The version of record of this article, first published in *Acta Analytica* (2019) 34:215–234 is available online at Springer Nature's website: https://doi.org/1007/s12136-016-0282-3

The author has inserted [page numbers] in square brackets to indicate the pagination of the version of record. Any citations of material herein should be to the version of record, not to this preprint version.

[419] On Knowing How I Feel About That – A Process-Reliabilist Approach

Larry A. Herzberg

Abstract: Human subjects seem to have a type of introspective access to their mental states that allows them to immediately judge the types and intensities of their occurrent emotions, as well as what those emotions are about or "directed at". Such judgments manifest what I call "emotion-direction beliefs", which, if reliably produced, may constitute emotion-direction knowledge. Many psychologists have argued that the "directed emotions" such beliefs represent have a componential structure, one that includes feelings of emotional responses and related but independent representations of what those feelings are about. I argue that such componentiality may help to explain how emotion-direction knowledge is achievable. I begin by developing a hybrid view of introspection that combines David Chalmers' phenomenal realism with Alvin Goldman's "partial redeployment" account of meta-belief content. I then provide a process-reliabilist account of introspectively gained emotion-direction knowledge that outlines the minimum conditions of reliably forming emotion-direction beliefs, and specifies several ways in which the warrant of such beliefs could be defeated by relevant counterfactual alternatives. The overall account suggests how distinct introspective processes might be epistemically synergistic.

1 Introduction

Experimental psychologists routinely use first-person reports to gather data on subjects' mental attitudes and emotions (Costall 2013). To collect such data, researchers use "Likert-like" scales consisting of a series of horizontal lines labeled by attitude type and evenly divided into numbered sub-units (e.g., from 0 to 5) to indicate intensity levels. In [420] some cases, contrasting attitude types (like joy and sadness) are placed at opposite ends of the scales, and the mid-point represents neutrality. In emotion research, the states to be measured are usually elicited by representations of scenarios that have either previously been correlated with emotions, or else are hypothesized to be so. Subjects are instructed to give their *immediate* judgments on how they feel about the scenario, the assumption generally being not only that they have direct (non-inferential) access to their emotions, but also that they have a reliable ability to recognize what their emotions are about. Measurements of somatic states, facial expressions, voice modulations, and other factors associated with emotion types can help to confirm whether the subject's report of her emotion type is accurate, and researchers may be justified in relying on such reports only given the availability of such corroborating data. But in practice, their reliance on such

reports reflects the common assumption that subjects' introspectively gained beliefs about their emotions are true and reliably produced, and hence (*ceteris paribus*) that they have a significant kind of self-knowledge.[1]

Given psychology's heavy reliance on such reports, it is surprising how little discussion there is in epistemology and philosophy of mind of how such "emotion-direction beliefs", as I call them, might be justified or warranted.[2] In fact, texts on self-knowledge tend to explicitly ignore such beliefs, analyzing instead the conditions of having self-knowledge of one's beliefs, intentions, desires, and similar propositional attitudes. This is to be expected in texts that approach the topic of self-knowledge from a rationalist perspective (e.g., Wright, Smith & Macdonald 1998), since rationalists are interested mainly in states for which the subject can be held directly responsible, and emotions are correctly assumed not to be such states. Even theorists who analyze emotions as evaluative judgments (e.g., Solomon 1984) or evaluative feelings (Prinz 2004) view subjects as being only *indirectly* responsible for them, via whatever control they have over the states that trigger them. Rationalists also tend to assume that rational subjects are epistemically *entitled* to know their own propositional attitudes, since they require such knowledge to exercise the critical reasoning abilities that constitute their rationality (Burge 1996). But almost no rationalist argues that such an entitlement extends to beliefs about one's emotional states.[3]

The scant attention self-knowledge of emotion has received from *non-rationalist* writers is harder to explain. For although their causal or acquaintance theories of self-awareness usually are meant to apply to simple sensations in addition to propositional attitudes, self-awareness of directed emotions is rarely if ever discussed.[4] Perhaps the omission here stems from the relative *complexity* of emotion. Many emotion theorists are "componentialists", taking *emotion* to be a "superordinate concept" that refers to clusters of other types of state that play distinctive causal or constitutive roles (e.g., Lazarus 1999). Componentialists disagree on *which* types of state play *which* types of role, and this lack of consensus might explain epistemologists' reticence on the subject. Also, *whatever* the relevant components turn out to be, epistemologists might reasonably think it best to analyze the conditions of self-knowledge of each component before [421] analyzing the conditions of the sorts of complex states that are apparently represented by emotion-direction beliefs. However, following such a "bottom-up" methodological principle may be misguided in the case of emotion-direction beliefs. For, as I will elaborate below, it may well be the *complexity* of the "directed emotions" they represent that best explains how emotion-direction beliefs can be warranted.

Along with many mainstream emotion theorists, I assume in what follows that emotional responses are mediated by the autonomic nervous system, and occur in response to characteristic sorts of mental representations. The initial responses include modulations of cognitive processing and modifications of hormonal, visceral, muscular, vascular, and related physiological conditions. Typical expressions of these conditions include facial expressions, modifications of posture, alterations of tone of voice, and so on. Importantly,

---

[1] 'Self-knowledge' may refer to knowledge of one's mental states, or to knowledge of some entity that *has* mental states, namely oneself. I use the term here only in the former sense. The same is true for 'self-awareness'. I discuss reliabilism about knowledge below.

