Expressivism, Moral Psychology and Direction of Fit

Carlos Núñez

Forthcoming in *Oxford Handbook of Metaethics*, edited by David Copp and Connie Rosati.
(Draft. Please cite only the published version)

Expressivists claim that normative judgments ('NJ') are—unlike regular descriptive beliefs—non-cognitive states of mind: attitudes whose function is not to represent the world as being a certain way, but to motivate agents to act, or react, in specific manners. This allows expressivists to avoid the metaphysical, epistemological and psychological problems of explaining what normative facts or properties are, how we can know anything about them, and why it is that NJ are so closely tied to motivation. The first two problems are avoided because expressivism does not appeal to normative facts or properties at all; the last one, because, if normative attitudes are motivational states, then no surprise they are so closely tied to motivations.[1]

This is enough to make expressivism a very attractive theory. But it does not come without its own explanatory burdens. Two broad issues need attention. One is the semantic task of providing a compositional model of the content of NJ and the meaning of normative sentences that can explain how the semantic properties of simple judgments and sentences can combine to determine the semantic properties of more complex judgments and sentences. This is largely the task of solving the so-called "Frege-Geach problem."[2] Much work has been devoted to this task. I will have nothing to add to it here.

[1] Here, I focus on NJ in general, rather than moral judgments specifically. Similarly, I focus exclusively on pure, rather than hybrid, expressivist views (according to which NJ are partly cognitive states (see Ridge and Köhler (2015)).
[2] See Schroeder (2008c).

But it's not the only issue that needs attention. There is also the psychological task of giving an account of the nature of NJ. Expressivism, after all, is a thesis about the mind as much as it is a thesis about the language we use to express it. One of its central claims is that these attitudes we call normative "beliefs" are not really what they seem. So what are they then? Expressivists need to provide an answer to this question. And here it isn't enough for expressivists to say that these are non-cognitive states, or to simply give a name to them (*e.g.* 'approval' or 'disapproval'). We need a rich-enough story of their nature. Such a story, moreover, must be able to do some explanatory work and account for central features we pre-theoretically take NJ to have.

Two of those features will be central here. The first is that NJ support relations of agreement and disagreement (from now on, I focus on disagreement, but everything I say applies to agreement). If, for example, you judge that Sam ought to have a beer and I judge he ought not, it's uncontroversial that we disagree. Any story of what NJ are should account for this. And here it is not enough to point to the formal apparatus developed as part of the semantic task mentioned above and say that we disagree because we hold this attitude of NJ, *whatever it may be*, towards contents that are rendered incompatible by the model. For there are attitudes we could take towards contents that are thus incompatible that would not land us in disagreement. Consider belief. If you believe that *p* and I believe that *not p*, we disagree. But we don't disagree if you doubt that *p* and I doubt that *not p*. So there must be something about belief—about the attitude itself—that explains why we disagree when we believe inconsistent contents. The same must be true of NJ, since we don't disagree if you doubt that Sam ought to have a beer and I doubt that he ought not.

The second feature is that some NJ are "agent-relative." I will explain this notion in more detail in §2.1, but the idea is that we often judge that there are reasons that pertain to individuals in particular, rather than to people in general. If, for example, you and I are heads of rival mafia families, then I might judge that you have a reason to get me assassinated. This is a reason that I attribute to you in particular, not to people in general. For starters, I don't think *I* have that reason. Much of our normative thought is, in this sense, agent-relative. Any proper story of NJ must account for such agent-relativity.

Unfortunately, expressivists have not really provided us with a story that can account for both features. This may not be due to neglect. On the contrary, I

think there are structural aspects that render the task rather daunting. Or so I will argue.

The guiding axis of my discussion will be the notion of an attitude's 'direction of fit' ('DF'), which I explain in §1. Expressivism is often identified by way of this very notion. Expressivism, it is often said, is a theory according to which NJ are mental states with a "world-to-mind" DF. In this, they would contrast with regular, descriptive beliefs, which have the opposite—"mind-to-world"—direction. Let me call states with a world-to-mind DF "conative," and states with the opposite DF "cognitive." Intentions and desires are paradigmatic conative states. Beliefs are paradigmatic cognitive states. The expressivist proposal, so understood, is that NJ are conative states of mind.

The first idea I will defend (in §2) is that NJ cannot be conative states. Conative attitudes have a distinctive causal profile—they are, *inter alia*, motivations to bring about their contents. No attitude with such a profile can play the role that NJ play in our minds. The reason takes the form of a dilemma. If NJ are conative states, then they are either attitudes *de se* (*i.e.* 'about oneself') or they are not. If they are not, then they will be unable to account for agent-relativity. If they are, then they will be unable to account for disagreement.[3] A theory of NJ should account for both. So NJ cannot be conative attitudes. Expressivists should deny that NJ have a world-to-mind DF.[4]

This doesn't mean they should claim that NJ have the opposite DF. But they could claim NJ are attitudes *without* a DF. These could still be non-cognitive states. Indeed, they could still be motivational. For instance, NJ could be affective attitudes, or, alternatively, *sui generis* non-cognitive, non-conative states.

[3] The problem I consider is not the "negation problem" identified by Unwin (1999) and popularized by Schroeder (2008b). See §2.3 below.

[4] I point to the difficulty expressivism has of making sense of both disagreement and agent-relativity in Núñez (2016, pp. 150–155). Since I was then assuming that expressivism was committed to the idea that NJ are conative states, and since I then thought that it was more important to make sense of disagreement than agent-relativity, I argued that the best version of the theory would have to render many agent-relative views incoherent. Ayars (2022) independently reaches a similar conclusion (although the expressivist view she proposes is significantly different). But, whereas I saw this implication as a genuine problem for expressivism, Ayars seems to see it as a virtue, since she seems to think such agent-relative views are incoherent anyway (see ft. 11). My considered view is that any metaethical view that implies this is unacceptable. I explain why in footnote 11.

Unfortunately, there are serious costs that come with this move. The second main idea I defend (in §3) is that, if NJ don't have a DF, then it's a mystery why they support relations of disagreement.

The take-home message is simple: expressivists owe us a satisfactory account of the nature of NJ. Without such an account, a dilemma looms. The structure of the paper aims to reflect it: NJ either have a world-to-mind DF or they don't. If they do (§2), then they are either *de se* attitudes or they are not. If they're not (§2.1), then they're unable to account for agent-relativity. If they are (§2.2), then they're unable to account for disagreement. If, on the other hand, they don't have a world-to-mind DF (§3), then it's a mystery how they support relations of disagreement. This is so whether we think of them as affective attitudes (§3.2), or as other *sui generis* non-cognitive, non-conative states (§3.3).

In the next section, I explain the basic psychological notions employed.

1. A basic psychological taxonomy

I will now explain the notion of an attitude's direction of fit, and then briefly explain what I understand conative and affective states to be.

The basic idea behind the DF metaphor is this: attitudes with a world-to-mind DF function to make the world fit their contents, whereas attitudes with a mind-to-world DF function to have contents that fit the world. Elisabeth Anscombe (1957 §32) illustrates this by pointing to the different roles played by a shopping list vs. the list used by a detective who is secretly tracking the consumer habits of a shopper. The function of the former is to make the contents of the shopping cart match the contents of the list. The function of the latter is to make the contents of the list match the contents of the cart. So, when the shopper notices that an item on his list is not in his cart, he reaches for the item and puts it in the cart, whereas when the detective notices this mismatch, he erases the item from his list. The idea, then, is that attitudes with a world-to-mind DF function like shopping lists, whereas attitudes with a mind-to-world DF function as detectives' lists.

