4
The Right and the Wren

Christa Peterson and Jack Samuel

A human being becomes human only among others.
Fichte, *Foundations of Natural Right*

4.1 Being human

We and other animals live in worlds richly colored by affect. A little wren has likes and dislikes, desires and aversions guiding it through the world. It does not know much of other minds, and so these attitudes presumably do not appear, to the wren, as personal, as significances things have only to it. They illuminate its world, and it follows where they lead.

Our own relationship to our desires is somewhat different. We understand our desires as our own; other people, we know, have different ones. But the world through our human eyes seems to come to light with affective and directive meaning far beyond whatever personal desires and preferences happened to show up in our bellies. Things are colored not just—or even primarily—as *wanted* and *unwanted* by us, but as *good* and *bad*. We may act on a natural desire, but between the pull and the action there's something more, something that gives us space to go in a direction other than wherever we'd be carried by our immediate inclinations: Is it truly good? Is it actually bad?

Ours is a valuer's world. And it's a shared world. We talk together of what's good and bad. Our world seems colored by other people's affect, not just our own. The grip of their pain seems immediate, not needing to be routed through any of our personal interests—as *bad*, as an urgent, pressing *reason* for us, just in virtue of our being there to see it, in the way we, as another person, can.

Christa Peterson and Jack Samuel, *The Right and the Wren* In: *Oxford Studies in Agency and Responsibility.* Volume 7. Edited by: David Shoemaker, Oxford University Press.
DOI: 10.1093/oso/9780192844644.003.0001

We will argue that these things are connected: that by coming to see others in a particular way, we gain a new evaluative capacity, one essential for our ability to take reflective distance from our own desires and set our own ends. The view we'll propose will be a kind of constructivism, in that it will portray moral reasons as arising from our very capacity to recognize ourselves as having reasons, but one in which that capacity itself is *essentially social*: it can develop only together with others, and, once developed, demands responsiveness to their perspectives.

In metaethics we often lose sight of the fact that what's at stake isn't just the status of moral *verdicts*, but the meaning and gravity of morally relevant *features*—things with visceral significance to us, like other people's pain, their deaths, love, compassion, trust, and betrayal—and their status in our practical thought.

Familiar forms of Kantian constructivism attempt to show that, in virtue of our capacity for practical reason alone, we are subject to a regimented moral law, furnishing in each case a determinate verdict. Our goal is more modest: to show that, in virtue of this capacity, we are answerable to the most essential and characteristically *moral* considerations—to other people's lives, pain, and well-being bearing on what we do. But this, too, would be an important success: to show that an appropriately minimal conception of the practical standpoint demands moral concern, that as soon as we *have* reasons, *other people* show up in them. We would have to sort out the tough cases amongst ourselves. But we'd escape what is, really, the deep threat of the idea of there being no universal moral order: people being free to treat us like we don't matter at all.

(The wren disappears into the brush.)

4.2 Outline

First, we'll describe two dilemmas the constructivist faces. One is the familiar dilemma between having an appropriately minimal conception of the practical standpoint and getting anything recognizably moral out of it, what we call the 'stowaway dilemma.' The other is a dilemma between two kinds of alienation: the constructivist characteristically tries to avoid agents being alienated from their reasons by driving them inside the agent, as rooted in their own practical thought. But, we'll argue, this risks alienating the agent from *other people*: other people showing up in the agent's reasons only derivatively, without an appropriately directive voice of their own. We'll

argue that these dilemmas recommend a picture of the practical standpoint prevalent in post-Kantian thought, where it is understood as somehow essentially social—in particular, as essentially requiring mutual recognition. We will develop one such picture, beginning with an account of mutual recognition as a pervasive feature of linguistic concepts, and culminating in what we call transformative expressivism.

4.3 Two dilemmas for the constructivist

4.3.1 Alienation

We navigate the world as a world of reasons, and we are reason-giving for each other: your suffering is a reason for me to help. This doesn't depend on my happening to care, or deciding to be open to it, but on your pain being what it is—and what it is to you—and my being the kind of creature that can recognize its significance. I'm accountable to you, and you to me, because of what we indelibly are.

But how? What makes your suffering a reason for me, in the immediate, insistent way that it is? What makes me answerable to it, whether I want to be or not? Non-naturalist moral realism provides a straightforward answer but little explanation: because facts about the reasons generated by the suffering of others are true. Normative reasons are part of the furniture of the universe, and one of them tells me to help you.[1]

This raises questions. The diversion through the deep facts of the universe seems to make your pain more distant from me than when it was in your voice. Non-natural reasons! What are these strange little things doing in our lives, with so central a role in our decisions about what to do? Are our lives, in so important a way as our ties to each other, really not a matter for us, but decided by the unthinking, unfeeling universe itself? How can these things written into the universe bind our practical thought, tell me what to value in a way I can't escape? What ensures these are even things I can act for, that I even can know? For most realists, these questions border on nonsensical. It is simply the nature of reasons to be fit to enter our

[1] The core of non-naturalist moral realism is that a complete ontology will make room for something normatively *sui generis*; see e.g. Shafer-Landau (2003), Parfit (2011), Enoch (2011), and Scanlon (2014). Naturalist realists take a different approach, but ultimately fall into one version of this dilemma or another, for reasons we don't have the space to discuss here. See (Samuel MS).

deliberation. This putatively additional property is in need of no more special explanation than their existence itself.

We want something more, and something more human. This is Korsgaard's central issue with realism. Traditional realism, she argues, leaves an explanatory gap: the existence of robust, mind-independent normative facts doesn't explain why these things count as reasons *for us*. It's just as though 'we have normative concepts because we've spotted some normative entities, as it were wafting by' (1996, 44).