[2] I use 'justified' and 'warranted' roughly synonymously, but I reserve the former term for cases of inferred beliefs, and the latter for cases of non-inferred beliefs. See note 10 below.

[3] Bilgrami 1998 is perhaps the main exception to this rule.

[4] See Gertler 2011 for discussion of the relations between theories of self-awareness and self-knowledge.

these initial responses are at least *registered* by the somatosensory and proprioceptive systems that subserve emotional *feeling* (LeDoux, 1996). However, emotional feelings can also result from activation of neurologically based body maps, or "as-if body loops" (Damasio 1994). The positive or negative valence of an emotional feeling, in association with a representation of what the feeling is about or directed at, can flexibly motivate coping behaviors (Lazarus 1991).

It is debatable whether emotional feelings themselves have representational content, and if so whether they represent bodily conditions or significant organism-environment relations (Prinz 2004). My view does not require them to be representational, and I tend to favor the view that they are non-representational sensory registrations of physiological emotional responses.[5] It is also debatable whether the term 'emotion' refers to the *feeling* of the neuro-physiological responses, or whether it refers instead to those initial responses themselves. Prinz follows James (1889) in arguing for the former; Damasio, LeDoux, and others argue for the latter. Ekman (1999) and Lazarus both view the neuro-physiological conditions and the subjective experiences caused by those conditions as normal components of emotion. It is a matter of still more debate whether the representation that triggers an emotional response (or the representation of what the emotional response is about or directed at) should be considered part of the emotion *per se* or not.[6] To some extent, this is a terminological issue on how to use the term 'emotion', and I will not take a position on it here.

Perhaps my most central assumption about emotion is that emotional feelings have a distinctive phenomenology, and that such "valenced" states have the psychological function of flexibly influencing the subject's behavior. Furthermore, I assume that the *motivational* functions of emotional feelings cannot reliably be fulfilled unless they are *directed at* what they are about, and that they are so directed via representational states. A feeling of fear caused by the sight of a snake cannot fulfill its function of causing the subject to move *away from* the snake unless it is associated (in a particular way) with a representation of the snake. Following Prinz, I call whatever an emotional feeling is about its "particular object", and assume that the feeling itself is both psychologically and neurologically distinct from any representation of its particular object (Damasio 1994). However, unlike Prinz and Damasio, I hold that there are many cases in which the [422] representation of the emotion's particular object is not the same as the representation of whatever causally triggered the emotion, as when my belief that it is raining causes me to be disappointed that I cannot have a picnic.[7] That is, as I have argued elsewhere, *emotion-direction* is distinct from *emotion-causation*.[8] It is largely because psychologists often

[5] For more discussion of this point, see Herzberg (under review). Cf. Burge (2010) for the distinction between perceptual representation and mere sensory registration.

[6] Lazarus (1999) goes to some lengths to argue that triggering representations become components of the emotions they trigger. Prinz (2004) tries to finesse the issue by distinguishing "state emotions" (emotional feelings without their triggering representations) from "attitudinal emotions" (such feelings plus their triggering representations).

[7] In this case the content of the state representing the emotion's particular object is different than the content of the belief that causally triggered the emotion. However, there are other sorts of case where the contents remain roughly the same but their modes of representation differ, as when one's fear caused by a visual percept a nearby snake is sustained (as one runs away) by a memory of having seen a nearby snake.

[8] Herzberg (2009). In that article I referred to *affect*-causation and *affect*-direction, but I have since decided that using the term 'affect', which is broader and perhaps less well-defined than 'emotion', raises unnecessary questions.

conflate the two relations that their subjects may indifferently be asked to report how *X* is *making them feel*, rather than how they *feel about X*. However, only the former question requires the subject to make a causal judgment, and there is no reason to believe that such judgments are as reliable as the corresponding emotion-direction judgments might be.[9] So in this paper I focus only on the epistemic warrant of judgments or self-reports having the form "I am feeling *E* about *P*" (and similar constructions), where *E* and *P* conceptually represent the emotion's type and particular object respectively. I take such reports to be manifestations of *emotion-direction beliefs*, which when warranted or reliably produced normally yield *emotion-direction knowledge*.[10]

Before I sketch out my analysis of emotion-direction knowledge, I should explain why I find process-reliabilism to be the epistemological framework that is best-suited for this purpose. In the late 1970s, Alvin Goldman developed the first systematic statement and defense of the view that, as he later put it, "a belief's justifiedness is fixed by the reliability of the process or processes that cause it, where (as a first approximation) degree of reliability consists in the proportion of beliefs produced by the process that are true." (2011, 8) In (1979) he argues that to be maximally informative, a theory of justification should not merely analyze one epistemic concept in terms of another, but rather should analyze epistemic concepts in non-epistemic terms. This strategy is common in other normative (in the sense of *prescriptive*) domains, such as ethics. Utilitarianism, for instance, is considered by many to constitute progress in moral theory precisely because it analyzes moral rightness in terms of presumably measurable states, such as happiness or material security. On Goldman's view, informative epistemological theories should similarly analyze epistemic concepts in *non-epistemic* terms, including semantic or representational relations like *truth* or *accuracy,* doxastic states like *belief*, *disbelief*, and *suspension of judgment*, psychological processes like *memory*, *perception*, *introspection*, or *inference*, and, perhaps most importantly, *causality*, *modality*, and *probability*. By contrast, an analysis of justification that appeals only to, say, the *rationality* of holding a belief, or to the subject's having access to *good evidence*, is not similarly informative, since 'rationality' and 'good evidence' are themselves epistemic terms. It should be noted, however, that one can agree that "naturalizing" epistemic concepts constitutes progress in epistemology without holding that successfully completing the project could rid epistemology of normativity, for it seems that normative issues must still arise in regard to, for instance, [423] the *sorts* of causal relation that are warrant-conferring, the *degrees* of reliability they must have, and the *frames of reference* within which their reliability is to be determined, all of which may differ by context.