At face value, the metaphor suggests that attitudes with a world-to-mind DF are (in part) motivations to make the world match their contents.[5] I will assume this is so. I will assume that, if you hold an attitude of such kind towards a

[5] I am thinking of these as "*pro*-attitudes." There might be reasons to postulate the existence of conative "*con*-attitudes," but I will ignore this complication here.

proposition *p* (or a property *F*) you are *thereby* (not necessarily in an overriding or conduct-controlling way) motivated to act in ways that (according to your beliefs) would be somehow conducive to the obtainment of *p* (or your instantiating of *F*). I call such attitudes 'conative.'

As I mentioned, expressivism is often described as postulating that NJ have a world-to-mind DF. I suppose that, mostly, when people say this, they mean to say that according to expressivism NJ are non-cognitive. But the claim is potentially misleading, because not all non-cognitive states have a world-to-mind DF.

Among the non-cognitive states, it's usual to distinguish between the conative and the affective. This distinction corresponds to the traditional tripartite picture of the mind, which divides mental states into the cognitive, the conative, and the affective.[6] Conative attitudes are typically understood as, *inter alia*, motivations to bring about their objects; affective attitudes as, *inter alia*, feelings (*i.e.* subjective experiences) or dispositions to experience certain feelings under suitable circumstances. I will assume that this is so.[7]

A few clarifications: first, it might seem that the difference between the conative and the affective is that, whereas the former are motivational states, the latter are not. But affective states may be motivational too, some perhaps essentially so. Pain, fear, disgust, anger, hate, etc., are all affective states. They all are, *inter alia*, feelings or dispositions to feel. But perhaps they all partly consist also of motivations to *e.g.*, to recoil from, escape, avoid, or hinder their sources.[8] The point, then, is not that conative attitudes are motivational and affective attitudes are not. It's rather that conative attitudes are motivations *with a specific aim*: to make the world match their contents. This is the idea contained in the DF metaphor. Affective attitudes do not have this aim.

Second, nothing in this picture precludes the existence of hybrid states composed of two or more attitudes of different kinds. One type of hybrid state that will concern us (§3.3) is that of the part conative, part affective. To keep

[6] For a historical overview in cognitive psychology, see Hilgard (1980).

[7] This isn't uncontroversial. Some maintain that conative or affective states reduce to value judgments, where these are explicitly characterized as cognitive states. Since an expressivist could not appeal to states so-conceived as explanatory resources, I will ignore such views.

[8] One could debate whether such motivations are partly constitutive of the state, or mere characteristic effects of it. We needn't take a position on this issue here.

things clear, though, I refer to such states as "hybrid," and reserve the labels "cognitive," "conative" and "affective," to pure specimens of these kinds.

Finally, I don't assume this traditional partition is exhaustive. There may be states that are neither cognitive, nor conative, nor affective, nor any combination thereof. I consider some of them in §3.3.

Non-cognitive attitudes, then, go beyond the conative. I take it that the claim essential to expressivism is that NJ are non-cognitive, not that they are specifically conative. Expressivists could maintain that NJ are affective or, alternatively, *sui generis* non-cognitive states. I will come back to these possibilities in §3. For the moment, however, I take the familiar description of expressivism at face value, and explore the prospects of specifically conative forms of expressivism.

2. Conative expressivism ("conativism")

I will now argue that expressivists should deny that NJ are conative states of mind. The reason is that conative attitudes are either unable to account for agent-relative NJ, or unable to account for disagreement.

2.1 First Horn: agent-centered NJ

Consider an expressivist view according to which for a judge J to judge that a subject S ought to φ is for J to have some kind of conative attitude in favor of S's φ-ing. Here is a generic version of the view: for J to judge that S ought to φ is for J to *want* that S φs, where I use "want" merely as a placeholder for whatever conative attitude (desires, intentions, etc.) the expressivist singles out as constituting NJ.

As different theorists have noted, there is an immediate problem with any view of this form.[9] Consider the following case:

> Jack and Sally are playing a match of chess. Jack wants nothing more than to win. At the same time, he is of the opinion that, when playing chess, one ought to play the move that maximizes one's chances of winning. Jack believes Sally would win if she castles. Because of this, Jack judges Sally ought to castle. However, he does not (in any sense

[9] E.g., Dreier (1996), Gibbard (2003), Ridge (2014).

> or to any degree) want her to do so. In fact, he very much wants her not to.

This scenario seems possible. That is, it seems possible that Jack might judge that Sally ought to castle and yet in no way want her to do so. This version of expressivism, however, is forced to say that it is metaphysically impossible. Since, according to it, J's judgment that S ought to φ is constituted by J's wanting that S φs, it cannot be the case both that Jack judges that Sally ought to castle and that he in no way wants that she does.

Jack's judgment that Sally ought to castle is bears a specific kind of relativity. Jack judges that Sally has a reason—indeed, decisive reason—to castle. But he doesn't think anyone else has *that* reason. Put in John Broome's (2013) terminology, Jack attributes "ownership" of that reason to Sally; he attributes an "owned" decisive reason—*i.e.*, an "owned ought"—to her. In the slightly odd but illuminating way Broome has of putting this idea, what Jack judges is that *Sally ought that Sally castles*. This construction helps distinguish what Jack judges about Sally from what he judges about another player (call him "Chester") who is a known cheater. Jack judges Chester ought to get kicked out of the chess tournament. But he does not judge that Chester ought that Chester gets kicked out. He thinks it's the tournament officials who ought that Chester gets kicked out. Similarly, when Jack judges that Sally ought to castle, he does not judge that he himself, or anybody else other than Sally, ought that she castles. As he sees it, it falls on Sally, and no one else, that she does so.

Following Broome, I will say that reasons and oughts that are owned by some, but not all agents—and the judgments that attribute them—are "agent-relative"; whereas reasons and oughts that are owned by all agents—and the judgments that attribute them—are "agent-neutral."[10]

Jack's judgment that Sally ought to castle, then, is agent-relative. This allows Jack to coherently judge that, while Sally ought to castle, he ought to prevent that she does. These judgments are consistent because they attribute owned oughts to different agents. It's not that Jack judges that it ought to be the case both that Sally castles and that he prevents that she does; or that there is anyone who ought both that Sally castles and that she does not. What Jack judges is that, although Sally ought that Sally castles, he ought that she does not. There is nothing rationally problematic or incoherent about holding both judgments.

[10] See also Ridge (2011).

Agent-relativity is a pervasive feature of normative thought. The normative egoist who thinks everyone ought to do what is best for themselves; the decision theorist who thinks agents ought to maximize their own expected utility; the deontologist who thinks, for instance, that parents ought to do what is best for their own children, they all hold agent-relative normative views. Many of these views are plausible. All of them are coherent.[11] And all of them are possible views someone like Jack could hold, even though he's an egoist who wants nobody's good but his own.

The view under consideration, however, can't make sense of this phenomenon, because it can't make sense of an agent who judges someone else ought to φ while not wanting, in any sense, that they do so. Given the prevalence of agent-relativity in our normative thinking, this is a fatal problem for the view.