We'll call this concern *normative alienation*: an anxiety about the possibility that we could have reasons to which we were motivationally indifferent, reasons whose relevance to our reflective deliberation was at best coincidental, reasons of which we could be systematically unaware. If we were so alienated from normative facts, morality seems irrelevant, unfit to play the role in our lives that it evidently does.

Constructivists take seriously the task of explaining how normative reasons are reasons for us, how they get their grip on us as agents. They begin by identifying a key characteristic or capacity of the kinds of creatures that are bound by normativity, and attempt to explain normative facts as somehow arising from it:

> Normative concepts exist because human beings have normative problems. And we have normative problems because we are self-conscious rational animals, capable of reflection about what we ought to believe and to do. (46)

Characteristically, these views hold that there are no normative facts independently of the practical thought of the creatures subject to them: moral facts are somehow the product of the very capacity that makes creatures morally responsible. But driving the source of normativity inside the agent risks driving us apart from each other. In their effort to avoid normative alienation, we will argue, constructivists fall into what we'll call 'social alienation.'

Sharon Street's (2008) Humean constructivism begins with the idea of a valuing creature: one that not only likes and dislikes, but also makes evaluative judgments. Evaluative *judgments*, as opposed to mere evaluative *attitudes*, are subject to the requirements of consistency and coherence. The evaluative judgments that survive such scrutiny, or arise from it, are true, relativized to the agent's particular practical point of view.

If your suffering is a reason for me, on Street's view, it's my attitudes that make it so. I might have started out valuing your well-being directly, or another of my initial values might have committed me to taking it up, for consistency and coherence's sake. Likely, many things I value will be in tension with indifference to your well-being. But if your suffering is a reason for me, it's because of something personal about *me*, because my attitudes happened to commit me to wishing you well. Even if you and I do value each others' well-being, we aren't reason-giving to each other directly—our reasons come from our own attitudes, not from each other. We can't, normatively, really reach each other.

Korsgaard's Kantian constructivism begins with a conception of a *reflective agent*: one that can endorse considerations as *reasons* and then act on their basis. Unlike Street's valuing creature, the reflective agent is supposed to be able to take reflective distance from her own initial desires, to ask who she wants to be, to set ends independently of what she happens to desire.

Reflective distance enables and requires us to decide what we will treat as a reason, and to have a basis for these decisions; we have to take up a 'practical identity,' a description under which we value ourselves. Taking something to be a reason, Korsgaard says, consists in adopting a certain maxim as a law governing my conduct, and we evaluate whether to do this under our practical identities: if the maxim is consistent with them, it's a reason for us, and if it's inconsistent, we are obligated otherwise, at the price of our identity.

One part of our identity, Korsgaard says, is inescapable. We *have* to value our humanity, Korsgaard says, if we are to value anything: that requires having practical identities, and our need to have reasons to live is what sustains our practical identities. And valuing our own humanity, she says, means valuing the humanity of others—and that makes your pain a reason for me.

The last step here is critical and opaque, and we will return to it. But for now, our focus is that this still seems to make your pain a reason for me in a way centered on myself.

To bring the worry into view, consider a similar concern that plagues perfectionist virtue ethics: that it turns morality into an exercise in self-improvement, that all of the normative force derives ultimately from the need to perfect one*self*. As Wallace (2019, 46) puts it, 'at the level at which normative requirements are explained, the interests of other people enter as occasions for the realization of virtue, rather than direct sources of requirements on the virtuous agent.'

On Korsgaard's picture, we might say, at the level at which normative requirements are explained, the interests of other people enter as stakes on which our identities may be sustained or torn, rather than themselves direct sources of requirements on us. All of our obligations come down to identity maintenance: 'An obligation always takes the form of a rejection against a threat of a loss of identity,' she says (1996, 102). Our accountability to others comes down to maintenance of the self.

We are supposed to be inescapably invested in the identity that generates our moral obligations; our concern for it isn't frivolous. To violate it is to lose our identity, Korsgaard says, 'to no longer be able to think of yourself under the description under which you value yourself and find your life to be worth living and your actions to be worth undertaking' (102). But it is about *us*.[2] It is almost as though the ultimate basis for your obligation not to harm others turned out to be that you would feel so guilty it would consume your life—that it would create an insurmountable problem *for you*, though in Korsgaard's case, the problem is practical, not emotional.[3] It is not just as though we have added, to Street's view, a sophisticated argument that we all actually do strongly desire the well-being of others. But it is not so far.

The threat of normative alienation calls for a theory of normativity that brings it closer to us, into the messy, embodied, and perhaps contingent features of our human lives. It pushes us toward 'agent-centered' theories, like constructivism: those that place the agent, the valuer, the reasoner at the center of their account of normative facts, emphasizing desires, values, preferences, or the capacity to practically self-determine as foundational to the explanation of how there is normativity at all. But in bringing normativity closer to ourselves we risk losing our moral grip on one another. Constructivists, we have seen, take the threat of normative alienation seriously, and fall into social alienation instead.

Korsgaard sees the importance of sociality. Her explanation of the key move between valuing one's own humanity and valuing the humanity of others relies on the fact that reasons, she says, are 'public in their very essence' (1996, 135). But it's not clear where that's coming from. Is the fact that reasons are essentially public somehow a result of the nature of reflective agency itself? If so, she doesn't show us *how*. If it's a further fact about

[2] Tarasenko-Struc (2019) similarly argues that, on Korsgaard's account, all of our obligations to others are ultimately derivative of our obligations to ourselves.