Goldman (1979) also argues that an adequate account of justification must focus first and foremost on the *actual* process or method that produces the belief, and not merely on the various resources available to the believer at the time the belief is formed (although these can enter into the account secondarily). For instance, even if the believer's having access to sufficient evidence *can* justify a belief, it can do so only if the believer – or, in the case of a non-inferred belief, the sub-personal belief-forming process – actually *uses* that evidence to form the belief. He notes that processes that produce paradigms of

[9] Nisbett and Wilson (1977) famously found that subjects sometimes "confabulate" erroneous conjectures about the causes of their preferences.

[10] I say 'normally' here because I am thinking of cases of warranted true beliefs in which the Gettier problem does not arise. Other conditions may need to be added to rule out Gettier cases.

unjustified beliefs, such as wishful thinking, guessing, hasty generalization, and perception under poor conditions, share the feature of being *unreliable*. Similarly, processes that produce paradigms of justified beliefs, including perception under favorable conditions, memory, introspection, and valid or strong patterns of reasoning, share the characteristic of being *reliable* (at least given reliably produced inputs). But, as Goldman makes clear, to judge the *reliability* of the belief-forming process that actually formed a particular belief, one must consider not only how that process performed in the actual circumstances, but also how it *would* perform in a range of relevantly similar circumstances. For even a long track record of actual success is not dispositive, given that the circumstances in which the belief was produced might be unprecedented. What is wanted, in other words, is a *propensity*, and not merely a *frequency*, analysis of reliability.

Over the last several decades, process-reliabilism has undergone significant refinements in response to objections and putative counterexamples that have been raised against it.[11] I cannot here recount that dialectic, but I do believe that process-reliabilism has survived as a viable epistemological framework, and I have two main reasons for finding it particularly well-suited to the task of analyzing the conditions of emotion-direction knowledge. First, process-reliabilism is *directly* applicable to *non-inferred* beliefs in a way that widely held alternatives are not, and, consistent with psychology's aforementioned presupposition that subjects can report their *immediate* beliefs regarding their emotions, it seems that emotion-direction beliefs are often non-inferred, resulting instead from more immediate, sub-personal conceptualization processes.[12] A somewhat apt (but admittedly controversial) analogy here is to beliefs about one's immediate environment, which often are not inferred from other beliefs, but result instead from processes that conceptualize aspects of the environment by way of perceptual states that presumably have only non-conceptual content. Similarly, beliefs [424] about the types and contents of one's mental states seem to be formed by conceptualization processes that have non-inferential, introspective access to the mental states they are about (although whether that access should be understood as being mediated by states with non-conceptual content is a question to be considered below). Of course, it does not follow from a belief's being *produced* non-inferentially that it is also *justified* non-inferentially, and some have argued that non-inferentially produced beliefs can indeed be justified only inferentially.[13] However, young children and cognitively impaired adults do seem to form beliefs about their mental states (as well as about their immediate environments), and it seems that they are at least epistemically *entitled* to use those beliefs in their reasoning or to guide their behavior, even if they lack the cognitive abilities needed to form a justificatory argument

[11] For Goldman's own summary of both the objections and his replies, see (2011).

[12] A note on terminology: as I use the term, 'inference' refers to an (often deliberate) activity governed by rational and logical norms, and for which the subject can properly be held responsible. Importantly, both the inputs and outputs of inferential processes are conceptual representations, most commonly *judgments*, which are manifestations of beliefs. Beliefs are *assertoric propositional dispositional* attitudes (as these terms are commonly analyzed). *Inferred beliefs* are produced by inference. *Non-inferred beliefs*, by contrast, are often produced by *sub-personal* processes over which the subject has little or no direct control, although the subject may properly be held responsible for endorsing or rejecting a belief that has been non-inferentially produced. Unlike inferential processes, the *inputs* to non-inferential processes are usually non-conceptual representations like perceptions, non-representational or pre-representational sensory registrations, or indeed nothing at all (e.g., random guessing). But there are also non-inferential processes – free association, for instance – that have conceptual inputs.

[13] Cf. Bonjour (1976).

in favor of their likely truth. This coheres with process-reliabilism's *externalism*: a subject who is warranted in believing (or entitled to believe) that *P* need *not* be justified in believing that she is warranted in believing (or entitled to believe) that *P*. By allowing for such epistemic warrant, process-reliabilism's externalism significantly *increases* its scope of application, since it is applicable not only to beliefs formed non-inferentially by sub-personal processes, but also to those formed inferentially by personal-level processes – the sorts of processes that preoccupy internalists about justification.