It is useful to see where the source of the problem may lie. One central motivation for expressivism is explaining judgment internalism—or why it is that

[11] Medlin (1957) assumes an expressivist view of the aforementioned kind and argues on its basis for the incoherence of egoism. But, as Dreier (1996) correctly notes, (1) this implication extends to any agent-relative view that allows—as egoism does—for situations in which each of two agents ought to do something that they can't *both* do (call these cases of "conflicting obligations"); and (2) since many of such views are plausible, and since they are anyway coherent, this really constitutes an argument against the metanormative theory, not against those normative views. Ayars (2022, p. 59) and Ayars and Rosen (2021, p. 1036) suggest there is independent reason to think those views are incoherent. But the reason they present is simply that they find it "odd" to say, of two people who are competing against each other, that each should win. This isn't a compelling reason. The oddity (if there is one) is peculiar to competitive contexts, and can be explained because saying this: (a) suggests that it's *optional* for each of them to win—something that is false in competitive contexts; and (b) is typically completely uninformative. In other contexts, however, attribution of conflicting obligations sounds straightforwardly true. Consider: you see a child drowning. You can easily save him. So you ought to save him. Suppose there's someone else around who could also easily save him. Then they also ought to save him. This is so even if (for whatever reason) you can't *both* do so. Given that (at the time) each of you can, it's true (then) that each of you should. Ayars (2022) and Ayars and Rosen (2021) suggest that, in cases of seemingly conflicting obligations, each agent only ought to try. This isn't plausible. Grant, for argument's sake, that you only ought to try when factors beyond your control would prevent you from succeeding. Still, when success is guaranteed, the restriction is unwarranted. To see this, suppose you're the only person around. Then nothing would prevent you from saving him. Clearly, you ought to save him then. Well, we can imagine this is true of *each* agent. Suppose neither of you will do anything to save him, because you're both narcissists who don't want to get your clothes dirty. Then nothing would prevent either of you. So you each ought to. If so, then, *a fortiori*, it isn't incoherent to think so, and any metanormative view that implies this can be rejected. Anyway, the mere fact that it sounds plausible to say this here is enough to block any argument for the general incoherence of such views that relies solely on the intuition that, other times, it sounds odd.

NJ are so closely tied to motivation. Roughly, the idea here is that, if J judges that she ought to φ, then she is, at least to some degree, and in the absence of irrationality or psychological pathology, motivated to φ. Why would this be?

Conativists have an easy answer: she would be motivated to φ because to judge that you ought to φ *just is* to be (somehow) motivated to φ. This answer, however, works only if J's judgment is constituted by some conative attitude that would somehow favor J's φ-ing. If so, then it seems like the theory will have problems when the judgment is about someone other than J. For the natural idea would be that, if J's judgment that J ought to φ is constituted by J's wanting that J φs, then J's judgment that S ought to φ would be constituted by J's wanting that S φs. But this is not possible. Jack's judging that Sally ought to castle cannot consist in his wanting that Sally castles, for Jack wants nothing of this sort.

This version of expressivism has the following feature: that, when a judge J1 and a different judge J2 both judge that the same subject S ought to φ, the attitudes that constitute their judgments are about the same issue: S's φ-ing. For reasons that will become apparent, I will call any version of expressivism of this sort "expressivism *de aliis*" (meaning "about others"). The suggestion, then, is that expressivism *de aliis*, regardless of what precise shape it takes (and I will explore different shapes below), has a problem accounting for agent-relativity. Now on to the second horn.

2.2 Second horn: disagreement

Because of the previous worry, some expressivists have formulated versions of the theory according to which for J to judge that S ought to φ is instead for J to have some kind of conative attitude in favor of J's *own* φ-ing, in case of being in a situation that is relevantly like S's.

I will call this kind of expressivism "*de se*" (meaning "about oneself"), because it construes NJ as attitudes that are about oneself and one's own actions (or responses, more generally). In contrast, I call the previous kind of expressivism "*de aliis*," because (apart from first-person NJ), it construes NJ as attitudes that are about other people and their actions.

The generic version of expressivism *de se*, then, says that for J to judge that S ought to φ is for J to want *herself* to φ in case of being in a situation relevantly like S's—where "want" is again a place-holder for the relevant conative state.

Unfortunately, there is also an immediate problem with any view of this form. Consider Jack and Sally again. Jack judges that Sally ought to castle. But suppose now that there is another person, Judy, watching the game. For whatever reason, Judy is of the opinion that one ought to let other people win at games. So, although she may agree with Jack that castling would be Sally's best chess move, she thinks Sally ought not play it—she judges Sally ought not castle. I take it as a datum that Jack and Judy *thereby* disagree. Part of what makes this possible, however, is that their respective judgments concern the exact same issue: whether Sally ought to castle. If they did not concern the same issue, then they wouldn't disagree.

The problem with expressivism *de se*, then, is that it construes the relevant attitudes as concerning different issues: Jack's attitude concerns himself and his own actions, Judy's concerns herself and her own actions. Jack wants *himself* to castle in case of being in Sally's situation. Judy wants *herself* not to castle if in that situation ('C' from now on). Because their attitudes concern different issues, there is no recognizable sense in which they thereby disagree.

In saying this, I am assuming, for the sake of argument, that there can be disagreement "in attitude"—to use C. L. Stevenson's (1944) famous phrase. That is, I am assuming that two agents can disagree, not in virtue of what they believe, but simply in virtue of what they (in some sense) want.[12] To use Stevenson's own example (1944, p. 3), two people who want to go to dinner together might disagree about where to go, if one wants that they go to a place with music and the other wants that they go to a place without. My point, then, is that, granting that disagreement in attitude is possible, there is no recognizable sense in which Jack and Judy disagree.

I think this is the intuitive diagnosis. It is amply reflected in philosophical treatments of these issues.[13] Simon Blackburn nicely illustrates this idea by noting that we do not disagree if you intend to prohibit smoking in your house and I intend to allow it in mine, even though we *would* disagree if we are married and you intend that we prohibit it and I intend that we don't (1998, p. 69).

This diagnosis also mirrors what we would say in parallel cases involving belief. Suppose Jack believes he would castle in C and Judy believes she would not.

[12] It might matter in what sense exactly they want this (*e.g.*, it seems more plausible that intentions ground disagreement than that desires do). I ignore this complication here.

[13] Among them, Blackburn (1998), Dreier (2009), Ridge (2014), Marques and García-Carpintero (2014), Marques (2016), Worsnip (2019), Ayars (2022).

Clearly, they would not thereby disagree. Such beliefs are perfectly consistent, because they concern different issues: what Jack would do *vs*. what Judy would do—both in C, yes, but still what different agents would do in that situation. I don't see why things would be different when we are considering what Jack and Judy want rather than what they believe. So I take it as an indication that they do not disagree "in attitude" that they would not disagree in belief in the parallel scenario.

Here's another way to see this: there are two main accounts of disagreement in the literature. According to one of them—which John MacFarlane (2014 ch. 6) calls "preclusion of joint satisfaction"—two people disagree, roughly, just if they hold attitudes (presumably, of the same kind) such that, necessarily, were one of them to be satisfied (*e.g.* true, in the case of beliefs; realized, in the case of intentions), then the other one would not. According to the other—which MacFarlane (2014 ch. 6) calls "non-cotenability" and Worsnip (2019) "interpersonal incoherence"—two people disagree, roughly, just if they hold attitudes that are "non-cotenable," that is, attitudes that cannot be coherently held by a single individual at a single time.[14]

I doubt that either preclusion of joint satisfaction or non-cotenability are sufficient for disagreement.[15] But it does seem plausible that something roughly like them is necessary. This is because both reflect the fact that disagreement involves a kind of conflict between the commitments we assume in virtue of the attitudes we hold. But if either condition is necessary, then Jack and Judy do not disagree. Their attitudes do not preclude each other's satisfaction, and they are cotenable, since it's obviously coherent for Jack to want to castle in C, while at the same time wanting that Judy does not.