[3] In her later work, the problem seems less practical and more of a threat to the metaphysical unity of the self (Korsgaard 2009).

the nature of reasons, we seem to not be understanding reasons in terms of the nature of the capacity that generates them after all.

This brings us to our second, more familiar dilemma.

4.3.2 Stowaway

The constructivist's idea is to make morality depend on us, but on some capacity pre-theoretically plausibly necessary for us to be subject to it. They often call this key capacity *agency*, but what is important is that it is a capacity truly necessary for a creature to be morally bound. But this idea seems to resolve into a dilemma: From any appropriately thin theory of agency, it seems one cannot derive requirements thick enough to approximate morality. Any picture of agency thick enough to yield anything like morality looks like it has smuggled in something it isn't entitled to.[4]

Street's Humean constructivism falls on the thin horn: with consistency and coherence as the only regulating norms, agents are free to have a practical perspective from which they really ought to kill their husband, on account of their particular desires. It seems to give up morality to save the *ought*.

Korsgaard's Kantian constructivism seems to fall on the thick horn: the claims that reasons have to be universal and public don't seem sufficiently motivated, other than to get the moral result.[5]

The particular problem with the thick horn will depend on what we want from our constructivism. It is often taken as coming in response to a stalemate between realism and antirealism, seeking to recover the *objectivity* of moral facts from the prevailing noncognitivism of the mid-twentieth century, without running afoul of the naturalistic worries about traditional moral realism. If unconstructed moral norms have snuck in, they've brought who knows what metaphysical baggage.

If our main concern is responding to—of all people—the amoralist skeptic asking why he should care about morality, we need a notion of agency that's thin enough to be *conceptually* inescapable. This is the heart of Enoch's

[4] That something like this is a deep problem for Kantian constructivism is approaching consensus; cf. Tiffany (2012) and Schafer (2015).

[5] Our pessimism reflects what we take to be the general attitude of Korsgaard's readers that one way or another the regress argument fails. See e.g. Cohen (1996), Bratman (1998), Gibbard (1999), Regan (2002), Ridge (2005).

(2006) 'shmagency' objection: if we can cut the normativity-generating part away and still be left with a practically functional capacity, the amoralist's question about why he should care about being moral persists as a question about why he should care about being an agent rather than a 'shmagent'—like an agent but without the extra bit. If whatever is generating normativity isn't actually a feature of everyone subject to morality, we haven't bound everyone we should have.

We're centrally looking for an explanation of why other people can give reasons to us. If sociality has been stuck at the end, we haven't explained it.

4.4 A post-Kantian thought

To answer the stowaway dilemma, the constructivist needs to generate something recognizably moral from something she's entitled to: from something genuinely inseparable from the core of the key capacity, from something not already moral, from something truly necessary to be bound by morality. To answer the alienation dilemma, our moral reasons need to be rooted *inside* the agent in the right way to not be normatively alienating, and yet somehow also ultimately focused *outside* the agent, on others.

We are deeply social animals! If we are going to find a source of normativity, a more basic kind of sociality is a compelling place to look. But, as Korsgaard notes, metaethicists have tended to treat this as an inappropriate basis for a vindicating account of morality, as though it would make it not really normative after all. If it were a contingent feature that wasn't necessary for us to be subject to morality's demands, it wouldn't give us an account that covered all the creatures it should.

In the post-Kantian tradition, there is a conception of agency that is intriguing here: agency as somehow *essentially social*. This seems like exactly the kind of thing the alienation dilemma demands: it could allow us to drive morality into the agent, without severing it from others. And if sociality were in some way *truly essential* to agency, in a way that somehow demanded ongoing responsiveness to others, we could have a way between the stowaway dilemma's horns: *other people* could be the source of the characteristically moral content, if our ability to evaluate and respond to reasons at all were somehow inseparable from their showing up in them.

Like the Kantian, the post-Kantian begins with our capacity for reflective agency. But the post-Kantian holds that that capacity is not possible alone. Reflective action requires self-consciousness, and self-consciousness is,

somehow, essentially social: for Fichte and Hegel, because it requires *mutual recognition*.[6]

We are *recognizers*: we are in the business of assigning significances to things, classifying them in ways that practically guide us—as *food*, as *sharp*, *fragile*, *dangerous*, *healthy*, *holy*, etc. In doing so, we adopt a particular orientation toward the object, manifested in how we treat, deliberate, and feel about it: we *eat* food, *crave* it when hungry; we avoid a *dangerous* thing when we can, or handle it with care, and might feel *afraid* around it.

We come to understand ourselves as recognizers by butting up against other recognizers: in more modern terms, we come to have theory of mind, the ability to understand our mental representations as our *own*, where other creatures' or the world may differ, by encountering those others.

Mutual recognition is something further. We don't just independently each recognize the other as a creature who attributes significances to things, in the way ravens might, trying to trick each other by pretending their caches are somewhere they're not. In mutual recognition we classify each other as *fellow* recognizers, granting each other not just a significance, but *meta*-significance: as having a perspective that directly bears on the significances *we* assign. Mutual recognition orients our minds to *meet*: we see each other as joint significance attributors and arbiters, who recognize each other as such, and are co-constitutors and co-determiners of a shared conceptual world.

On the post-Kantian picture, sharing a world awakens us to ourselves. It enables us to have reasons, instead of mere inclinations, to go beyond appearances, to judgments. We will develop a linguistic version of this claim.

4.5 Shared concepts

The idea of granting each other any kind of authority in assigning significances perhaps seems already moral. But it is mostly mundane. In fact, it seems like a pervasive part of language.

The wren's classifications are personal. The behavior of other wrens might offer evidence, but what significances it assigns are ultimately a private matter. Things changed for us, in coming to have language.