Indeed, even if process-reliabilism falls short of one's favored standard of knowledge or justification, there is a good reason to use it to better understand the conditions of merely *reliably formed* beliefs, and this is my second reason for using the framework here: unlike what Goldman calls "current time-slice" views of justification that focus entirely on conditions of the subject at the time of belief, process-reliabilism (unsurprisingly) focuses on potentially interrelated *processes*, and particularly on *causal* processes. This emphasis allows for a synergistic cooperation between epistemology and psychology that might otherwise be lacking. For a process-reliabilist analysis should begin by mapping out, at a functionalist level, the likely sub-processes that seem to be required to produce a belief about a mental state with complex properties or even a componential structure, especially when those properties or components are as diverse as they seem to be in directed emotions. Since each sub-process might have its own degree of reliability, and reliability is in principle a *measurable* quantity, such an analysis seems capable of suggesting empirical hypotheses worthy of investigation. For if the existence of the relevant sub-processes can be verified, and their degrees of reliability in various circumstances estimated, the resulting information might suggest sub-process-specific diagnoses and treatments of systemic cognitive breakdowns. It is hard to imagine how a non-process-oriented, "current time-slice", purely conceptual analysis of the evidence accessible to a subject at a given time could play such a potentially useful role. Of course, to play this role, the epistemological analysis itself must be responsive to developments in cognitive psychology, but it goes without saying that any philosophy that deals with mental states should be informed by the best current psychology.

One might still wonder, however, why I have chosen to focus on the reliability conditions of introspectively formed beliefs about one's *emotions*, rather than on beliefs about other sorts of mental state that feature more prominently in the epistemological literature: beliefs about one's own beliefs, intentions, or desires. The simple answer is [425] that I am interested in investigating how different modes of introspection might be required to achieve self-awareness and conceptualization of different sorts of mental state, or of different properties of the same state, and how these different processes might positively or negatively influence each other. In particular, I am interested in how self-awareness and conceptualization of states with salient *phenomenal* properties might influence self-awareness and conceptualization of states with salient *representational* properties, and *vice versa*, and there seem to be no mental states with more clearly distinct phenomenal and representational properties or components than directed emotions.[14] Beliefs, intentions, and at least some desires arguably lack phenomenality, or at least their phenomenal aspects

[14] The reader may have noticed that I am hedging here a bit between considering *properties* and *components*. As I have already indicated, I favor a componential view of directed emotion structure, but my epistemological interests extend to self-knowledge of perhaps more singular states with both phenomenal and representational properties.

are less salient those of fear, sadness, anger, disgust, joy and the like. So it is precisely because directed emotions have a complex or componential structure that include both salient phenomenal and representational properties that I find them particularly useful for present purposes.

That is a quick summary of my guiding presuppositions. In the next section I outline my view of introspection by contrasting it with Goldman's (2006) "quasi-perceptual" view. My view differs from his in allowing *non-causal* – mereological, constitutive, or "embedding" – relations to ground the reliability of applications of phenomenal concepts to phenomenal states or properties (consistent with Chalmers 2003). It also abandons Goldman's preference for a *unified* theory of introspection; I argue that a "hybrid" theory allowing both causal and embedding relations to ground introspective beliefs is at least as plausible as Goldman's theory. However, at least in terms of its form, the analysis of emotion-direction knowledge I outline in section 3 is heavily indebted to Goldman's (1986) analysis of non-inferential perceptual knowledge. Dissimilarities here arise from the fact that introspection involves awareness of one's mental states, rather than awareness of objects in one's environment. My analysis is also more complex than Goldman's, since I am assuming that directed emotions are composed of an emotional feeling and a separable representation of that feeling's particular object. The upshot of this complexity is that if we do have emotion-direction knowledge (as is normally supposed), our having it is a considerable cognitive achievement, and may suggest how different modes of self-awareness can be epistemically synergistic.

2 A hybrid view of self-awareness

2.1 Goldman's "quasi-perceptual" theory of introspection

Goldman (2006) develops his "quasi-perceptual" theory of introspection by contrasting it with non-perceptual views and defending it against well-known objections to similar views. Such objections center on purported disanalogies between self-awareness and sense perception. For instance, Shoemaker (1996, 207) argues that, unlike visual perception, introspective awareness of one's own mental states fails to have a [426] distinctive phenomenology – a *way things appear*. Goldman concedes this point, arguing that introspection need not resemble sense perception in *all* respects for it to be at least quasi-perceptual (228).[15] He then points out that introspection, like sense perception, "requires or is at least facilitated by attention," and uses this point to reply to Shoemaker's further objection that introspection lacks an "orientating organ". Attention, Goldman argues, *is* the "organ" of introspection, "the orientation of which puts a subject in an appropriate relation to a targeted state" (244). As I discuss below, alternative views of self-awareness also view attention as playing a key role in introspection.[16]

Goldman goes on to attack "*pure* redeployment" or "monitoring mechanism" views of first-person access.[17] Such views hold that if a proposition *P* is already in one's "belief

[15] Unless otherwise noted, all subsequent references to Goldman are to (2006).

[16] Goldman (2006, 244) also defends his view from Shoemaker's famous "self-blindness" objection. However, Gertler (2011, 149-159) mounts a more broadly applicable defense by arguing that the conceptual requirements of rationality are orthogonal to the viability of "inner-sense" views of self-awareness.

[17] Goldman is thinking specifically of Nichols and Stich (2003).

box", the monitoring mechanism underlying first-person access simply copies ("redeploys") *P*'s content, appends "I believe that..." to it, and then adds the resulting "I believe that *P*" to the belief box. This sort of process seems capable of operating syntactically and non-representationally; its operations could easily be programmed into a computer. Goldman's concern here is that such a view has difficulty explaining self-ascriptions of states other than beliefs. For any process that takes a proposition from, say, the *fear* box, prefixes it with "I fear that", and adds it to the *belief* box must be able to discriminate between the "boxes" themselves. But if the boxes are understood functionally or dispositionally (as they usually are), and functions or dispositions *per se* are causally inert (as they are thought to be), it is unclear how such discrimination could occur unless the monitoring mechanism were able to at least quasi-perceptually discriminate between the dispositions' *categorical bases*, which are presumably neurological. It is precisely such a quasi-perceptual discriminative ability that Goldman will build into his own theory of introspection.