Now, there's a familiar formal trick that has been proposed as a solution to this problem. It consists in not representing the relevant agent in our model of the content of these attitudes. The proposal follows David Lewis' treatment of attitudes *de se* (1979). Lewis suggests that we can think of *de se* beliefs as the self-ascription of properties. Analogously—as Dreier (1996) and Gibbard (2003) suggest—we could think of the state that constitutes NJ as something like the self-*prescription* of a property. If so, then the idea would be that Jack self-prescribes having the property of *castling if in C*, while Judy self-prescribes

[14] See also, Marques (2016).
[15] See MacFarlane (2014, p. 21 n.124).

having the property of *not castling if in C*. What we've done with this move is to pull the agent out of our model of the content of these attitudes. And this may give the impression that the two attitudes suddenly do concern the same issue. Have we thereby secured disagreement?

Obviously not. Consider again the corresponding beliefs. Jack believes he would castle if in C. This is a *de se* belief. Following Lewis, we could say that Jack self-ascribes the property of *castling if in C*. Judy, in turn, self-ascribes the property of *not castling if in C*. We've taken the relevant agents out of our model of the content of these attitudes. Have we thereby secured disagreement? Obviously not. These beliefs are perfectly consistent; they are co-satisfiable, and, to the extent that we want to say that they are non-cotenable (because, the thought would be, it is incoherent to self-ascribe both *F* and *not F*), then we must say that non-cotenability is insufficient for disagreement (for, obviously, two people do not disagree if one self-ascribes *F* and the other self-ascribes *not F*).

Allan Gibbard (2003) tries to get around this problem by going a step further: it's not just that Jack self-prescribes (as he puts it, 'plans') the property ('the plan') of castling if in Sally's situation. He self-prescribes the property of castling in case of *being Sally* in her exact situation. Similarly with Judy. Because Jack and Judy self-prescribe incompatible properties in case of being *the very same person in the very same situation*, Gibbard thinks that we do have the "stability of subject matter" that he himself takes to be necessary for disagreement (2003, p. 66).

Unfortunately, this won't do. Suppose Jack believes he would castle *were he Sally in C*, and Judy believes she would not *were she Sally in C*. Do they thereby disagree? Not intuitively, and not according to *any* account of the meaning of the statements that would express such beliefs (Kocurek, 2018). Such accounts render the claims consistent, precisely because they are about different issues.[16]

The lesson is simple: we don't secure disagreement simply by deciding to limit the information that gets represented in our model of the content of these states. We can choose not to represent the fact that these are attitudes about oneself in our model of their contents. But then we will need to keep track of this fact in our account of their nature: their *self*-ascribing or *self*-prescribing nature. Since there are two different people doing the relevant self-ascriptions or self-

[16] If they don't render them trivially consistent. See Kocurek (2018).

prescriptions, we won't get disagreement even if they are self-ascribing or self-prescribing incompatible properties.

This illustrates why expressivism *de se* has a problem with disagreement. Now let me explain why the dilemma generalizes to any conative version of expressivism.

2.3 Why the dilemma generalizes

The distinguishing feature between expressivism *de aliis* and expressivism *de se* is that, when J1 judges S ought to φ and J2 judges S ought not φ, the former construes the attitudes that constitute these judgments as being about the same issue, while the latter does not. Regardless of the peculiarities that different conativist theories might take, they all must take one of these two shapes. If a particular theory says the attitudes concern the same issue, it will have problems with agent-relativity. If it says they concern different issues, it will have problems with disagreement.

To illustrate, consider some of the different shapes a conativist theory could take. I continue to leave it open which conative attitude we are dealing with, because it does not really matter which one we pick. What matters is what content such attitude takes. There are two main variables we could tinker with when it comes to the content: (a) the agent-variable (so far I've considered only S and J); and, (b) the response-variable (so far I've considered only φ-ing). I explore each in turn.

(a) Tinkering with the agent-variable

There are only two plausible options here. First, expressivists could render J's judgment that S ought to φ as J wanting, not that S or J φ, but rather that anyone who is relevantly like S φs. R. M. Hare seems to hold a view of this form.[17] The issue is complicated because Hare does not say much about the attitude that constitutes NJ. His focus is normative language, not normative thought. He maintains that normative claims should be understood as universal prescriptions. But he sometimes seems to suggest that, just as the attitude that corresponds to the sincere utterance of a statement is a belief, the attitude that corresponds to the sincere utterance of a prescription is a preference for people to act in the

[17] Dreier (1996) interprets Hare in this way.

manner prescribed.[18] If so, then if—as Hare maintains—normative claims prescribe how every agent is to act if in a given situation, then NJ would be constituted by a preference that every agent acts in that way if in that situation.

Suppose this is so. This is an attractive theory. For one thing, it would avoid the problem with disagreement identified above. When J1 judges S ought to φ and J2 judges S ought not φ, their attitudes would be about the same issue: whether everyone in S's circumstances is to φ. The strategy has other virtues.[19] But it is a version of expressivism *de aliis*, and, as such, it has the same problem with agent-relativity explored above.[20] Jack judges that Sally ought to castle, but he does not want (prefer) that every agent castles if in Sally's situation. For one thing, he does not want (prefer) Sally to do so. Nor (we can stipulate) does he want (prefer) anybody else to do so.[21]

The second way of tinkering with the agent variable would be to render Jack's judgment as his wanting, not that *everyone*, but that people *in general* (or the generic 'one') castle in that situation. Appealing to generics has its advantages, since generics allow for exceptions.[22] This would allow Jack to judge that Sally ought to castle even when he in no way wants that she does. Unfortunately, the proposal does not really avoid the problem, since Jack doesn't even want people *in general* to castle in that situation.

As far as I can tell, there is no other way of tinkering with the agent-variable that has any plausibility, so let me turn to the other strategy.

(b) Tinkering with the response-variable

We could instead construe J's judgment that S ought to φ as J wanting that someone (be it S, J, everyone, people, etc.) ψs, where ψ-ing would be some response that is somehow conducive to (someone's) φ-ing.

---

[18] See his (1952, p. 20) and his (1981, pp. 22, 91).

[19] I explore some of them in Núñez (2016).

[20] Dreier makes this point in his (1996).

[21] It may be tempting to read Gibbard in this way, as claiming that J's judgment that S ought to φ amounts to J's planning that anyone in S's circumstances φs. But this wouldn't work. If you plan that everyone in C φs, you are disposed to intend the means necessary for, and not intend anything inconsistent with, everyone's φ-ing in C (see Bratman (1987)). Jack is not so disposed. Gibbard seems perfectly aware of this problem (2003, pp. 68-69). This is why he appeals to attitudes *de se*: plans for what to do oneself.

[22] See Leslie and Lerner (2016).

Consider first some possible versions of expressivism *de aliis* according to which J's judgment that S ought to φ amounts to J wanting that S (or everyone, or people):

- *wants* to φ, or
- *rules out* not φ-ing as a possible way of acting, or
- *blames* (herself/everyone?) for not φ-ing, etc…

The possibilities here are endless. But it doesn't really matter which ψ we choose. As long as these are versions of expressivism *de aliis*, all of them will face the same problem. Jack judges Sally ought to castle, but he does not want that Sally (everyone/people) wants to castle, or that she (they) rule it out, or that she (they) blames (herself?/everyone?) for not doing so, etc.

We could instead have a version of expressivism *de se* that tinkers with the response-variable. We could say that J's judgment that S ought to φ is constituted by J's wanting that J *herself*:

- *wants* that S (J/everyone/people) φs, or
- *blames* S (J/everyone/people) for not φ-ing in S's situation, etc…

But this gets us nowhere. It doesn't matter which ψ we select. If J1's attitude is about J1's own ψ-ing, then J2's attitude will be about J2's ψ-ing, and we won't have the "stability of subject matter" that is necessary for disagreement.