[6] This discussion draws loosely from Fichte (1797) and Hegel (1807), filtered through the interpretive work of Pinkard (1994), Brandom (2007, 2019), and Pippin (2008). See also Clarke (2009) and McNulty (2016).

The wren communicates with alarm calls, calling out to others when it senses danger and hiding away when other birds call. It might be oriented to make some adjustments to its underlying category according to the attunement of other birds' alarms, to learn about dangers from them. But if it notices another little bird calls alarms in response to things that are stably outside the wren's own category, that just means the other isn't an especially reliable indicator of what concerns the wren—and things end there.

That's no way to get language off the ground. Compare how we would respond: When you say something is dangerous that I really think is not, I don't respond to you like I would respond to a robot trained to identify hazards that has suddenly spit out a bizarre result. We are drawn to give reasons, to explain our judgments. Some of these practices could be construed as just gathering information for ourselves, with more means than are available to the wren. But our concern outstrips that. If we don't have different background information, we still want to settle the issue: Can we distinguish a sense you're using from the one we are? What standards are you using? What standards should we use? We treat each other as though we both have a kind of standing in how our shared concepts are assigned. We try to get on the same page.

The birds' communication isn't conceptually generative; they have the categories they do and don't seem to take on new ones for the sake of understanding each other.[7] But learning language involves taking on an enormous number of new ways of classifying things—it requires responding to the mere fact that others categorize things in a certain way by creating your own mental representation that does the same. Our linguistically mediated mental representations—what we'll call *concepts*—are at their core tied to other people's. Communicating with words requires sharing meanings, which requires coordinating concepts: we each have to maintain our own mental representation in such a way that they *correspond*, they 'mean the same thing,' are, somehow, 'the same concept.'

Our concepts correspond centrally because they're *meant* to. We understand them as matching up with other people's, and they're programmed to make it work. We perhaps shouldn't say that conceptual coordination strictly requires mutual recognition. The universe is a big place; maybe some kind of creature can have a language that one of them is the queen of and all the rest fit their mental representations to hers. But the way *we* get

[7] It is nevertheless better to be a bird!

the requisite coordination, and do so organically, on our feet, is through mutual recognition: we treat each other as *joint* authors of our shared concepts, in a deep and automatic way. It has to be deep, because our concepts have to be stuck together in a way more fundamental than any particular characterization if they are to be brought back into correspondence when they diverge. It has to extend beyond speakers of our own language, because the point is to come to speak the same one. It has to be largely automatic because of the sheer scale of coordination that needs to occur.

Our concepts are designed to function in a shared classificatory and conceptual project, and are structured in a way that is suitable for shared, not private, application: they don't include things that couldn't be included in other people's as well, and they have application conditions that are in principle sharable. 'It bit me' is an appropriate ground for the judgment that a creature *bites* insofar as I could be interchanged with any of us; 'it bit *me*', where special weight is given to the fact that it was *me*, is not.

We of course don't have to accept the judgment of someone deploying a shared concept. But our use of them is naturally answerable to others: others can challenge it, and in reply we need to account for our use in a way that they can, from their perspective, assess, using standards they could adopt themselves.

This looks like the kind of deep but basic sociality that we might be able to build from: a generic kind of accountability to others, one bound up with our judgment.[8]

So far, of course, it doesn't give us moral concern: what we're after is other people being pushed to give standing to what really matters to us, not whether we think a hot dog is a sandwich. And it doesn't give us reflective agency: it might be bound up with reflective judgments about concepts, but that in itself doesn't give us any kind of control over our desires or ends.

But this would change if we had a lexicalizable concept such that giving standing to other people's judgments about it constituted giving standing to what really matters to them; a concept such that applying it to something constituted setting that thing as an end.

In metaethics we have a view that posits something like that: expressivism.

The proposal we offer, a development of Peterson (MS), is expressivist in that evaluative judgments like 'that's good' are construed as a matter of

[8] Compare Korsgaard's use of Wittgenstein's 'Private Language Argument' (*Sources* §4.2.3). Our point here is more modest: regardless of whether a private language is possible, we can't use one to communicate with each other. To do that, we have to coordinate.

taking an affective and directive stance, not describing anything in the world. Unlike more familiar expressivist views, it takes the lexicalization of these affective significances to be transformative: they become predicable concepts, and work in ways that their parent attitudes could not.[9]

4.6 A human language

Some of our lexicalized concepts—*passenger*, *prime*—are categories only acquired through language in the first place. But others look like versions of pre-linguistic significances: *object*, for example, appears to be a primitive kind of mental representation (Carey 2009); *food* looks like a pre-linguistic practical significance orienting us to eat a thing when hungry; *blue* a phenomenological visual significance; *sweet* and *bitter* phenomenological and appetitive significances; *sad*, *disgusting*, *suspicious*, *exciting*, *surprising* emotive significances.

Philosophers talk about 'representation' as though it must be representing things in the world, *as they are in the world*. But, of course, in the pretheoretic sense of the term, the painting 'Guernica' represents the bombing of Guernica, not as it is to the universe, but as it is to us—or first to Picasso, and conveyed to us: as horrifying, as devastating, as incomprehensible, as a nightmare. It's a representation in a human language, a conveying of visceral significances things can have *for us*.

We can also convey human significances with language: the phenomenal significance of a thing being *blue*, the emotive significance of a thing being *sad*. When we accept these judgments, we classify the thing under those significances; we 'see' the thing as if it looked blue, or as if we felt sad about it.

