored in that organ. And this isn't an association he could break without substantially changing his own self-conception.

---

[114] "Mac Is a Serial Killer" 4:24
[115] "Charlie Has Cancer" 5:51

I think careful consideration of dominant gender norms and dominant understandings of gender groups' membership conditions will reveal a non-detachability of the kind Dummett refers to here. These norms aim to dictate not just what we wear or how we address one another, but nearly every aspect of our lives including (perhaps most fundamentally) sex and reproduction.

Consider media representations of heterosexual encounters: men routinely pick women up, press them forcibly against walls, etc. People's experience of heterosexuality is governed by the belief and expectation that men are larger and stronger than women. More pertinently to the case of Carmen and Mac, people's understanding of heterosexuality places penile-vaginal intercourse at its very center; arguably this act plays a central role in defining for people what men and women are and how they're to relate to one another. For the act of 'penetrating' has a certain metaphorical significance as violence and domination. Most men don't want to be 'penetrated,' even by a woman, since this would be to take on a subordinated and hence unmanly role. Thus, though Mac is sometimes able to see and treat Carmen as a woman — he's able to activate certain associations and deploy certain kinds of learned scripts — he can do this only as long as he ignores her penis. This is because that whole cluster of concepts, beliefs, associations, and scripts is based on the idea that women are naturally different from men in virtue, principally, of their genitals.

Given what I've said so far, I hope it's clear why someone like Mac would feel *discomfort* about someone like Carmen. But this doesn't exactly prove my point, which is that Mac (and most other cis straight men) would have to revise their own identities if they were to recognize those of trans women. You might think Mac could simply not associate with Carmen.

One problem with this suggestion is just that it isn't fully tenable in the context of sharing a social world governed by a mostly shared set of social meanings. As Bettcher points out, one central social function of gendered clothing norms is to impose on everyone a norm whereby they're expected to signal their genital status to those around them. Assuming this is what such norms are for, they work only to the extent that gendered clothing functions as at least a reasonably reliable indicator of genital status. So "live and let live" doesn't exactly work here.

Another, perhaps deeper, problem is that acknowledging Carmen as a woman would undermine Mac's own claim to manhood. Most people, I think, view their gender as *entailed by* or *readable from* their bodies. Unlike e.g., Bettcher or Korsgaard, they do not view gender as a matter of *identity* in the sense that it's something they claim for themselves and must take responsibility for. It's just an undeniable fact. Shifting from an understanding of manhood and womanhood as bodily states to an understanding of gender as normative status — whether that status is taken up voluntarily, ascribed by others, etc. — destabilizes any identity predicated on the former sort of understanding.

What all this means is that disputes over application conditions can't be nearly separated from disputes about the normative upshots of application. So, the kind of pragmatist picture I'm advocating for doesn't work as a way of simplifying disputes or breaking them down into a typology. But it does help to illuminate the ways in which labels connect to social practices and why disputes over application conditions matter. Because these words aren't mere labels but rather signifiers of normative status, the lack of a shared understanding of both application conditions and normative upshots — or the lack of perceived harmony between these — presents an obstacle to life in a shared social world.

While this observation is, I think, fatal for Identity Voluntarism, I don't think of it as bad news. In fact, as the discussion of Dworkin and MacKinnon suggests, I think we in many cases will actually *want* to tear down the system of norms. But the challenge remains, as before, to show how we can recover notions of identity and autonomy from the rubble.

3.10 Recognition & The Relationality of Identity

I'm going to careen off into discussion of Hegel and Beauvoir here. The point of doing this is not to wholly endorse their views, but to show how they resolved roughly what I'm seeing as the Kantian paradox of autonomy and how Beauvoir applied it to the examples of gender identity and sexuality. This latter point is especially important because it's going to blend with what I go on to say about how our concepts and identities change through direct, recognitive encounters with other people.

3.10.1 Hegel & Beauvoir

Hegel thought that self-consciousness requires the awareness of another conscious being.[116] For him, the most primitive form of consciousness is one which looks directly at the world as a means of satisfying its desires.[117] Only when it finds itself treated as an object by another being does it learn to take up an outside perspective on itself and thereby become self-conscious. This process is painful, though, because the image of itself that it sees reflected in the eyes of the other is refracted by the other's own judgments. This means the other has a kind of power over it. The newly self-conscious being then tries to shape that reflection into its own preferred image of itself — which means exerting control over the other. This relationship is

[116] Cf. Hegel (1977), 111: "Self-consciousness exists in and for itself when, and by the fact that, it so exists for another; that is, it exists only in being acknowledged."
[117] See Section A of Hegel (1977).

reciprocal, with both beings struggling to dominate the other and force them to recognize them in the way they wish to be recognized.

The winner of this struggle becomes the Lord, and the loser becomes the Bondsman. The Lord forces the Bondsman to do his bidding, where this crucially involves practically recognizing the Lord as the supreme subjectivity, i.e., as possessing a kind of absolute mastery over the world of objects — a world that includes the Bondsman himself.

But this Lord/Bondsman relationship is dialectically unstable, because the recognition that the Lord extracts from the Bondsman is hollow. Recognition is always recognition *from* another agent — another autonomous being. By placing himself in this position of supremacy and deprecating the Bondsman as a mere instrument for his own fulfillment and aggrandizement, the Lord contradicts himself. So, the situation is unsatisfactory even from the Lord's own perspective; changing it requires acknowledging the Bondsman as a free and equal being.