To illustrate, consider Mark Schroeder's (2008a) favored version of expressivism, the most developed version that pursues this strategy. Schroeder uses "being for" as I am using "want," that is, as a placeholder for whatever conative attitude the expressivist eventually picks.[23] Furthermore, following Gibbard (1990), he suggests (again, as a placeholder) that the relevant ψ could be *blaming for*. The proposal, then, is that we can understand J's judgment that S ought to φ as J's being for blaming S for not φ-ing.

This is an attractive view. Among other things, it allows Schroeder to explain what J's judging that it's not the case that S ought to φ amounts to. It amounts to J being for not blaming S for not φ-ing. This solves what Schroeder—following Unwin (1999)—calls "the negation problem" for expressivism: the

[23] Since he describes the attitude as having a world-to-mind DF (2008a, p. 92).

problem of explaining what external negation could amount to under an expressivist view.[24] This is a major feat.

Despite its virtues, it does not avoid the current dilemma. Consider: when J judges that S ought to φ, does this mean that J is for S (everyone/people) blaming (herself/others) for not φ-ing? This would be a *de aliis* version of Schroeder's view. Or does it mean that J is for J blaming (herself/others) for not φ-ing? This would be a *de se* version of Schroeder's view.

I take Schroder to favor the latter idea.[25] But, either way we go, familiar problems await. If we take the first route, we'll have problems with agent-relativity. Jack judges Sally ought to castle, but he is not for Sally (or anyone else's) blaming anyone for not castling. Alternatively, if we take the second route, we'll have problems with disagreement. Jack would be for he himself blaming for not castling in C. Judy, in turn, would be for she herself blaming for castling in C. These attitudes are about different issues, so Jack and Judy do not disagree in virtue of holding them. Either way we go, we have a problem. And that is, precisely, the moral of the story.

2.4 Taking stock

If NJ are conative states, then, in order to account for disagreement between J1 and J2, their attitudes must be about the same issue. However, to allow for agent-relativity, they must not be about the same issue. Such attitudes must either be or not be about the same issue. So, if NJ are conative attitudes, they will be either unable to account for disagreement, or unable to account for agent-relativity.[26]

NJ, then, cannot have a world-to-mind DF. This would be fatal for expressivism if not having such a DF implied having the opposite one. But it does not. Not all attitudes have it as part of their function to either have contents

[24] Schroeder (2008b).

[25] See Schroeder (2008a, p. 58).

[26] Ayars (2022) suggests this problem can be avoided by identifying NJ with an attitude she calls "decision" (*i.e.* J's judgement that S ought to φ is J's decision that S is to φ). Decisions sometimes constitute (or give rise to) motivations to promote their objects, but other times they don't. They typically do when one decides what to do oneself. They typically don't when one decides what someone else is to do (p. 44). So decisions are—in my sense—sometimes conative and sometimes not. This to me sounds more like a description of the problem than a solution to it, since decisions seem to play no causal role in the judge's mind when they aren't conative states. Regardless, the view explicitly—even gladly—renders egoism and similar agent-relative normative views incoherent (p. 59). So it doesn't avoid the first horn.

that fit the world or to make the world fit their contents. Expressivists could point to such attitudes. I turn to this issue now.

3. Non-conative forms of expressivism ("non-conativism")

Expressivists could claim that NJ are affective states. Alternatively, they could claim that they are *sui generis* non-cognitive states, be they reducible or irreducible to other, independently recognizable attitudes. I consider each of these paths in turn, and argue that it remains mysterious why such states would ground disagreement. I present the guiding reason for my skepticism in §3.1, and then turn to consider affective attitudes in §3.2, and *sui generis* states in §3.3.

3.1 Disagreement between attitudes without a DF

Here's a natural thought about disagreement: that it is a matter of adopting conflicting positions on a certain issue, or of giving conflicting answers to a certain question (Stroud, 2019). This explains why it's intuitive to think that there is disagreement between beliefs and disagreement between intentions. After all, as different theorists have argued, it's natural to think of belief and intention as states that consist of being somehow settled or decided on an answer to a certain question (Hieronymi, 2009). When you believe that *p*, you take a stance on the theoretical question whether *p*: you are, in a sense, *cognitively committed* to it being the case that *p*. When you intend that *p*, you take a stance on the practical question whether *p* is (as I shall put it, to signal that this is a practical rather than predictive question) *to be*: you are, in a sense, *conatively committed* to (roughly) making it be the case that *p* (Brandom, 1998; Bratman, 1987). Since it's natural to think of belief and intention as comprising answers to such questions, and since such answers can conflict in the familiar sense that they can't both be true, or can't both be implemented, it's easy to locate such cases within the coordinates of what we naturally think of as disagreements.

However, part of what allows us to think of these attitudes as comprising answers to such questions, and so as sustaining relations of disagreement, is that they are in the business of either having contents that match the world, or of making the world match their contents. Because of this, there is something definite an agent commits to by holding them: she commits to the world being, or to making it be, a certain way.

This isn't so with attitudes without a DF. Since such attitudes are neither in the business of having contents that match the world, nor of making the world match their contents, there is nothing definite an agent commits to by holding them. In any case, she does not commit to the world being, or to making it be, any definite way. Because she doesn't, it's difficult to think of such attitudes as comprising an answer to a question (what question could that be?), and so to understand them as grounding disagreement in this natural sense.[27]

Put differently: attitudes without a DF neither preclude each other's satisfaction, nor seem non-cotenable. They trivially don't preclude each other's satisfaction, because they don't have satisfaction conditions. They don't, because such conditions are just those in which the fit obtains. And they don't seem non-cotenable, because incoherence is typically understood in terms of the conflicting commitments our attitudes carry, and there doesn't seem to be anything we commit to by holding such states that would ground that conflict. So, if either preclusion of joint satisfaction or non-cotenability are necessary for disagreement, attitudes without a DF don't ground disagreement at all. Or so I will suggest.

I will now illustrate these ideas in relation to affective attitudes, and then, in §3.3, *sui generis* states.

3.2 Affective expressivism ("affectivism")

I will now argue that it's doubtful whether there is disagreement between affective states. Before I do, however, I should note that this issue is controversial in a way that it just isn't controversial whether there is disagreement between cognitive and—crucially—conative states.

Suppose you love cooking and hate doing the dishes, whereas I'm exactly the other way around. Do we disagree? It's not clear that we do. We like and dislike opposite things, surely, but such differences might form the basis of a rather harmonious agreement. Traditional wisdom would seem to support this skepticism. As the saying goes: "*de gustibus non est disputandum*" ("in matters of taste, there can be no dispute"). Different theorists concur: "Suppose I like ice cream and you don't"—says Michael Ridge—"So far, we have a difference, but no obvious disagreement." (2014, p. 180). Even among theorists that

[27] I don't mean to suggest that having a DF is sufficient for sustaining disagreement. But I do think it's likely necessary.

countenance the possibility of affective disagreement, there is often recognition that the sense in which such differences really amount to disagreements is "rather thin" (MacFarlane, 2014, p. 123) and admittedly "less clear" (Worsnip, 2019).