On this account, these concepts, in the first instance, aren't attempts to get at anything out in the world, but mental significances that we share with each other.[10] It's a little odd to say that 'the ocean is blue' *expresses* blue phenomenology. But it *conveys* it, *records* it, and *presents* things with it. If I tell

[9] We intend to use 'transformative' here in something like the sense in Boyle (2016), though not everything we say here is likely compatible with a transformative conception of rationality as he envisions it.

[10] 'Mental significance' isn't meant to suggest a single unified kind: it includes a wide variety of things, from plausible pre-linguistic mental *categories*, like *food*, to qualitative features that are not obviously pre-linguistically sortal, like color in visual experience. *Food* orients us to treat the thing in a specific way, color doesn't have any intrinsic practical significance. But both are ways that a thing can be consciously marked to us.

you that alpine forget-me-nots are blue and in considering it you summon an experience somehow like the one you'd have if you were actually perceiving blue flowers, you've understood me well.[11] The significance your *blue* concept has to you, and what you hope to convey to others in calling things blue, is this phenomenological content, not anything about wavelengths.[12] The significance your *sad* concept has to you, and what you use it to convey, is an emotive content.

Non-linguistically, these significances don't manifest only in the hot, vivid renderings we feel in the first moment; we record them in cooler forms. The lexicalized manifestations seem to look more like the latter. So accepting that something is *blue* in a linguistic mode doesn't necessarily raise a sensation of blueness; accepting that something is *sad* or *disgusting* doesn't necessarily raise a full-blown feeling of sadness or disgust—though it might. But it marks the object in the way sadness or disgust would: with an emotive significance. It orients you to treat it in the way you're oriented to treat things you've felt sad about or disgusted by.

The lexicalization of these pre-linguistic significances commits us to a particularly deep kind of mutual recognition. The only sense in which we can *share* phenomenological and other psychological contents is by admitting other people's as corresponding: *your* color sensations, not just my own, have to be admitted as rendering the significance—as characterizing the concept—for you to get a phenomenological grip on my claims about color. For each of us to have a concept that connects up to a pre-linguistic significance in the right way, we have to treat those significances, in each of our heads, as having the same fundamental relationship to the shared concept. We can't assign any phenomenological content to *orange* if you'll only admit your visual experiences as bearing on it.

Pre-linguistically, these significances are seemingly applied automatically, by characteristic faculties: things feeling cold, looking a certain shade, tasting sweet. To the wren, that's it. It's not obvious that the bird ever has occasion to evaluate competing color representations: With these significances rendered only by its perceptual system, it's never confronted with

[11] This is arguably Hobbes' account of terms for 'fancies': privately, they function as 'notes of remembrance' that allow us to resurrect a decayed sense impression—see *Leviathan* iv.3, cf. ii.3 and ii.10.

[12] Not everyone we speak with *does* have access to the phenomenal content: blind and colorblind people can talk to color sighted people about color. The concepts are held together by being *meant to correspond*, even when some speakers don't have access to the phenomenal content that characterizes the concept for those that do.

inconsistent representations of a single thing at a given time—the fact that a thing can't simultaneously look both ways is what makes them inconsistent. And it's not inconsistent for color phenomenology to change over time; the purple of the sky at dusk doesn't compete with the gold of it in the morning. The bird particularly doesn't seem likely to develop any sense of its emotive significances being mistaken. They're just not the kind of thing that the world might push back on.

Things change when the significances are rendered into lexicalizable concepts—they change under mutual recognition. To play their communicative role, the concepts have to be assignable in a different way than the pre-linguistic significances, and to things the pre-linguistic versions couldn't reach: we have to be able to record things we have never ourselves seen or felt as *cold* and *blue*, on the basis of testimony. Truly *competing* significance assignments become pervasive when we start sharing them with other recognizers. You saw something one way, I saw it another, and once we present those impressions under concepts that are supposed to correspond, we have an inconsistency, and a conflict to negotiate.[13]

The concepts tell us to work it out. The need is both interpersonal and internal. The standing others' non-linguistic impressions and conceptual assignments by necessity have on the shared concept means each of our representations is called into question by the others' report, even to ourselves: my concept *to me* is not just a matter of my own impressions, nor yours to you. How will we resolve the question? We have the automatic non-linguistic assignments to guide us, and the kind of meaning, and any practical import, the significance has to us. But we need more: we start appealing to different standards that others could apply as well, offering grounds for concept application that others could accept.

The space made for these reflective judgments is a space between us, and functions accordingly. As with other linguistic concepts, the standards we adopt have to be at least in principle sharable and subject neutral. The fact that *I* was the one that saw or felt something of course tends to give it more grip, but my concept doesn't itself prioritize me, because the concept builds itself to be shared.

Lexicalizable concepts don't *replace* the corresponding pre-linguistic significances, but they *transform* them: they change their status. A thing

[13] This discussion picks up on a theme in Brandom's 'semantic' reading of Hegel, according to which 'determinate negation' drives conceptual determinacy and refinement. See especially 1.IV and 3.II.

marked to the wren with color phenomenology is not construed as either *looking* or *being* a color; the little bird is not called to make such distinctions. But we are. Our non-linguistic mental faculties continue to mark things with the same significances as always, no more on the basis of our conceptual standards than before. But if we *reflectively approach* these classifications, it's in the form accessible to and suited for that kind of thought: the lexicalized, sharable concept, with interpersonally neutral standards for application. The pre-linguistic significances become *reflectively* governed by interpersonally accessible standards, not exclusively attached on their basis. With our color concepts in hand, automatic color phenomenology becomes the *'appearance'* of color, which may or may not be inscribed in the corresponding conceptual color *judgment*.