But Goldman recognizes that, by itself, a process's being causally sensitive to some set of "input properties" is insufficient to justify viewing it as perceptual. For perceptual systems must also *transduce* the "input properties" to which they are causally sensitive into output states with non-conceptual representational content. For instance, as Goldman uses the term, the visual system transduces patterns of light intensities registered on the retina into visual percepts that non-conceptually represent objects and their properties. In cognitively capable subjects, aspects of percepts can be further "translated" into concepts from which further inferences can be drawn. So Goldman must explain (1) just what *are* the input properties to which introspection is causally sensitive, and (2) how we should conceive of introspection's immediate, non-conceptual output.

To answer (1), Goldman assumes that there are only four plausible candidates for the input properties: functional, phenomenal, representational, and neural.[18] He quickly rules out functional properties for the same reason that he ruled out pure redeployment views, namely their subjunctive or dispositional nature: they specify what *would* occur *were* certain conditions to be satisfied. Fragility is a paradigm dispositional property; a [427] fragile object is one that *would* break *were* a sufficient amount of force applied to it. Just as vision is (and must be) causally insensitive to an object's fragility, it seems that introspection must be causally insensitive to a mental state's dispositional properties.

Goldman then rules out phenomenal properties, partly because some introspectable states – such as certain thoughts – seem not to have them, and he is seeking a *unified* theory of self-awareness. He also rejects phenomenal properties on the grounds that if, as many believe, they supervene on physical properties, they may well be *epiphenomenal* (since their causal power could be adequately explained physically) and hence – *qua* phenomenal properties – causally inefficacious. Since no process can be causally sensitive to causally inefficacious properties, he concludes that a mental state's phenomenal properties cannot be the input properties he is looking for. I will return to this point below, for it is precisely the point on which Goldman and I most disagree.

Finally, Goldman rules out introspection's being causally sensitive to representational properties on the grounds that such properties seem incapable of explaining one's ability to introspectively distinguish, say, a *belief* that *P* from a *desire* that *P*, assuming that *P* has the same content in both states, and that its content exhausts the content of both the belief

[18] The following summary is of (246-253).

and the desire. Also, a belief or desire's representational content seems to carry no information regarding an attitude's *intensity*, which Goldman, like many psychologists, assumes to be introspectable.

By a process of elimination, then, Goldman concludes that introspection can reveal the types and intensities of first-order mental states only by being causally sensitive to their *neural* properties. Classification occurs, he speculates, "on the basis of which groups of cells are activated." In this way, "the representation of mental state types is accomplished by a perception-like recognition process, in which a given occurrent token is mapped into a mental category selected from a smallish number of types." (252) However, given the indeterminately large number of *content* types, a mental state's content cannot be similarly recognized. Here Goldman accepts a limited redeployment theory. *Judging* that *I hope that P* requires *P*'s content to be redeployed: "The hope's content is replicated by the metarepresenting state." (254) Simple redeployment from a first-order state to a meta-belief is possible if the first-order state's content is conceptual. But if it is non-conceptual, "There must be an intramental *translation*, from one mental code to another." (254) So, on Goldman's view, introspection involves at least three sub-processes, each of which (we should note) could have different degrees of reliability: quasi-perceptual recognition of attitude type through causal sensitivity to neural properties, redeployment of conceptual content, and, when necessary, translation of non-conceptual to conceptual content.

As for the sorts of non-conceptual representations immediately output by introspection, Goldman hypothesizes that an *introspective code* ("I-code") with a proprietary "introspective vocabulary" represents a state's doxastic, valence, and bodily feeling dimensions. For instance, "HOPE may represent a mental-state category that combines desire on the valence dimension and doubt, or uncertainty on the doxastic dimension." (261) Importantly, Goldman holds that transitions from I-coded representations to concepts do not occur in isolation. Rather, "the suggestion is that I-coded representations are *among* the representations that figure in mental-state concepts like BELIEF, DESIRE, FEAR, and LOVE." (263) Other dimensions, including functionalist ones, enrich such concepts. But Goldman asserts that "introspective representation serves as default evidence for the token state being [of a certain type] and, absent defeating [428] evidence, yields that classification." (263) In what follows, I agree with Goldman that ordinary mental state concepts include functionalist parameters in addition to introspectively derivable ones. I also agree with him that introspection can yield "default evidence" for a classification of mental state type. However, I argue that there is currently no convincing reason to adopt his I-code hypothesis, or, more broadly, his "unified" quasi-perceptual approach to introspecting a mental state's type, given that there is a viable alternative.

2.2 Concerns about Goldman's view of introspection

I believe that the plausibility of Goldman's view that introspection outputs states with non-conceptual representational contents analogous to those of sense percepts ultimately depends on whether a future science of introspection discovers that introspection incorporates any *perceptual constancies*. In perceptual systems like vision and touch, perceptual constancies – adaptive biases derived from environmental regularities – are used to process early sensory registrations, yielding "objectifications" of distal objects and their properties (as opposed to mere registrations of proximal effects on sensory surfaces). Burge

(2010) argues that a sensory system's incorporation of a perceptual constancy is the most "reliable mark" of its being perceptual; it might even be a necessary condition. Although it is certainly conceivable that *neurological regularities* could provide bases for perceptual constancies in introspection in somewhat the same way as *environmental regularities* provide bases for perceptual constancies in vision and other modes of perception, to my knowledge there is so far no empirical evidence that they do.