This is important because it shows why "affectivists" bear an explanatory burden that neither cognitivists, nor—crucially—conativists do. The latter two point to states most of us pre-theoretically recognize as grounding disagreement. (This is where the dialectical force of Stevenson's original case resides: most of us recognize we can disagree about what to do, where this isn't a disagreement in belief). Because of this, they don't bear the burden of explaining why we disagree in virtue of such states. In contrast, affectivists point to attitudes about which there just isn't consensus whether they ground disagreement or not. Hence, it's not unreasonable to ask why we should think such states ground disagreement.[28]

Anyway, suppose that NJ are affective states. Anger or disgust, for instance, may prove to be attractive candidates.[29] Here is a generic version of the view: for J to judge that S ought to φ is for J to be somehow *repelled* by (the possible state that consists of) S's not φ-ing, where "repelled" stands for whatever affective attitude (disgust, anger, distress, etc.) the expressivist signals as constituting NJ. If so, then, presumably, for J2 to judge that S ought not φ would be for J2 to be *repelled* by S's φ-ing. What could J1 and J2 disagree about in virtue of such states?

If you believe we will go to a place with music and I believe we will not, we disagree about what we'll do. And if you intend that we go and I intend that we don't, we disagree about what (we are) to do. But if you are repelled by our not going and I am repelled by our going, what could we thereby disagree about?

We could not disagree about what is the case. In being repelled by our not going, you do not thereby take a stance on whether anything is the case. Among other things, you do not take a stance on whether we will go. Of course, you may be repelled by that possibility because you believe that something is the case, but that is a different matter. The point is that, in being repelled by our not going, you do not thereby take a stance on whether *p*, for any *p*.

It might be tempting to think that you take a stance on whether that possibility is, in a normative or evaluative sense, *repulsive*, and that this might be the issue about which we disagree: we disagree about whether our going would be repulsive. But this is problematic for several reasons.

---

[28] See Dreier (2006, p. 220)

[29] On the intimate connection between such attitudes and specifically moral judgments, see Haidt, et al. (1997), Rozin, et al. (1999), Prinz (2006).

First, it seems false. It seems you might be repelled by something you do not judge to be repulsive (you might be going to therapy to correct this). Second, even if true, it would be circular in the present dialectic to appeal to normative or evaluative judgments as explanatory resources. So, even if it were true that in being repelled you judge it to be repulsive (because, as an expressivist treatment of these judgments would go, to *judge* repulsive is to *be* repulsed) we can't, in the present dialectic, explain the alleged disagreement in repulsion in terms of a disagreement in normative or evaluative judgments. Third, because, even ignoring the two previous points, the corresponding judgments would not explain the disagreement, since we don't disagree if you judge that it would be repulsive for us not to go and I judge it would be repulsive for us to do so. These judgments are consistent, since it's perfectly possible for both alternatives to be, in fact, repulsive.

On the other hand, we could not disagree about what is to be the case. In being repelled by our not going, you do not thereby take a stance on whether you, or we, are to go, or do anything at all. Naturally, your repulsion may lead you to intend that we go. But this, again, is a different issue. The point is that, in being repelled by our not going, you do not thereby take a stance on whether S is to φ, for any S and any φ.

So we can't say that we disagree about what we *will* do or are *to* do. In fact, we can stipulate that we agree on both counts: we both believe we will go and we both intend that we go. Still, you are repelled by our not going and I am repelled by our doing so. What then do we disagree about?

The worry, naturally, is that, if there is no recognizable issue about which we disagree, then there is no recognizable sense in which we disagree. We could emphasize this by going further and stipulating that we are both repelled by *both* possibilities. Since both possibilities may be genuinely repulsive, such reactions may be appropriate and, crucially, perfectly coherent. But then, if we agree both about what we'll do and about what to do, and, moreover, have the same affective reactions to all relevant possibilities, what then do we disagree about?

This is crucial because it shows that disagreement between such states is not only dubious; it's actually non-existent. This is because, if repulsion towards *p* and repulsion towards *not p* grounded disagreement, then there would be a sense (no doubt, metaphorical) in which we could be said to "disagree with ourselves" when we are each repelled by both possibilities. However, there is no sense (metaphorical or otherwise) in which we disagree with ourselves in such

situations: we are in no way incoherent or at odds with ourselves—such states are perfectly cotenable, and they trivially don't preclude each other's satisfaction. This shows that they don't conflict in the way required for disagreement. Indeed, it shows that they don't conflict at all.

What goes for repulsion, moreover, seems to go for affective attitudes in general: there's nothing incoherent about holding the same affective attitude towards inconsistent (or otherwise incompatible) contents. This is obvious with positively valanced affective states. It's not incoherent to enjoy going out and enjoy staying in, or to be happy with the prospects of both rain and sunshine. Having such sentiments does not make you incoherent, it makes you easygoing and self-possessed. But if instead you dislike, dread, are angered, saddened or disgusted by all those possibilities, then you're a bore and a grouch, but you needn't be incoherent (Baker & Woods, 2015, p. 409).

Affective attitudes, then, are perfectly coherent in cases where the judgments they would supposedly constitute are clearly not. This means that, quite independently of worries having to do with disagreement, the former cannot constitute the latter.

Some theorists argue that, although it's not incoherent to hold the same affective attitude towards inconsistent (or incompatible) contents, it is incoherent to hold "opposite" affective attitudes towards the same content (Baker & Woods, 2015). For example, it would be incoherent to both like and be disgusted by licorice, or to both love and hate Bob (MacFarlane, 2014, pp. 122-123). If so, then affectivists might hope to reduce contrary judgments to opposite attitudes. For example, *repulsion* at S's φ-ing for the judgment that S ought not φ, and *attraction* to S's φ-ing for the judgment that S ought to φ. Unfortunately, this won't do, for two reasons.

First, it's unclear that it's necessarily incoherent to hold opposite affective attitudes towards the same content. As the literature on "benign masochism" (Rozin et al., 2013; Strohminger, 2014) or "hedonic ambivalence" (Strohl, 2019) suggests, people commonly enjoy things that disgust, scare, sadden, anger or hurt them. The phenomenon is pervasive (thus horror movies and roller-coasters). It's implausible that such behavior necessarily reflects irrationality. In any case, theorists that study this phenomenon do not frame it in this way. Similarly, the literature on emotional ambivalence suggests you might coherently experience

opposite emotions towards the same object.[30] Two reasons suggest this. First, both emotions may be warranted. You might be justified to feel both happy and sad about a friend getting a position you wanted (Greenspan, 1980), or about the death of a loved one who had been ill and suffering for a while (Maguire, 2017). More generally, it seems like the reasons (or facts) that warrant experiencing one emotion need not count as reasons against experiencing its opposite (Maguire, 2017). Second—and this, I think, is the explanatorily fundamental point—to experience an affection is not yet to take a stance on whether anything either is, or is to be, the case.[31] Because it isn't, you need not have assumed any conflicting commitments in virtue of experiencing opposite affections. As far as those affections go, it's still open to you what to think or do about them. Among other things, it's still open to you—as Harry Frankfurt (2004) has eloquently argued—to decide to "side" with one of them, and do so wholeheartedly. You may be completely resolved on all relevant practical questions even if you remain subject to conflicting emotional pulls. You might love and hate someone, or something, you are wholeheartedly resolved to love (Frankfurt, 2004, p. 91).

Naturally, opposite affections might give rise to inner turmoil. "*Odi et amo*" ("I hate and I love"), Catullus famously laments. He finds this tormenting, as it surely must be. But so is sticking to your diet when you'd really like to have that second piece of cake, and we don't consider this inner turmoil necessarily irrational.