This is tantalizing. Lexicalization seems to enable us to take reflective distance from significances that were, pre-linguistically, automatically applied, by making them subject to mutual recognition—by giving other people standing. With emotive significances like *sad* and *joyful*, we are reaching into the things we really care about. For our account of reflective agency and moral concern, we need one more step: we need something more general.

4.7 Transformative expressivism

The wren, now, has found a beetle to eat.

The beetle being something the wren especially likes to eat appears, to the wren, just as the importance of the beetle *simpliciter*, not a significance the beetle has *only to it*. Without much theory of mind, the wren can't represent its desires and preferences as constrained to itself, with space held for other creatures' perspectives to differ. Its preferences are free to project out into the world, to appear like any other significance a thing can have to the wren—as being another wren, another wren's song, a place to nest, a stick suited to build, an egg to push out of the nest, a chick of its own, the way to fly when it gets cold.

Some of our attitudes—aversion, fear, grief, anger, dread—and sensations—pain, nausea, breathlessness—are significant to us in part by having negative valence. Other attitudes—desire, satisfaction, pride, affection—and sensations—comfort, rest for exhausted muscles, catching one's breath—are significant to us in part by having positive valence. These valences tend to project onto the associated object: the objects of our fear, grief, and anger

appear to us as having negative significance; the objects of our desire, pride, and affection appear to us as having positive significance. Negative valence orients us *away* or *against*; positive valence orients us *toward* and *for*.

The wren likes some things and dislikes others, but without much awareness of other minds; it doesn't have a concept, or even a proto-concept, of evaluative attitudes. What it has, if anything, is a proto-concept of *good*: mental representations of things as positively and negatively valenced, as to-be-pursued and to-be-avoided, that don't make any particular reference to itself. The 'language' of the significance is affective, the message is directive, and it's about the object, not the wren. Representing attitudes as personal comes later: our beliefs in the first instance present themselves to us as about the world, representations we think of as *our beliefs* only if reason arises. Similarly, the world still appears to us with valence, valence we can, when appropriate, sequester as the issuance of our personal attitudes.

How could a proto-concept of 'good' like the wren's—valences and directives issued by mere attitudes, marking up the world—become the full concept *good*? It is simple, and alchemy: in the same way any other mental significance becomes a full, lexicalizable, concept, it gains a coordinative center of gravity.

The normal workings of mutual recognition have special consequences here, because *good* and *bad* are special concepts: assigning *good* to something orients us toward it and for it, assigning *bad* to something orients us away from and against it. It's a matter of taking a practical stance, of setting something as an end, and so reflective control over it amounts to reflective agency, and other people having standing on the question of what's *good* amounts to their having standing on what we do.

Why would we have started doing that—letting other people in to influence our desires, to draw our interests to their own? Affective significance is a central part of the natural meaning we experience in the world—and the meanings that matter most to us. It's so nice to share them. It's so nice to be and feel together. More temperately, it's of course very useful. Our valenced experiences include the most urgent messages our bodies send us, and it's good to be able to share them, and share the practical thrust before we have a precise explanation. The valenced sensations we experience largely overlap, and in critically important ways: what tastes good and bad, for example, is largely recreational now, but a matter of life and death when we were in the business of eating unknown and imprecisely identified things in the forest. As creatures that constantly spontaneously collaborate, it helps to be able to coordinate priorities and aims.

What guides what we call *good* and *bad*? Like other lexicalized pre-linguistic significances, three things: First, the spontaneous valenced attitudes that are its foundation. The mental significances unreflectively projected by these attitudes become something like the 'appearance' of *good* and *bad*: provisional assignments we may or may not ultimately inscribe. Second, other applications of the lexicalized concept, the things we've already judged *bad* or *good*. With those two alone, we might have a tool to better systematize our desires, to affirm ones that fit together nicely and reject ones that undermine. But it wouldn't enable us to measure a desire by anything but our others; it wouldn't enable us to truly, from a perspective not controlled by our immediate inclinations, set our own ends. What allows us to take the actual distance characteristic of reflective agency is the third thing: the way our lexicalizable concepts are built to be shared—to correspond to the concepts of others, so I can get, with words, a thought out of my head and into yours.

These criteria for application might seem to leave things too open. The generic *good* and *bad* do leave things wide open. We can use *good* to describe the relief of a dying child's pain, or an ice cream cone. It's not surprising that people have developed many more specific evaluative vocabularies—in English, for example, *duty*, *courage*, *right*, and *wrong*—with further constraints on their application. We will discuss them soon. But now, our focus is on our freewheeling, generic positive and negative evaluations, that admit the exceptionally wide variety of things we can count as a reason.

And that demand we count certain things. Other people's reflectively endorsed ends bear directly on our own, because reflectively endorsing an end is a matter of applying a shared concept that, like any other lexicalized concept, we by necessity understand as not personal, not a matter of our sole discretion, but corresponding and tied to others', to be coordinatively applied.

As a concept rendering a pre-linguistic mental significance, it also goes deeper: other people's non-linguistic presentations of it also have to be admitted as bearing on the concept—as the 'appearance' of it—for it to have the right significance to each of us. When the significance at hand is valence, that means other people's suffering and despair appear as *bad*, and their joy, the relief of their pain, appear as *good*. You're not forced to affirm these 'appearances' into judgments any more than you are with your own unreflective valenced representations, though if denying that serious suffering is itself bad is conceptually possible, it is just barely: the way real suffering

presents itself to the sufferer—the overwhelming, inescapable negative feeling, the desperation to escape—is a paradigm of what, at its extreme, *bad* means to us, of the kind of significance we, in drier moments, use the concept to evoke.[14] But most importantly, other people's valenced feelings and sensations relate to the shared concept in the same way and with the same essential standing as your own. You can't admit your pain as characterizing what's *bad* and deny that others' does, just because they're not you, any more than you could admit your own color experiences as characterizing *orange* and deny that anyone else's do. Your reasons for rejecting the apparent badness of someone else's suffering have to be interpersonally available: they can't prioritize you for the sake of being you.