This lack of empirical support is not enough to reject the view outright, but I do believe that Goldman too quickly dismisses the possibility that phenomenal properties play a central role in introspection. His dismissal of them seems to be primarily motivated by his view that "there is a strong prima facie case for a (substantially) *unified*, or *homogeneous*, account of first-person privileged access." (227) This is the source of his concern that *some* introspectable states *lack* phenomenal properties. For if phenomenal properties cannot be used to account for one's access to *all* of one's mental states, Goldman's goal of a unified account requires that they cannot be used to account for access to *any* of one's mental states. But this restriction comes at the cost of intuitive appeal, for while there currently is no reason to believe that an I-code exists (other than the explanatory role it plays in a theory like Goldman's), the phenomenal qualities of one's sensations certainly seem to be both introspectable and discriminable. I have no reason to doubt that I can reliably discriminate an itch from an ache, or a feeling of sadness from one of fear or anger.[19] A theory of introspection should *accommodate* and hopefully explain this fact, rather than rule it out. It is also worth noting that Goldman does allow for *some* disunity in his account, insofar as he distinguishes between introspection of *content* via redeployment (with or without translation) and [429] introspection of *attitude type* via sensitivity to neurological properties. So why should he not also allow that different mental state types, or even different properties of a single state, could be accessible via different types of introspective process?

More fundamentally, Goldman's concern that phenomenal properties might be causally inefficacious were they to merely supervene on physical properties seems to me to put the cart in front of the horse. A *psychological* theory of introspection should be guided not primarily by metaphysical concerns, but rather by the need to provide adequate psychological explanations of behavior, and surely one's awareness of phenomenal properties *per se* can be psychologically relevant to explaining one's behavior. One's taking aspirin can often be satisfactorily explained by citing one's desire to get rid of a headache (and one's belief that the drug will relieve it). Similarly, citing the unpleasant phenomenology of generalized anxiety can play a key role in explaining why a patient seeks therapy. Unlike explanations of behaviors motivated by one's immediate environment, such explanations do not cite intermediate stages of mental processing requiring non-conceptual representations. It is because we take such explanations to be at least contextually adequate that we should seek an account of how introspectively formed beliefs might represent sensational or phenomenal qualities in a more direct way than

[19] If there are such differences to be felt, they stem from somatosensory registrations of the neurophysiological profiles associated with emotion types. Both Prinz (2004) and LeDoux (1996) argue that the profiles of at least basic emotions involve enough parameters to produce discriminable differences. In cases of non-basic emotions, I believe that the subject's awareness of the emotion's particular object may play a key role in accurate type recognition, and allowing for this is a key feature of my analysis of emotion-direction knowledge.

Goldman's theory allows. Conveniently, David Chalmers offers just such an account, and I will adopt it here after summarizing its most relevant features.

2.3 Chalmers' account of phenomenal concepts and phenomenal beliefs

Chalmers' (2003) discussion of phenomenal qualities and their direct conceptualization focuses almost entirely on color properties, but he explicitly intends his view to cover the phenomenal qualities of emotional experiences as well (235), so for present purposes I will illustrate his view using cases of emotional feeling. Note that the discussion below only partially describes his view. For a more complete exposition and defense of it, one should consult the original essay.

It is consistent with Chalmers' view that when one attends to the phenomenal qualities of one's emotional feelings and thinks, for instance, *I am feeling sad*,[20] there are several concepts of *feeling sad* that might yield a true belief.[21] First are concepts of emotional feelings that have their references fixed relative to either community or individual usage. Such concepts can respectively be glossed as "the phenomenal quality of the feeling typically caused *in normal subjects within my community* by paradigmatically sadness-eliciting events", and "the phenomenal quality of the feeling typically caused *in me* by paradigmatically sadness-eliciting events". Both are plausible (albeit competing) interpretations of the phenomenal concept that underlies the linguistic sense of 'feeling sad'. Following Chalmers' convention, we can label these two concepts feeling sad_c and feeling sad_i. Importantly, insofar as these are *phenomenal* concepts, one's ability to use them to report one's emotional experiences to others apparently depends on a more fundamental ability to pick out a phenomenal property indexically, [430] via a demonstrative concept *E* that can rigidly designate the phenomenal quality of a present feeling by simply *attending* to it ostensively while thinking *this experience*. Call such a demonstrative concept *E*. With the ability to use *E*, one could introspectively form the belief that *I am feeling E, and E is* sad_c (or sad_i).

In addition to these three ways of *relationally* determining the reference of a phenomenal concept like *feeling sad*, Chalmers argues for a fourth type of phenomenal concept that picks out a phenomenal property "in terms of its intrinsic phenomenal nature", and hence *directly* or *non-relationally*. If we have such a "pure" phenomenal concept of the phenomenal quality of feeling sad, it picks out the phenomenal quality of sadness *as the phenomenal quality that it is*, and not merely as 'this experience' or in terms of any relations it bears to normal occurrences in oneself or others. Chalmers admits that such pure phenomenal concepts are "difficult to express directly in language", since language must rely on community-based concepts. But he points out that some philosophers have at least *stipulated* uses for pure phenomenal concepts (e.g., Chisholm 1957), and notes that Russell (1910) might have had pure phenomenal concepts in mind when he claimed that we have a special capacity for direct reference to our experiences, i.e., that we are in some special way *acquainted* with them, and that this sort of acquaintance could be a source of

[20] In what follows, I follow Chalmers' convention of *italicizing* concepts or the conceptual contents of beliefs.