How it's possible to move from this moment of hierarchy to a state of mutual recognition that's satisfactory to both parties is something of a mystery. Can each party regard both himself and the other as both subject and object at one and the same time? Or must they, as Sartre thought, eternally struggle to objectify one another?[118]

Beauvoir appropriated this parable for her own purposes in *The Second Sex*, positioning men as Lord/Subject and women as Bondsman/Object. While Beauvoir is often interpreted here as using Hegel's Lord/Bondsman dialectic as a shallow metaphor or (perhaps worse) simply rehashing Sartre's own appropriation of it, Nancy Bauer shows that Beauvoir in fact departs from both Sartre and Hegel in significant ways. She is more optimistic than Sartre but less so than Hegel. Unlike Hegel, she denies that this will ever take the form of a stable, final state of the

[118] For more on Sartre's views on this matter and how they contrast with Beauvoir's, see Chapters 4 and 5 of Bauer (2001).

kind Hegel identifies in *Geist*.[119] There will always be some form of struggle. But she departs from Sartre when it comes to the nature of this struggle. Instead of an eternal struggle for us to subordinate one another, Beauvoir envisions a struggle that we undertake *together*. It is an unending project of understanding ourselves and one another, where it's understood that this requires both of us to be vulnerable to the other's defining us. My only real adversary in this struggle is myself: "I struggle to let go of a fixed picture of myself, to risk letting the other teach me who I am."[120]

It's instructive that Beauvoir turns to (hetero)sexual relationships as a plausible site at which this kind of project can be undertaken in the oppressive world in which we live. This is because "the erotic experience is one of those that discloses to human beings in the most poignant way the ambiguity of their condition. In it they experience themselves as flesh and spirit, as the other and as subject."[121] It's instructive for at least two reasons.

First, it appears on the surface to contrast sharply with the Dworkin/Mackinnon view of the social meaning of heterosexuality. But I think the pictures can be made consistent. While boys and men often learn to express their gender identity by framing themselves as subject and women as objects,[122] another familiar dimension of sexual experience is the desire to be desired. This always, at least implicitly, means seeing oneself as an object for the enjoyment of another conscious being. But by desiring and even delighting in this sense of oneself as an object of desire, one of course exercises one's own subjectivity. So, in erotic love, we find an opportunity

[119] Bauer (2001), 185
[120] Bauer (2001), 236
[121] Beauvoir (2011), 416
[122] Cf. MacKinnon (1989), 124: "Man fucks woman; subject verb object."

to meet with another being on equal footing: we are both irretrievably ambiguous, subjects and objects at one and the same time.[123]

Second, and keeping both the pessimistic and optimistic cases (i.e., the MacKinnon and Beauvoir images of sexuality) in mind, it suggests that sexuality provides a model for how interpersonal relations can transform and be transformative for us. If Hegel's parable of the Lord and Bondsman outlines schematically why interpersonal relationships of domination are dialectically unstable and must, even from the viewpoint of the Lord, be transcended, then the transition of sexuality from a use of others as mere objects to a delighting in mutuality provides a case study in how that transcendence can occur. It occurs through concrete exchanges with particular people — through questions and disclosures, requests and invitations. In this process one can learn to take responsibility for one's desires by submitting them to others for scrutiny. To do this is to risk judgment and rejection. But it's a risk we must take if we're to see a reflection of ourselves in the eyes of others that we can at the same time recognize as authentically ours.

Think again of Mac and Carmen. The negotiations they undertake in their sexual encounters are implicitly negotiations of their gendered and sexual identities. What brings them together and motivates this negotiation is their desire for one another. But they cannot both satisfy those desires as they initially conceive of them, because Mac's desire for Carmen articulates with his identities in a way that's incompatible with accepting her wholly as she is. This initiates the struggle Beauvoir talked about. Their intimacy cannot be fully realized unless one or both of them changes who they are — psychologically, physically, or both. In fact, both changes happen over the course of the series. Carmen chooses to have bottom surgery sometime

[123] This is not to say that Beauvoir's picture of heterosexual relations is, on the whole, a rosy one. It isn't.

after she and Mac have stopped seeing one another. And Mac, after years of increasingly undeniable (but desperately denied) fascination with male bodies and with penises in particular, comes to identify as gay. At the level of sheer plot, then, the sexual relationship between Mac and Carmen does not function as a site of transformation in the way Beauvoir might have imagined — they undergo their respective transformations after they've stopped seeing each other, and after the transformations they're more obviously sexually incompatible than they were before. But thematically, *It's Always Sunny* positions Mac's relationship with Carmen as the inaugural step in the destabilizing of his identity as a straight man. His attraction to Carmen presented him with an unworkable situation; he resolved it by changing his self-understanding.

The point here isn't that sexual relationships are the only, or even a uniquely good, context in which to do this. But phenomenologically it provides for most people an already-experienced point of contact between one's own desires and genuine concern for the subjectivity and desires of another person.