The second reason why this strategy won't do is that, even if it turned out that it's necessarily incoherent to hold opposite affective attitudes towards the same content, it's still coherent to hold the same affective attitude towards inconsistent (or incompatible) contents. This means that we would need a non-*ad hoc* explanation for why, although J's repulsion at S's φ-ing constitutes J's judgment that S ought not φ, J's repulsion at S's not φ-ing would *not* constitute J's judgment that S ought to φ. I doubt such an explanation is possible. In any case, it hasn't been given. In its absence, however, we would have two different states that constitute the judgment that S ought not φ: repulsion at S φ-ing and attraction at S's not φ-ing. However, only one of them would be incoherent with the judgment that S ought to φ. This means, intra-personally, that it would only sometimes be incoherent to hold contrary NJ, and, inter-personally, that such

[30] See *e.g.* Greenspan (1980), Koch (1987), Swindell (2010), Maguire (2017), Cecchini (2021).
[31] Baker & Woods (2015, p. 407) suggest that "to like or dislike something […] is to take a stand on something." Unfortunately, they don't tell us what that something could be.

judgments would only sometimes ground disagreement; other times, they would not. This is obviously absurd.

The lesson is simple: the fact that having an affective attitude does not yet constitute a way of taking a position on whether anything either is, or is to be, the case, suggests that affective attitudes don't ground relations of disagreement. But, even granting, for the sake of argument, that some of them do (namely, opposite attitudes directed at the same content), it suggests that NJ cannot reduce to affective attitudes, because affective attitudes don't ground disagreement—or incoherence—in cases in which the judgments they would presumably constitute obviously do (namely, same attitudes directed at incompatible contents).

I now turn to *sui generis* non-cognitive attitudes.

3.3 *Sui generis* expressivism

Expressivists could claim that NJ are *sui generis* non-cognitive states, be they irreducible attitudes, identifiable simply by the dispositions that realize them, or hybrid states, made up of more basic and independently identifiable conative and/or affective states. (I ignore hybrid states that are part cognitive, because I'm focusing on pure rather than hybrid expressivist views).

Consider the former possibility first. The idea, for instance, could be that NJ are constituted by an attitude that—following Gibbard's (1990)—I'll call "norm-acceptance." Following the spirit of Gibbard's proposal, let me say that, for J to judge that S ought to φ is for J to accept a norm that requires anyone in S's circumstances ('C') to φ. And that for J to accept such a norm is for J to be disposed to:

1. φ herself were she in C,
2. avow, in appropriate conditions, a norm that requires φ-ing in C, and
3. experience certain feelings (say, disgust or anger) at the sight or thought of people not φ-ing in C.[32]

Alternatively, expressivists could point to a hybrid state made up of independently identifiable conative and/or affective components. We could also think of norm-acceptance in this way. In fact, we could say that the dispositions 1-3 mentioned above obtain precisely because norm-acceptance consists of the following conative and affective components:

[32] Gibbard (1990, pp. 71-75).

1*. an intention (or policy[33]) to φ oneself were one in C,
2*. a desire to avow, in relevant conditions, a norm that requires φ-ing in C, and
3*. a standing disgust at the thought or sight of people not φ-ing in C.

These, then, are some examples of the type of *sui generis* states expressivists could point to. The possibilities are obviously endless. But, for any attitude the expressivist picks, we can ask why two agents disagree when they hold inconsistent NJ. Why would J1 and J2 disagree if J1's judgment amounts to J1's acceptance of a norm that requires everyone to φ in C, and J2's judgment amounts to J2's acceptance of a norm that requires everyone to not φ in C?

It may seem that the answer here is evident: they disagree because they accept conflicting norms. But this appearance is illusory, a mere function of the name we've given to the attitude. We can, of course, grant that these norms conflict in the familiar sense that they can't both be followed. But we still need to know what it is about the attitude itself that explains why two people who accept conflicting norms disagree. All we know so far is that J1 and J2 have the dispositions, or attitudes, listed, *mutatis mutandis*, in 1-3 or 1*-3* above. That is all. The question is why people who are so disposed, or who hold such attitudes, disagree. And there is nothing obvious about *this* question.

For one thing, they would not disagree in virtue of any single one of the components of such a state, whether we think of them as 1-3 or 1*-3* above. They would not disagree in virtue of 1 or 1*. As we've seen, two agents do not disagree simply in virtue of one of them intending—and so, *a fortiori*, being disposed—to φ in C and the other intending—and so being disposed—to not φ in C.

Likewise, J1 and J2 would not disagree simply in virtue of 2 or 2*. Now, Michael Ridge (2014 ch. 6) maintains that two people A and B disagree about S's φ-ing in C "just in case in circumstances of honesty, full candor, and non-hypocrisy, A would advise φ-ing in C and B would advise ψ-ing in C, where φ-ing and ψ-ing are incompatible." (p. 187). If so, then, if avowing a norm is—as Ridge thinks—a way of advising, then two people who are disposed to avow conflicting norms would disagree. Ridge does not explain why this would be so. But he does argue that the account is extensionally adequate. I doubt that it is.

[33] See Bratman (1989).

Suppose you're fond of Susie and so are disposed to try to convince her to join you for dinner in case you run into her tonight. I'm also fond of Susie, and so I'm disposed to try to convince her to join me in case I run into her. Let's say Susie can't both dine with you and dine with me. On Ridge's account, it would turn out that we disagree. This isn't plausible. What do we disagree about? We can stipulate you don't think Susie ought to join you rather than me. We can stipulate you haven't given the issue any thought at all. The same is true of me. Then there would be nothing for us to disagree about.

We are left with components 3 and 3*. This is the affective component. I've argued that such states don't ground disagreement. So no individual component of 1-3 or 1*-3* accounts for disagreement.

One may think this shouldn't be of much concern. On the one hand, the overall state might ground disagreement even if none of its individual components does. On the other, this is just one way of construing the relevant attitude. Nothing prevents us from construing it differently.

However, it follows from what's been said already that neither the overall state, *whatever it may be*, nor any one of its components, *whatever they may be*, will account for disagreement in any familiar sense. This is because, given our concern with agent-relativity, neither the overall state that constitutes the judgment that S ought to φ, nor any one of its components, can be a conative state directed at anyone other than the judge herself φ-ing in those circumstances. Recall that NJ have to be such that Jack could judge that Sally ought to castle while wanting no one (except himself) to do so. More generally, NJ have to be such that an egoist could judge that everyone ought to do what is best for themselves while wanting no one but himself to do so.

Call this "the egoist constraint" on any possible account of the nature of NJ. The constraint is there to secure compatibility with agent-relativity. But it virtually ensures that disagreement between NJ won't be explained by the kind of clash that constitutes conative disagreement. Since, by hypothesis, the disagreement won't be explained by a cognitive disagreement either, neither the overall state, *whatever it may be*, nor any one of its components, *whatever they may be*, will account for disagreement in any familiar sense.

So, if it's norm-acceptance that constitutes NJ, then, given the egoist constraint, accepting a norm that requires everyone to φ in C cannot involve wanting anyone but oneself to φ in C. This means that, when J1 accepts that norm and J2 accepts the contrary, there won't be any agent (or group) A (not S,

not J1, not J2, not everyone, not even the generic 'one') such that J1 thereby wants A to φ and J2 thereby wants A to not φ. This means that, in accepting conflicting norms about what everyone is to do, they don't disagree about what anyone is to do. What, then, do they disagree about?[34]

Naturally, the expressivist may simply hold her ground at this point and say that what J1 and J2 disagree about is, precisely, what people ought to do, and that having the attitudes she identifies as NJ towards the contents her formal apparatus identifies as inconsistent is what such disagreement *fundamentally* consists in. This may be so. Nothing I've said would prevent it from being so. But we must be clear about the role that appeal to normative disagreement would be playing here, and it is not an explanatory role. In other words, if this is all the expressivist has to say at this point, it will remain a mystery why these attitudes support disagreement.[35]

4. Conclusion

Expressivists have tended to pay more attention to the semantic task of providing a formal apparatus to model the contents of NJ than to the psychological task of giving a proper account of its nature. There is nothing wrong with this if it stems from a healthy division of philosophical labour. But the labour must eventually be completed if expressivism is to discharge its own explanatory burdens.