[21] Identifying the types of one's emotional feelings need not be based entirely on attending to their phenomenal qualities. Just as Goldman assumes that emotion type concepts like *hope* are not exhausted by their I-code representations, I assume that they are not exhausted by their phenomenal properties.

warrant for our phenomenal beliefs.[22] (233-34)

When the content of a phenomenal concept is partly *constituted* by a phenomenal quality of a present experience, it is a *direct phenomenal concept*. As Chalmers puts it, "The clearest cases of direct phenomenal concepts arise when a subject attends to the quality of an experience, and forms a concept wholly based on the attention to the quality, 'taking up' the quality into the concept." (235) I refer to such "taking up" as the *embedding* of the quality by the concept.[23] Such concepts do not characterize their objects *as* objects of attention, but their formation *requires* acts of attention, and *the same act of attention* can be used to form both a demonstrative concept *and* a direct phenomenal concept of a given quality. *Direct phenomenal beliefs* can thus be acquired "when the demonstrative phenomenal concept...and the direct phenomenal concept...are *aligned*: that is, where they are based in the same act of attention." (236) Importantly for our purposes, not all phenomenal concepts are direct, since one can *retain* a phenomenal concept long after the experience that instantiated the quality to which it refers has ceased. Although such *standing phenomenal concepts* may be more "coarsely grained" than direct phenomenal concepts, they can be used to recognize new instances of a phenomenal quality as being roughly the same quality as a previous instance. Chalmers speculates that the content of a standing phenomenal concept might be determined by "some combination of" (1) cognitive states that "bear a relevant relation to the original phenomenal quality in question" (perhaps a memory image of the quality), "(2) dispositions to have such states; and (3) dispositions to recognize instances of the phenomenal quality in question." (238-239)

[431] Importantly, Chalmers defends his brand of phenomenal realism against Shoemaker's (1975) charge that when such a view is combined with epiphenomenalism, phenomenal beliefs cannot play a *causal* role in the production of introspective knowledge – the same concern that Goldman raised against the possibility that introspection is causally sensitive to phenomenal properties. As Chalmers glosses it, Shoemaker relies here on the premise that if the phenomenal qualities are *causally* irrelevant to phenomenal beliefs, no such belief can count as knowledge. But Chalmers replies that if there are direct phenomenal concepts whose referents are determined through careful acts of attention, then "the connection between experience and phenomenal belief is tighter than any causal connection: it is constitution. And if a causal connection can underwrite knowledge, a constitutive connection can certainly underwrite knowledge too." (256)

Direct phenomenal beliefs formed through *careful* acts of attention might be infallible. However, not all acts of attention are careful, and it is important to note that there is no reason to believe that most phenomenal beliefs are direct; rather, most would seem to involve comparisons between *standing* and *demonstrative* phenomenal concepts, where there is clearly room for error. For instance, the first time I suffer a serious loss and experience sadness about the loss, I might form a direct phenomenal concept of the sensation I am experiencing, *S*. My memory of *S* (with its embedded phenomenal quality) might provide me with a standing phenomenal concept, S_s, which on a future occasion

[22] Of course, it was precisely Russell's view of acquaintance that inspired Wittgenstein (1953) to develop his argument against any "private language", particularly one that could refer to a sensation type. Here I will simply concur with Chalmers' comment on Wittgenstein's argument: "I can say only that I have seen no reconstruction of it that provides a strong case against the view I have laid out." (241)

[23] Gertler (2001) provides a detailed metaphysical account of the embedding relation, and relates it to her "demonstrative attention" account introspection. Chalmers describes the difference between their accounts in terms of the relative priority of attention and embedding. I take no position on that issue here.

allows me to form the phenomenal belief, *this feeling is* S_s (that is, *the phenomenal quality to which the demonstrative concept 'this feeling' refers is of the same type as the phenomenal quality embedded by* S_s). Such a belief would clearly be fallible, for it may be false that the phenomenal quality of the demonstrated feeling is of the same type as the phenomenal quality embedded by S_s.[24] Still, if the conceptualization and comparison processes are reliable, I might eventually come to form the warranted belief that S_s is S_i – the sort of feeling I *normally* experience on occasions of loss. Later I might infer with some warrant that S_i is S_c, the sort of feeling that normal members of my community typically experience on occasions of loss, feelings called 'sadness'.[25] Similarly, if on some occasion of loss I become introspectively aware that the feeling I am experiencing is *not* S_s, I am able to form the warranted phenomenal belief that *this feeling is not sadness* (S_i or S_c). It is one's ability to form such phenomenal beliefs that often explains one's ability to recognize when one's emotions do not match those that either we or our community deem normal or appropriate in the circumstances.