3.10.2 Negotiation & Authority as Interpersonal Justifiability

In this section I'm going to try to spell out the ways in which identities are products of negotiation. The fundamental idea here is that claiming an identity is a social act undertaken by a particular, embodied, and socially embedded person. It implicates other persons both directly (as audiences for a given communicative act) and indirectly (as members of the discursive community for which an identifying aspires to be universally valid).

Issuing a claim always presupposes some kind of authority on the part of others to challenge it. If they couldn't challenge it, then they couldn't genuinely recognize it. This was the problem the Lord faced in trying to coercively extract recognition from the Bondsman. In other words, I have to be prepared to justify my claims to my interlocutors. This is as true of

identifying as it is of any other claim. This of course raises the possibility that I might *fail* to justify myself, in which case I'll be obliged to drop the claim — that is, to alter my identity. This alteration may involve a change in whether I think myself to fall within a concept's application conditions. But, for reasons we've already seen, it can also involve change in which concepts I view as legitimate. (For example, the considerations that lead to respect for the identities of trans women also require me to make changes in how many genders I think there are.) Thus empathizing with people, trying to understand their experiences from a perspective that takes them seriously as my equal, can change my perspective in ways that I couldn't have articulated previously. Dialogue produces deep conceptual change.

But how expansive is the authority we have to make claims of one another? What determines whether others are obliged to accept my claims? My answer is: To regard someone as my equal is to regard them as having the very same standing that I have. So the demands they can make of me are just the same ones I can make of them. A particular claim is authoritative only if it stands up to scrutiny from the point of view of a community of mutually accountable equals. The mutual accountability and commitment to interpersonal justification here outlined are inescapable insofar as one adopts a social identity, because a social identity is always an identity *as* a member of a group the other members of which one regards as having the authority to make demands of one.

It's clear how this framing resolves the paradox of autonomy, at least as regards social identity. Normativity doesn't come solely from within, but neither is it arbitrarily imposed from the outside. Rather, identities quite generally are other-regarding in such a way that the involvement of other people cannot be viewed as irrelevant, as merely incidental, or as an unwelcome constraint on one's freedom. If my identity demands that I bear certain normative

relations of reciprocal recognition to other people, then accountability to others is built into the activity of identifying from the start. Autonomy vis-a-vis one's social identity is thus inseparable from the authority of others to demand justification for one's identity. The first-personal perspective can only play the role it plays for Bettcher and Korsgaard — the role as the locus of identity and source of normativity — because it already has built into it a demand for recognition by others and a correlative commitment to justifying oneself to those others.

Implicit in my assertion that identity is subject to demands for justification is the claim that an identity may fail to be justified or justifiable. How might this be so? This is another matter on which I think reactionary rhetoric has shown a glint of awareness. In 2016, the University of Michigan implemented a system whereby students could write in a preferred pronoun which would then appear in class rosters available to instructors. Some students abused the system by entering terms like 'Lord' and 'Your Majesty.' There are many things to say about this practice, few of them good. Probably most of these students simply took themselves to be making a joke.

But I think there's something that at least some of these students were sensing, even if they couldn't quite articulate it. Why were terms signifying hierarchy — especially high-ranking positions in systems not presently in force in the United States — so popular? I suspect it has to do with students' sense that there's something wrong with Identity Individualism, and that this something is illustrated by their demand that they be so addressed. 'Words have meanings,' they might say, gesturing toward the fact that Identity Individualism enshrines identity and address as very important things while ignoring what makes them so important in the first place, namely their role in conferring and tracking normative status. In asking to be called 'Your Highness,' a student might aim to point out this lacuna by making those around him uncomfortable. That

discomfort is meant to undercut the anything-goes approach to identity. If anything goes, then the student ought to be able to demand to be addressed as a noble. The comparison with the Lord/Bondsman dialectic is plain.

We needn't turn toward especially politically charged cases to make this point, though. Social identities are negotiated all the time. Taking up an identity as someone's friend, for instance, involves commitments and entitlements with respect to that person. If I fall short of my commitments, my friend has the right to call me to account. If my failures are severe or persistent enough, they may decide that my commitment is not authentic — that I'm not a true friend. In other words, they can judge that I'm not entitled to claim that normative relation to them. In just the same way, someone might call my identity as an activist into question if they've never seen me at any of the relevant meetings or marches.

Children whose parents remarry often find themselves in awkward positions where their relationships with their parents' new spouses are unclear. The processes by which these relationships and attendant identities are negotiated are often messy and painful. Things are further complicated by the fact that there's no clearly defined way these relationships are *supposed* to look. To what extent does a stepparent take on the role of a *parent*? Different families can come to different understandings here, all of which they may find equally satisfactory.

The thing to take away from these examples is that the processes by which these relational identities are forged are ones of interpersonal justification. They involve calling on one another to enter into relations of reciprocal recognition. Resistance to that call can be legitimate, and demands to show that the call is warranted — whether by background facts or by the caller's

authentic forward-looking commitments — are often in order. We answer these demands in an effort to forge a shared understanding of ourselves and one another.