Here, I have argued that there are considerable difficulties associated with this task. On the one hand, if NJ have a world-to-mind DF, they will be either unable to account for agent-relativity, or unable to account for disagreement. Since a proper theory of NJ must account for both, expressivists should deny that NJ have such a DF. On the other hand, however, if NJ do not have a DF, then it's a mystery why they sustain relations of disagreement. Of course, an account may be given that explains this. But it hasn't been given.

---

[34] Following Gibbard's (2003) motto, couldn't we say that they disagree, not about what S, one of them, both of them, everyone, or the generic *one*, is to do, but just about what to do *simpliciter*? But "what to do?" is not a complete question, and—as the standard syntax for infinitives suggests (see *e.g.* Radford (1988))—it cannot receive an answer until its "understood" or "covert" subject is specified. You can decide what *you* are to do, a mother can decide where *her child* is to go to school, and an urban planner can decide what *one* is to do in an emergency. Such decisions are possible because they have a psychological imprint: they lead agents to think and act in characteristic ways. What you cannot do is decide what to do *simpliciter*, because there is nothing such a decision would lead you to think or do.

[35] Dreier makes a similar point in (2006, p. 220).

Bibliography

Anscombe, G. E. M. (1957). *Intention* (Vol. 57). Harvard University Press.

Ayars, A. (2022). Deciding for Others: An Expressivist Theory of Normative Judgment. *Philosophy and Phenomenological Research*, *105*(1), 42-61.

Ayars, A., & Rosen, G. (2021). Noncognitivism and agent-centered norms. *Philosophical Studies*, *179*(4), 1019-1038.

Baker, D., & Woods, J. (2015). How Expressivists Can and Should Explain Inconsistency. *Ethics*, *125*(2), 391-424.

Blackburn, S. (1998). *Ruling Passions: A Theory of Practical Reasoning*. Oxford University Press UK.

Brandom, R. (1998). Action, Norms, and Practical Reasoning. *Philosophical Perspectives*, *12*, 127-139.

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (1989). Intention and Personal Policies. *Philosophical Perspectives*, *3*, 443-469.

Broome, J. (2013). *Rationality Through Reasoning*. Wiley-Blackwell.

Cecchini, D. (2021). Experiencing the Conflict: The Rationality of Ambivalence. *The Journal of Value Inquiry*.

Dreier, J. (1996). Accepting agent centered norms: A problem for non-cognitivists and a suggestion for solving it. *Australasian Journal of Philosophy*, *74*(3), 409–422.

Dreier, J. (2006). Negation for Expressivists: A Collection of Problems with a Suggestion for their Solution. *Oxford Studies in Metaethics*, *1*, 217-233.

Dreier, J. (2009). Relativism (and expressivism) and the problem of disagreement. *Philosophical Perspectives*, *23*(1), 79-110.

Frankfurt, H. G. (2004). *The Reasons of Love* (Vol. 74). Princeton University Press.

Gibbard, A. (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Vol. 41). Harvard University Press.

Gibbard, A. (2003). *Thinking How to Live*. Harvard University Press.

Greenspan, P. (1980). A Case of Mixed Feelings: Ambivalence and the Logic of Emotion. In A. O. Rorty (Ed.), *Explaining Emotions* (pp. 223--250). University of California Press.

Haidt, J., Rozin, P., Mccauley, C., & Imada, S. (1997). Body, Psyche, and Culture: The Relationship between Disgust and Morality. *Psychology and Developing Societies*, *9*(1), 107-131.

Hare, R. M. (1952). *The Language of Morals*. Oxford University Press.

Hare, R. M. (1981). *Moral Thinking: Its Levels, Method, and Point* (Vol. 93). Oxford University Press.
Hieronymi, P. (2009). Two kinds of agency. In L. O'Brien & M. Soteriou (Eds.), *Mental Action*. Oxford University Press.
Hilgard, E. R. (1980). The trilogy of mind: Cognition, affection, and conation. *Journal of the History of the Behavioral Sciences*, *16*(2), 107–117.
Koch, P. J. (1987). Emotional ambivalence. *Philosophy and Phenomenological Research*, *48*(2).
Kocurek, A. W. (2018). Counteridenticals. *The Philosophical Review*, *127*(3).
Leslie, S.-J., & Lerner, A. (2016). Generic Generalizations. *Stanford Encyclopedia of Philosophy*.
Lewis, D. (1979). Attitudes de dicto and de se. *Philosophical Review*, *88*(4), 513–543.
MacFarlane, J. (2014). *Assessment Sensitivity: Relative Truth and its Applications*. Oxford University Press.
Maguire, B. (2017). There Are No Reasons for Affective Attitudes. *Mind*, *127*(507), 779–805.
Marques, T. (2016). We can't have no satisfaction. *Filosofia Unisinos*, *17*(3).
Marques, T., & García-Carpintero, M. (2014). Disagreement about Taste: Commonality Presuppositions and Coordination. *Australasian Journal of Philosophy*, *92*(4), 701–723.
Núñez, C. (2016). *The will and normative judgment* [Doctoral Thesis, Stanford University]. Unpublished. http://purl.stanford.edu/pq483mk7065
Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, *9*(1), 29–43.
Radford, A. (1988). *Transformational grammar : a first course*. Cambridge University Press.
Ridge, M. (2011). Reasons for Action: Agent-Neutral vs. Agent-Relative. *The Stanford encyclopedia of philosophy*, 1–48.
Ridge, M. (2014). *Impassioned Belief*. Oxford University Press.
Ridge, M., & Köhler, S. (2015). Metaethical theories, hybrid. https://www.rep.routledge.com/articles/thematic/metaethical-theories-hybrid/v-1
Rozin, P., Guillot, L., Fincher, K., Rozin, A., & Tsukayama, E. (2013). Glad to be sad, and other examples of benign masochism. *Judgment and Decision Making*, *8*(4), 439–447.
Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). *The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three*

moral codes (community, autonomy, divinity) [doi:10.1037/0022-3514.76.4.574]. American Psychological Association.
Schroeder, M. (2008a). *Being For: Evaluating the Semantic Program of Expressivism* (Vol. 70). Oxford University Press.
Schroeder, M. (2008b). How Expressivists Can and Should Solve Their Problem with Negation. *Noûs*, *42*(4), 573-599.
Schroeder, M. (2008c). What is the Frege-Geach problem? *Philosophy Compass*, *3*(4), 703-720.
Stevenson, C. L. (1944). *Ethics and Language*. Yale University Press.
Strohl, M. (2019). Art and painful emotion. *Philosophy Compass*, *14*(1), e12558.
Strohminger, N. (2014). Disgust Talked About. *Philosophy Compass*, *9*(7), 478-493.
Stroud, S. (2019). Conceptual Disagreement. *American Philosophical Quarterly*, *56*(1), 15-27.
Swindell, J. S. (2010). Ambivalence. *Philosophical Explorations*, *13*(1), 23-34.
Unwin, N. (1999). Quasi-Realism, Negation and the Frege-Geach Problem. *The Philosophical Quarterly*, *49*(196), 337-352.
Worsnip, A. (2019). Disagreement as Interpersonal Incoherence. *Res Philosophica*, *96*(2), 245-268.