Reflective agency requires being able to take an evaluative perspective somehow external to your own desires. As far as we can tell, the universe's standpoint is one of hydrogen, helium, and rocks—a standpoint poorly suited for these purposes. Even if the universe somehow *has* a normative perspective, it won't help us gain the capacity necessary for reflective agency, because it's not forthcoming. Instead, we find reflective distance in the space between us: in coming to be inclined and able to consider other affective perspectives, without being obligated to accept any particular judgment; to be able to reflectively grant things we don't spontaneously desire this status on the basis of the kind of reasons that *could* motivate such a desire; to hear others' responses to the values we have; to develop them, together, where they accord.

Our reflections happen in this shared space, but our judgments are ultimately our own. We can build, gradually, a foundation of values we've reflectively chosen, that were considered not just in light of our other immediate inclinations, but on the kind of considerations that might move many of us.

4.7.1 Dilemmas

We were after a *social* constructivism, in hopes that it would provide a way out of our dilemmas. The transformative expressivist proposal does well.

It's not normatively alienating, because moral reasons are raised within the agent's own capacity for reflective deliberation. It's not socially alienating, because we have standing in each other's practical deliberations directly,

[14] Cf. Manne (2017), for whom certain bodily states constitute moral claims directed at any creature close enough to notice.

as fellow recognizers: other people's pain bears directly on the most basic concepts, *good* and *bad*, we use to set our ends. The significance your pain has to you is what makes it a reason for me, given my being a creature that can recognize it.

The account is entitled to its resources twice over. First, the sociality it appeals to is not already moral, because it is taken to be a feature of all lexicalizable concepts: you have to be social about *itchy* and *potato* in the same way. Second, it gives a genealogy on which the very feature that makes you reason-giving for me is also the foundation of reflective agency: I get reflective distance from myself by becoming open to other people. They are inseparable.

And it gets a recognizably moral result: not a full moral law, but moral *concern*. Other people have standing with regard to what we reflectively take as our ends. Their suffering presents itself as a reason against, and in my reflective deliberation, I am pushed to give it the standing of my own.

This is, of course, not a story of how we all became good. It is a story of how other people got into our reasons, by default and from the start. It's a story of how we became accountable to other people's pain, how the justificatory burden for disregarding it is legitimate and high.

The concept is not itself inconsistent with oppressive or genocidal ideologies that devalue some kinds of people: it demands interpersonally accessible reasons for treating people differently, but can't itself ensure the beliefs people take up are good or decent.

The account validates our moral appeals but does not provide any new arguments with which to convince the amoralist, which seems just as well. The appropriate response to someone declaring that they don't intend to treat others as though their lives and well-being amount to anything is to get ready to fight them, not debate them, and certainly not to base our ethics around them. Debate will never be an effective response to people who don't care to treat others decently, and metaethics can't protect us from cruelty and cold disregard. We have to protect each other.

4.7.2 Normative gardens

What the nature of our practical thought gives us isn't a determinate moral law, but moral *concern*, and an opportunity: the ability and readiness to build, together, rich normative worlds that grip us and that bring meaning to our lives. In addition to lexicalizing pre-linguistic, generic positive and negative evaluations, language enables us to develop novel evaluative

concepts together, wrapping the affective and directive force of the generic *good* and *bad* into more complex wholes, all with the social features of linguistically mediated concepts that drive us together in conception. We can build normative and affective worlds, and we are prolific. We can imbue negative evaluation into concepts picking out specific kinds of action, like *murder* or *kinkshaming* or—especially forcefully—people who perform those actions, like *rapist* or *scab*. We can develop concepts that constrain *how* we evaluate: *right* and *wrong*, for example, seem to ask for a rational, principled justification; some modes of evaluation like *efficient*, *productive*, and *strategic* cleanly excise evaluation of the goal from evaluation of the action. Rather than living under a tidy, singular moral order, we find ourselves in an overflowing garden of normative concepts.

The generic affective content by default grips us wherever it's imbued, but unlike *good* and *bad*, these more particular concepts aren't cognitively inescapable. Which we habitually use has practical consequences: different concepts with different application conditions will lead us to imbue different features with their motivational force. The push we feel to rescind *right* and *wrong* judgments when it looks like we won't be able to offer a rational, completely generalizable defense of them does not extend to *fucked up* or *not ok* judgments; we will relinquish the former in situations where we could maintain the latter with full motivational force.

Other people's suffering is not a paradigmatic instance of all of these more complex concepts. In addition, evaluative concepts with additional content don't all share the evaluative egalitarianism that the collaborative demands of shared concepts generate in a concept, like *red* or *good*, whose essential content is just a fundamental way things 'appear' to each of us. The fact that we are pulled to coordinate on judgments about what is *sinful*, *just not done*, or *insubordinate* doesn't give each of us this kind of 'equal say,' because the concepts are in their content oriented toward particular sources of authority: God, established convention, hierarchy.

4.8 Real alienation

All this raises the specter of a final genre of alienation, what we'll call *real alienation*.

There are both normative and social varieties. Real normative alienation arises when a person experiences the normative framework within which

they exercise their agency as alien, as not their own, incapable of being their own. It can arise when, because of our cultural context, we habitually use evaluative concepts that involve content we don't, or wouldn't, endorse: for example, we might not identify with a hegemonic mode of practical evaluation but give it a central action guiding role in our thought in a defensive way, because we expect others will evaluate us in its culturally dominant terms. Or we might use an inherited concept without scrutinizing its content.