3.11 Tying Things Together

To sum up: Identities are normative statuses consisting of clusters of relationally defined commitments and entitlements. Concepts are functions on deontic scores whose successful deployment takes us from one set of normative statuses to another. Gender-concepts are concepts with an interpellative function — i.e., their successful application slots the individual to whom a gender is ascribed into a gendered normative status or identity. Whether a given deployment of a concept *is* successful — whether it's genuinely entitled — is an irreducibly normative question. While the answer to this question will in many cases be reasonably clear. But disputed cases can only be resolved by an interpersonal process of justification in which we issue further claims, adducing and producing further reasons from our shared standing as mutually accountable equals. Disputes over whether someone is entitled to claim an identity, then, can only be answered by complex dialectical processes that will implicate not only what application-conditions our concepts should have and what it takes to be entitled to a given identity, but also what identities should be on offer.

The remainder of this paper will be to use this theoretical framework to generate some substantive normative conclusions. In doing so, I'll start with what we might think of as an ideal-theoretic approach. I believe such an approach can capture what was right about Identity Voluntarism and its prizing of autonomy while acknowledging both that autonomy is necessarily socially embedded and that the social meanings in which we live are themselves legitimate objects of critique. This alone will be enough to secure the result — surprising to some and perhaps obvious to others — that many people, and probably most men, have illegitimate gender

identities. I'll then transition to talking about a non-ideal approach on which particular identities are understood and evaluated as dialectical responses to unjust social arrangements.

3.11.1 Ideal Justice and Nonideal Justification

How might we appeal to the idea of interpersonal justification to adjudicate between these rival sets of norms and identities? One way — perhaps the most natural given the Kantian affinities I've expressed — would be to appeal to a kind of conceptual Kingdom of Ends, or a discursive practice that could be an object of a stable agreement for a community of mutually accountable equals. Such an ideal could give us a powerful way to criticize objectionable concepts and their attendant identities.

A case in point is provided by what Raewyn Connell has called "hegemonic masculinity," a kind of masculine identity "centered on a single structural fact, the global dominance of men over women." Hegemonic masculinity is defined in relation both to subordinated alternative masculinities and to *emphasized femininity*, which is "defined around compliance with this subordination and is oriented to accommodating the interests and desires of men."[124] Central to hegemonic masculinity is its commitment to heterosexuality, where — as we saw earlier in the brief discussion of Dworkin and MacKinnon — heterosexuality itself largely amounts to the eroticization of inequality and of bodily difference as a symbol of inequality.

If we were to characterize hegemonic masculinity and emphasized femininity as normative statuses, they would plainly fall short of the ideal we're considering here.[125] The

[124] Connell (1987), 183

[125] My treatment here is arguably a *mis*treatment of Connell's theoretical framework. Throughout *Gender and Power*, Connell is at pains to emphasize the inadequacy of what she calls 'sex role theory,' i.e. theories according to which gendered practice is best understood in terms of unitary male and female roles, constituted by norms, that govern people's actual behavior. Connell's criticisms of sex role theory are numerous; I couldn't address them here even if I wanted to. For now, I'll say that I believe many of the qualities she attributes to sex role theory to be

significance these identities place on (reproductive) bodily difference means there's strong pressure to shuffle people into one or the other identity depending on facts of reproductive anatomy. This is an injustice in application-conditions because it makes life-chances dependent upon morally arbitrary anatomical facts over which we mostly lack any control. And, more fundamentally, these identities are intrinsically objectionable because the norms in which they consist instantiate relations of domination and subordination. To the extent that gendered identities and practices as we know them match Connell's descriptions of hegemonic masculinity and emphasized femininity, it will follow that gender identities as we know them are illegitimate.

While that last point shows that the framework I've developed here has some bite, it also illustrates (at least) two problems. First, the construal of gendered practices in terms of stable systems of norms is static and ahistorical; gendered practices as we know them today are products of an ongoing process in which conflict and contestation play central roles. Second, the account seems to condemn hegemonic masculinity and emphasized femininity as *equally* illegitimate. The latter problem is especially pressing because it points to a more general inability of the account so far to distinguish between the oppressor and the oppressed, dismissing identities adopted *in opposition to* injustice as themselves illegitimate because they (presumably) would not exist in an ideal world. Ultimately, though, I think both of these objections can be answered without making any fundamental changes to the account.

The first thing to emphasize is that, while I do want to talk about stable systems of norms, those norms can only be realized by concrete social practice — by people making, and

detachable from the basic idea of gendered practice being structured by the application of roles defined in terms of normative relations.

responding to, demands. A particular system of norms can only be said to govern a practice to the extent that the participants recognize those norms as binding on them, where that recognition amounts to treating certain demands as legitimate or authoritative. Demands can challenge and they can be challenged, and nothing I've said places any *a priori* constraint on what demands can be recognized, or on what basis.[126] Economic circumstances, historical antagonisms, and violent coercion can all influence what systems of norms are in force — and what identities are justifiable — in a particular context.

Contingent features of a social-historical context may also generate a non-ideal justification for certain identities that would not be possible under conditions of ideal justice. One clear example here would be identities defined specifically in opposition to injustice; a perfectly just world would not need activists or freedom fighters. Other identities, while perhaps not defined in opposition to the status quo, present less-objectionable alternatives to it. Some members of some queer subcultures see themselves as trying to create spaces that more closely match their ideals, while acknowledging that they can only do so with the resources a non-ideal world has given them. While these identities may only have their point against a background of injustice, they do not — like hegemonic masculinity — express contempt for those of whom they make their demands.

It's helpful here to think of Dembroff's idea of "critical gender kinds," or social groups whose members "collectively destabilize one or more elements of the dominant gender ideology."[127] I'm especially interested in what they call *existential destabilizing*, which "stems from or otherwise expresses individuals' felt or desired gender roles, embodiment, and/or

[126] To clarify: I do place *a priori* constraints on what demands can be legitimate or authoritative. But that's not the same as placing constraints on what demands a community might *treat* as legitimate or authoritative.
[127] Dembroff (2020), 12

categorization."[128] The idea here is that, rather than taking on an identity with some principled intent to destabilize dominant gender ideologies, some individuals and groups destabilize as a consequence of trying to be true to themselves. While Dembroff's paper focuses on the category *genderqueer*, they emphasize that many different gender groups can existentially destabilize in this way, including basically all LGBT subgroups.

Taking this idea on board and situating it in the dialectical framework I've developed, here's what we'll say about these. Gender identities can only be forged and expressed against a background of already-given conceptual resources. Perhaps from some ideal-theoretic perspective, we can say that our given (i.e., bioessentialist and binary) gender regime is objectionable, but that observation alone doesn't create workable identities for us. For some people, the best way to make sense of themselves against such a background will be to claim a ready-made normative status that dominant ideologies would deny them — this is perhaps the case for "binary trans" individuals.[129] In these cases, trans people make moves that are intelligible to their interlocutors because they draw on existing ideologies in relatively standard ways: "I am a woman" is a claim that everyone is used to hearing and acknowledging. Even if the background ideology on which this claiming draws is objectionable, the claiming itself can be seen as a way of grasping for greater autonomy — it's a claim we're obliged to respect so long as the ideologies that make it intelligible are still in force because, given those background conditions, it is for some people the truest way in which to act out their autonomy. Again, it's not that binary trans identities can be validated from a perspective that transcends all particular social arrangements — no identities can be validated in that way. It's rather that we, here and now,

[128] Dembroff (2020), 13

[129] I know from conversation that many trans people will object to this phrase. I use it here because it's the one Dembroff uses when contrasting trans people who exclusively identity as either men or women with genderqueer people who do not.

confronted by trans people demanding our recognition, have no good reason to deny them unless we can come up with a reason why they shouldn't be permitted to live the kind of life they want to live.

These binary trans identities, while essentially embedded in a binaristic gender system, are nonetheless destabilizing in Dembroff's sense and for the reasons I explained above. We have an ethical obligation to recognize binary trans identities because failure to do so involves an unjustified imposing of our will on trans people. But following through on this obligation is incompatible with keeping our binary gender concepts as we had them. This is how dialectical change works: resolution of a contradiction — in this case between our gender concepts and our general commitments to one another as free and equal — requires transformation of our thoughts and practices.

Similar things can be said about nonbinary individuals. These individuals respond to existing gender arrangements with something more like wholesale rejection. It's especially obvious how this can destabilize, and we can see this through the fact that those familiar only with dominant gender practices seem to view nonbinary identities as truly unintelligible.

This latter fact about intelligibility puts me in a position to say something about why I've focused so much on trans men and (especially) trans women and said little about nonbinary people. My aim throughout the dissertation has been to secure the result that trans-exclusive gender-ascriptions are false — that trans people are who they say they are and that those who deny this are wrong factually as well as normatively. Things are less straightforward when it comes to nonbinary identities, as these are more likely to represent a more radical break from dominant conceptual schemes. Remember: the hallmark of deep conceptual difference is that the claims made in one scheme will not be evaluable as either true or false in the other. While Kevin

Williamson is perfectly capable of denying that Laverne Cox is a woman, he may not be able to do the same thing with Sam Smith. He can't really deny what he can't understand.

What he *can* do is incorrectly attribute a binary gender to a nonbinary person. We might compare such cases to ones in which a child possesses a concept of natural number but not of rational number. The child cannot meaningfully deny that ¼ is ½ of ½. But they can assert, falsely, that there is no number between 0 and 1. Because their conceptual scheme is strictly less expressive than ours, all of their claims will be intelligible to us while some of ours will be unintelligible to them. Similarly, someone with a binary scheme can be said to speak falsely when they say that a nonbinary person is (exclusively) a man, even though they're incapable of understanding what the nonbinary person really is.

3.11.2 Limitations and Objections

There are two, roughly opposite, potential objections to my view I'd like to explore. The first is that it overgenerates: if the primary reason hegemonic gender concepts are objectionable has to do with their curtailing people's autonomy, then shouldn't that provide a reason for loosening our application-conditions for *all* identity concepts? The second is that it, in a sense, *under*generates by implying that it's at least in principle legitimate to challenge the identities of trans people. In both cases I actually accept the claim that the objector treats as bad: we *do* have some reason to loosen application-conditions for all identities and trans identities — like all identities — are in principle challengeable. I don't think either of these is a bad result when they're properly understood.

For the overgeneration objection it helps to think of social positions and attendant identities that we very clearly don't want to make available to people simply on the basis of first-personal identification. There should, for example, be some institutional barriers one has to

overcome in order to be recognized as a medical doctor. My answer to this sort of case is pretty simple and, I think, commonsensical. While generic considerations of autonomy do indeed generate *prima facie* reasons to respect people's claimed identities, those reasons will be of varying strength in different cases and will always be defeasible. Both things can be said in the doctor case.