Real social alienation can arise when you habitually use only modes of practical evaluation that don't connect you to others in the right way. If by default you only evaluate your actions for whether they're *efficient* (or *admirable*), you're not, in your actual practical thought, connecting to others, even though you have other—more primitive, inescapable but neglected—concepts that would.

Importantly, this final kind of alienation is not a feature of the metaethical theory, but a troubling feature of life it does not explain away. The picture we've provided can make sense of radical criticism—though it might be difficult to make, and to have heard, under certain hegemonic frameworks. But metaethics can't prevent us from being entangled in evaluative frameworks that are disenfranchising, oppressive, or otherwise inhospitable to our lives. What we can ask of theory is to equip us with resources that might help us identify and resist it.

> In acquiring one's conception of the world one always belongs to a particular grouping which is that of all the social elements which share the same mode of thinking or acting. When one's conception of the world is not critical and coherent but disjointed and episodic, one belongs simultaneously to a multiplicity of mass human groups. The personality is strangely composite: it contains Stone Age elements and principles of a more advanced science, prejudices from all past phases of history at the local level and intuitions of a future philosophy which will be that of a human race united all over. . . . The starting point of critical elaboration is the consciousness of what one really is, and is "knowing thyself" as a product of the historical process to date which has deposited in you an infinity of traces, without leaving an inventory. (Antonio Gramsci, *Selections from Prison Notebooks*)

The wren sings a song it seems to have learned from its neighbor.

References

Boyle, Matthew. 2016. 'Additive Theories of Rationality: A Critique.' *European Journal of Philosophy* 24 (3): 527–55.

Brandom, Robert. 2007. 'The Structure of Desire and Recognition: Self-Consciousness and Self-Constitution.' *Philosophy and Social Criticism* 33 (1): 127–50.

Brandom, Robert. 2019. *A Spirit of Trust*. Harvard University Press.

Bratman, Michael. 1998. 'Review of Korsgaard's The Sources of Normativity.' *Philosophy and Phenomenological Research* 58 (3): 699–709.

Carey, Susan. 2009. The Origin of Concepts. Oxford University Press.

Clarke, James Alexander. 2009. 'Fichte and Hegel on Recognition.' *British Journal for the History of Philosophy* 17 (2): 365–85.

Cohen, G.A. 1996. 'Reason, Humanity, and the Moral Law,' in Christine Korsgaard, ed., *The Sources of Normativity*, pp. 167–88. Cambridge University Press.

Enoch, David. 2006. 'Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action.' *Philosophical Review* 115 (2): 169–98.

Enoch, David. 2011. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford University Press.

Fichte, Johann Gottlieb. 1797. Grundlagen Des Naturrechts Nach Prinzipien Der Wissenschaftslehre. C.E. Gabler.

Hegel, G.W.F. 1807. Phänomenologie Des Geistes. Würzburg.

Gibbard, Allan. 1999. 'Morality as Consistency in Living: Korsgaard's Kantian Lectures.' *Ethics* 110 (1): 140–64.

Korsgaard, Christine, ed. 1996. *The Sources of Normativity*. Cambridge University Press.

Korsgaard, Christine. 2009. *Self-Constitution*. Oxford University Press.

McNulty, Jacob. 2016. 'Transcendental Philosophy and Intersubjectivity: Mutual Recognition as a Condition for the Possibility of Self-Consciousness in Sections 1–3 of Fichte's Foundations of Natural Right.' *European Journal of Philosophy* 24 (4): 788–810.

Manne, Kate. 2017. 'Locating Morality: Moral Imperatives as Bodily Imperatives,' in Russ Shafer-Landau, ed., *Oxford Studies in Metaethics 12*, chapter 1. Oxford University Press.

Parfit, Derek. 2011. *On What Matters: Volume Two*. Oxford University Press.

Peterson, Christa. MS. Dissertation in progress. University of Southern California.

Pinkard, Terry P. 1994. *Hegel's Phenomenology: The Sociality of Reason*. Cambridge University Press.

Pippin, Robert. 2008. *Hegel's Practical Philosophy: Rational Agency as Ethical Life*. Cambridge University Press.

Regan, Donald H. 2002. 'The Value of Rational Nature.' *Ethics* 112 (2): 267–91.

Ridge, Michael. 2005. 'Why Must We Threat Humanity with Respect? Evaluating the Regress Argument.' European Journal of Analytic Philosophy 1 (1): 57–73.

Samuel, Jack. MS. 'Alienation and the Metaphysics of Normativity: On the Quality of Our Relations with the World.'

Scanlon, Thomas. 2014. *Being Realistic About Reasons*. Oxford University Press.

Schafer, Karl. 2015. 'Realism and Constructivism in Kantian Metaethics: Realism and Constructivism in a Kantian Context.' *Philosophy Compass* 10 (10): 690–701.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence*. Oxford University Press.

Street, Sharon. 2008. 'Constructivism about Reasons,' in Russ Shafer-Landau, ed., *Oxford Studies in Metaethics 3*, chapter 8. Oxford University Press.

Tarasenko-Struc, Aleksy. 2019. 'Kantian Constructivism and the Authority of Others.' *European Journal of Philosophy* 28 (1): 77–92.

Tiffany, Evan. 2012. 'Why Be an Agent?' *Australasian Journal of Philosophy* 90 (2): 223–33.

Wallace, Jay. 2019. *The Moral Nexus*. Princeton University Press.