First, the ways in which denying someone access to an occupational identity like *medical doctor* can curtail people's autonomy are structurally different from the ones in which denial of access to a gender identity do so. In the context of a gender-structured society, gendered normative statuses and their attendant identities are pervasive, persistent, and mandatory. They're imposed on people at birth on the basis of morally arbitrary physical characteristics and stick with those people throughout their lives across nearly all social contexts. Professional identities, in contrast, offer a wide range of possibilities (as many identities as there are professions), as well as the ability to transition between professional positions over time. Moreover, the significance of one's professional identity is much less 'sticky' than their gender identity — wedding ceremonies, for example, don't usually look different for doctors than they do for lawyers.

Second, even if some individuals may have a strong interest in being recognized as doctors, that interest can be defeated by society's interest in the category corresponding at least roughly to a certain body of knowledge and cluster of skills. The normative consequences of being a doctor involve patients putting their lives in the doctor's hands; it's fair for us to ask people to demonstrate competence before we entrust them with this responsibility. Compare this to the gender case, where recognizing a trans woman as a woman is costly for transphobes only in the sense that it may require them to stop buying into hegemonic conceptions of gender. But

those conceptions are illegitimate on principled grounds. Whereas medical licensure has the legitimate social function of identifying individuals who are competent to perform a kind of highly skilled and socially necessary labor, sex- and gender-assignment's principal function is male supremacy. So, we have good reasons to deny someone's claim to be a doctor, we generally won't have good reasons to deny their claim to be a woman.[130]

This leads us to the undergeneration worry. In what sense are trans identities implicitly open to challenge or committed to justification? This question is pressing because, as several in this literature have pointed out, queer subcultures typically treat first-person gender identifications as authoritative in a way that precludes their being challenged.[131] I've made clear why I think that we can't say that *all* gender identities are legitimate across the board, but I should nonetheless say something about how it could possibly be permissible to challenge someone's trans identification. I can think of two sorts of cases: insincerity and uncertainty.

First, and most obviously, someone can simply lie about their gender. Reactionaries are known to do this — think of Steven Crowder pretending to be a trans woman at the gym or the Colorado Springs shooter briefly claiming to be nonbinary after his arrest. The point of these moves is to set a kind of discursive trap for queer people and their allies: if we're obliged to respect any and all gender identifications, then don't we have to respect those of people we know are acting in bad faith? My answer to this is: no. While there's some risk involved here (e.g. we didn't *know for certain* that the Colorado Springs shooter didn't really identify as nonbinary), we

[130] The example of racial identity is something of an elephant in the room here. I don't have a fully articulated theory of racial identity - of its social function or of what a racialized normative status amounts to - so I can't give an adequate answer in terms of my theory of what entitles someone to claim to be Black, for instance. But as this section makes clear, my theory has room to claim that (e.g.) Rachel Dolezal is not Black regardless of the sincerity and seriousness of her claims to be so. I suspect we'll want to say that these are cases in which the inaccessibility of a racial identity does (or at least can) represent a fairly serious curtailment of individual autonomy, but that this curtailment is contextually justified by considerations having to do with the social functioning of race and what that person's claiming their desired identity would mean for others.

[131] Cf. Bettcher (2009) and Kukla and Lance (2022).

simply are not obliged to take someone seriously when they ‘identify’ as a gotcha. And it’s perfectly fine to challenge their claims out loud.

Second are harder cases in which people are simply undecided, or even mistaken, about their identities. I’ll give an example. Some years ago, a friend of mine was trying to decide whether they were asexual. They were unsure because, while their recent experiences seemed to be those of an asexual person, they weren’t sure this identity was a good fit for the range of experiences they’d had over the course of their life. They approached me to talk about this, and I offered my perspective. I did not try to tell them what they were, but I did challenge certain assumptions I thought they were making. I thought, for instance, that they were assuming a sexual identity could be legitimate only if it could be seen as steady over the course of one’s lifetime. So here I took on some limited authority to adduce considerations that my friend might consider relevant to their sexual identity. Importantly, this was only possible because of the bonds of trust we’d forged over time and because they invited my comments.

Ultimately, of course, it was up to them to decide whether they were asexual or not — I wasn’t going to tell them they were wrong once they’d decided. But this kind of case suggests a more general lesson: while various facts about the social function of gender and the ways in which transphobia is discursively enacted mean that it will very rarely be appropriate to call someone’s sincere transgender identification into question, this doesn’t mean that it’s illegitimate in principle. Rather, it’s that doing so appropriately will require a great deal of trust, sensitivity, and care — and can probably only happen in very limited contexts in which someone invites others to help them dig into themself.

Bibliography

1. *Suppose You Call a Sheep's Tail a Leg, How Many Legs Will the Sheep Have?* Quote Investigator 2015 [cited 2023 8/8/2023]; Available from: https://quoteinvestigator.com/2015/11/15/legs/.
2. *Chris Hayes: If This Isn't Terrorism, What Is?* 2017, MSNBC.
3. *Archived State Statistics*. 2017 8/8/2023]; Available from: www.shiftnc.org/data/state-statistics/archived-state-statistics.
4. *U.S. Code*, U.S. Congress, Editor. 2021.
5. *Waukesha: Tragedy Exploited by White Supremacists*. 2021 8/8/2023]; Available from: https://www.adl.org/resources/blog/waukesha-tragedy-exploited-white-supremacists.
6. Anderson, E., *Value in ethics and economics*. 1993, Cambridge, Mass.: Harvard University Press. xiv, 245 p.
7. Anderson, E., *Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce.* Hypatia, 2004. **19**(1).
8. Ásta, *Categories we live by : the construction of sex, gender, race, and other social categories*. Studies in feminist philosophy. 2018, New York, NY, United States of America: Oxford University Press. x, 140 pages.
9. Baldwin, J., *The evidence of things not seen*. 1st ed. 1985, New York: Holt, Rinehart and Winston. xiv, [2], 125 p.
10. Barnes, E., *Realism and social structure.* Philosophical studies, 2017. **174**(10).
11. Bauer, N., *Simone de Beauvoir, Philosophy, and Feminism*. Gender and Culture Series. 2001, New York, NY: Columbia University Press. 1 online resource (321 p.).
12. Beauvoir, S.d., C. Borde, and S. Malovany-Chevallier, *The second sex*. 1st American ed. 2010, New York: Alfred A. Knopf. xxi, 800 p.
13. Bettcher, T.M., *Evil Deceivers and Make-Believers: On Transphobic Violence and the Politics of Illusion.* Hypatia, 2007. **22**(3): p. 43-65.
14. Bettcher, T.M., *Trans Identities and First-Person Authority*, in *You've Changed: Sex Reassignment and Personal Identity*, L. Shrage, Editor. 2009, Oxford University Press.
15. Bettcher, T.M., *"Trans Women and the Meaning of 'Woman'"*, in *Philosophy of Sex: Contemporary Readings, Sixth Edition*, A. Soble, N. Power, and R. Halwani, Editors. 2013, Rowan & Littlefield. p. 233-250.
16. Blackburn, S., *Ruling passions: a theory of practical reasoning*. Vol. Oxford: New York. 1998, Oxford: New York: Clarendon Press ; Oxford University Press.
17. Brandom, R., *Freedom and Constraint by Norms.* American philosophical quarterly (Oxford), 1979. **16**(3).
18. Brandom, R., *Making it explicit : reasoning, representing, and discursive commitment*. 1994, Cambridge, Mass.: Harvard University Press. xxv, 741 p.
19. Brigandt, I., *The epistemic goal of a concept: accounting for the rationality of semantic change and variation.* Synthese (Dordrecht), 2010. **177**(1).

20. Brunner, J. *Trump supporter in state Senate says some protests are 'economic terrorism,' should be felonies*. The Seattle Times, 2016.
21. Burgess, A., H. Cappelen, and D. Plunkett, *Conceptual engineering and conceptual ethics*. 2020, Oxford University Press: Oxford. p. 1 online resource (x, 461 pages).
22. Burgess, A. and D. Plunkett, *Conceptual Ethics I.* Philosophy compass, 2013. **8**(12).
23. Burgess, A. and D. Plunkett, *Conceptual Ethics II.* Philosophy compass, 2013. **8**(12).
24. Butler, J., *Gender trouble: feminism and the subversion of identity*. 10th anniversary ed. Vol. New York. 1999, New York: Routledge.
25. Cappelen, H., *Fixing language : an essay on conceptual engineering*. First edition. ed. 2018, Oxford: Oxford University Press. 1 online resource (350 pages).
26. Carey, S., *The origin of concepts*. Oxford series in cognitive development. 2009, Oxford ; New York: Oxford University Press. viii, 598 p.
27. Chalmers, D.J., *Verbal Disputes.* The Philosophical review, 2011. **120**(4).
28. Connell, R., *Gender and power : society, the person and sexual politics*. 1987, Stanford, Calif.: Stanford University Press. xvii, 334 p.
29. Darwall, S.L., *The second-person standpoint : morality, respect, and accountability*. 2006, Cambridge, Mass.: Harvard University Press. xii, 348 p.
30. Davidson, D., *Inquiries into truth and interpretation*. 2nd ed. 2001, Oxford: Clarendon Press. xxiii, 296 p.
31. Dembroff, R., *Real Talk on the Metaphysics of Gender.* Philosophical topics, 2018. **46**(2).
32. Dembroff, R., *Beyond Binary: Genderqueer as Critical Gender Kind.* Philosophers' imprint, 2020. **20**.
33. Dembroff, R. and C. Saint-Croix, *'Yep, I'm Gay': Understanding Agential Identity.* Ergo (Ann Arbor, Mich.), 2019. **6**(20201214).
34. Dembroff, R.A., *What Is Sexual Orientation?* Philosophers' imprint, 2016. **16**.
35. Diaz-Leon, E., *Woman as a Politically Significant Term: A Solution to the Puzzle.* Hypatia, 2016. **31**(2).
36. Dummett, M., *Frege : philosophy of language*. 2d ed. 1981, Cambridge, Mass.: Harvard University Press. xliii, 708 p.
37. Dworkin, A., *Intercourse*. Vol. New York. 1987, New York: Free Press.
38. Eklund, M., *Choosing normative concepts*. First edition. ed. 2017, Oxford, United Kingdom: Oxford University Press. ix, 219 pages.
39. Fischer, H. *Arizona Senate votes to seize assets of those who plan, participate in protests that turn violent*. Arizona Capitol Times, 2017.
40. Fodor, J.A., *The language of thought*. Language and thought series (New York). 1975, New York: Crowell. x, 214 p.
41. Frey, R.G. and C.W. Morris, *Violence, terrorism, and justice*. Cambridge studies in philosophy and public policy. 1991, Cambridge ; New York: Cambridge University Press. x, 319 p.
42. Gibbard, A., *Thinking how to live*. Vol. Cambridge, Mass. 2003, Cambridge, Mass.: Harvard University Press.
43. Grover, D.L., J.L. Camp, and N.D. Belnap, *A Prosentential Theory of Truth.* Philosophical studies, 1975. **27**(2).
44. Haslanger, S. and J. Saul, *PHILOSOPHICAL ANALYSIS AND SOCIAL KINDS.* Proceedings of the Aristotelian Society, 2006. **106**(1).

45. Haslanger, S.A., *Resisting reality : social construction and social critique*. 2012, New York: Oxford University Press. xi, 490 p.
46. Hegel, G.W.F., A.V. Miller, and J.N. Findlay, *Phenomenology of spirit*. 1977, Oxford: Clarendon Press. xxxv, 595 p.
47. Hieronymi, P., *Responsibility for believing.* Synthese (Dordrecht), 2008. **161**(3).
48. Hirsch, E., *Quantifier variance and realism : essays in metaontology*. 2011, New York: Oxford University Press. xvi, 261 p.
49. Jenkins, K., *Amelioration and Inclusion: Gender Identity and the Concept of Woman.* Ethics, 2016. **126**(2).
50. Jenkins, K., *How To Be A Pluralist About Gender Categories*, in *The Philosophy of Sex: Contemporary Readings. 8th Edition*, R. Halwani, J.M. Held, N. McKeever, and A. Soble, Editors. 2022. p. 233-259.
51. Kant, I., P. Guyer, and A.W. Wood, *Critique of pure reason*. The Cambridge edition of the works of Immanuel Kant. 1998, Cambridge ; New York: Cambridge University Press. xi, 785 p.
52. Korsgaard, C.M. and O. O'Neill, *The sources of normativity*. 1996, Cambridge: Cambridge University Press. 1 online resource (xv, 273 pages).
53. Kuhn, T.S., *The structure of scientific revolutions*. 3rd ed. Vol. Chicago, IL. 1996, Chicago, IL: University of Chicago Press.
54. Kukla, Q. and M. Lance, *Telling Gender: The Pragmatics and Ethics of Gender Ascriptions.* Ergo: An Open Access Journal of Philosophy, 2022. **9**(n/a).
55. Kukla, R.a. and M.N.a. Lance, *'Yo!' and 'Lo!': the pragmatic topography of the space of reasons*. 2009: Harvard University Press.
56. Legg, C.a.C.H. *Pragmatism*. The Stanford Encyclopedia of Philosophy 2021 8/8/2023]; Available from: <https://plato.stanford.edu/archives/sum2021/entries/pragmatism/>.
57. Lewis, D., *Scorekeeping in a Language Game.* Journal of philosophical logic, 1979. **8**(3).
58. MacKinnon, C.A., *Feminism unmodified: discourses on life and law*. Vol. Cambridge, Mass. 1987, Cambridge, Mass.: Harvard University Press.
59. Manne, K.a., *Down girl: the logic of misogyny*. 2018: Oxford University Press.
60. McElhenney, R. *It's Always Sunny in Philadelphia*. Charlie Has Cancer. [4]. R. McElhenney. 2005.
61. McElhenney, R. *It's Always Sunny in Philadelphia*. Mac Is a Serial Killer. [10]. R. McElhenney. 2007.
62. Mendez, A. *NC bill has tough penalties for disruptive protesters, 'economic terrorists'*. Fox News Baltimore, 2017.
63. Mock, J., *Redefining realness : my path to womanhood, identity, love and so much more*. 2014, New York: Atria Books. 1 online resource (161 pages).
64. Scharp, K., *Replacing Truth*. 2013: Oxford: Oxford University Press.
65. Stahl, T., *Criticizing Social Reality from Within: Haslanger on Race, Gender, and Ideology*. Krisis: Journal for Contemporary Philosophy, 2014(1): p. 5-12.
66. Steele, K. and H.O. Stefánsson. *Decision Theory*. The Stanford Encyclopedia of Philosophy 2020 [cited 2023 8/8/2023]; Winter 2020:[Available from: <https://plato.stanford.edu/archives/win2020/entries/decision-theory/>.
67. Strawson, P., *Freedom and Resentment.* Proceedings of the British Academy, 1962. **48**: p. 187-211.

68. Taylor, C., *Interpretation and the Sciences of Man*. The Review of metaphysics, 1971. **25**(1).
69. Wynn, N., *Pronouns*. 2018.
70. Yousef, O. *Domestic terrorism charges in Georgia are prompting concern over political repression*. NPR, 2023.
71. Ziporyn, B.a.o.i.t., *Zhuangzi: the complete writings*. 2020: Hackett Publishing Company, Inc.