I find Chalmers' view of how beliefs about phenomenal qualities are formed to be at least as plausible as Goldman's quasi-perceptual view of how introspective beliefs about emotional states like hopes are formed. Obviously, empirically verifying Chalmers' view would be difficult, but no more so, I think, than verifying Goldman's view. So I assume below that introspectively forming an [432] emotion-direction belief can involve, among other factors, the use of demonstrative and standing – and hence on some occasions *direct* – phenomenal concepts, as well as the ability to relate such concepts to the community-based concepts (with their functionalist parameters) that ground emotion-type vocabulary. However, I also retain Goldman's "partial-redeployment" account of how meta-beliefs inherit the conceptual content of the propositional attitudes they are about, as well as his assumption that conceptualization of non-conceptual content requires translation. Finally, I remain neutral on the question of whether introspective awareness (and subsequent meta-conceptualization) of the types and intensities of one's occurrent, *non-emotional* propositional attitudes is generally best explained by a quasi-perceptual or phenomenal realist model.[26]

3 A Process-Reliabilist Analysis of Emotion-direction Knowledge

As I mentioned in section 1, Goldman argues that an adequate characterization of *reliability* must include a subjunctive condition regarding how the process that was actually used to form a true belief would perform in a range of counterfactual circumstances, and there has been much debate in the epistemological literature concerning the exact form such a condition should take.[27] The condition I adopt below is most closely related to

[24] Even if the belief that *this feeling is* S_s is literally true, the belief might be faulty in a rather different way. For, being recalled from memory, the phenomenal quality that S_s embeds might no longer match the phenomenal quality originally embedded in *S*, and insofar as the function of S_s is to memorialize that quality, it will have failed to fulfil that function. So the belief, although true, might be misleading.

[25] Such an inference would of course require me to draw from a great deal of background knowledge about the similarities between myself and others, as well as some basic assumptions about the relationship between phenomenal and neurological properties.

[26] I say 'generally' here to allow that meta-beliefs in epistemically rational subjects might *constitute* the first-order beliefs they are about. See brief discussion below.

[27] There have been competing characterizations of the subjunctively expressed conditions on reliability,

Goldman's (1986) view that a belief-forming process is reliable only to the extent that it can discriminate between *relevant counterfactual alternatives*, that is, one that tends to produce a belief with the same content when and only when the belief would be true. Obviously, there is plenty of room for discussion about the range of counterfactual circumstances that are to count as *relevant*, as well as the required precision of 'same content'. But since I am here simply adopting Goldman's relevant counterfactual alternatives approach, which he develops mainly to evaluate the reliability of *perceptually* produced beliefs about one's environment, the question that most immediately concerns us is whether the approach requires any modification in order to be useful in evaluating the reliability of *introspectively* produced beliefs about one's mental states, where introspection is understood along the "hybrid" lines sketched out above.

Goldman's well-known wolf/dachshund case illustrates why, in the case of visually-produced beliefs, a notion of *perceptual equivalence* is required for a reliabilist analysis of perceptual knowledge, rather than a broader notion of relevant counterfactual alternative. That case exploits the intuition that after Oscar forms the true belief that *the object over there is a dog* (caused by his seeing a *dachshund*), neither the mere possibility nor even the high probability of the object's counterfactually being a *wolf* defeats his claim to knowledge, even if Oscar tends to mistake wolves for dogs. This is because, given the difference between the visual appearances of dachshunds and wolves, seeing a wolf is not a perceptual equivalent of seeing a dachshund. Since the case stipulates that there are no *non-dogs that look like dachshunds* nearby, we feel [433] confident affirming that Oscar *knows* that the object over there is a dog. So for a relevant counterfactual alternative to defeat a claim to perceptual knowledge, it must be a *perceptual equivalent* of the actually perceived object or state of affairs.

To adequately analyze introspective knowledge, do we need a notion of "introspective equivalence" similar to that of perceptual equivalence? Perceptual equivalence is required in the case of visually produced beliefs because a belief with relatively general content (e.g. one that includes the concept *dog*) can be caused by a perceptual representation with more specific or "finely grained" non-conceptual content (e.g., the specific visual appearance of a particular dog). So if (as Goldman believes) the conceptualization of a mental state's type similarly depends initially on non-conceptual I-code representations of a token state's type that are more finely-grained than the *conceptual* representation of that state's type (as, say, a *belief, desire,* or *hope*), it seems that some notion of introspective equivalence will indeed be required. Similarly, on the phenomenal realist view outlined above, whether a notion of introspective equivalence is required would seem to depend on how "coarsely grained" various sorts of community-based phenomenal concepts are relative to the specific phenomenal properties that are embedded by direct and perhaps demonstrative phenomenal concepts. However, given how little is currently known about the degrees of graininess in question, in what follows I rely only on a notion of relevant counterfactual alternative that does not require introspective equivalence.

With these points in mind, I can now outline my process-reliabilist analysis of *emotion-direction knowledge*. The analysis is restricted to introspectively-based processes. I am not ruling out other ways of achieving emotion-direction knowledge, such as by being

---

starting with Dretske's (1971) "conclusive reasons" view, and extending through Goldman's (1976, 1986) "relevant alternatives" view, Nozick's (1981) "sensitivity" requirement and Sosa's (1996) closely related "safety" requirement. Goldman (2011) defends his "relevant alternatives" view, which I adopt here.

convinced by an authoritative source like a clinical psychologist. Also, although the clauses are written so as to allow the processes to be sub-personal and non-inferential in a way that is consistent with epistemological externalism, they do not rule out the subject's reliable use of inference, and so do not exclude internalist routes to knowledge. Finally, it should be noted that this analysis presupposes a "hybrid" view of introspective self-awareness that combines elements of Goldman's approach to meta-belief formation (in clauses 4 and 6) with Chalmers' phenomenal realism (in clause 5). Goldman's epistemological influence is perhaps most evident in clause 7, which requires the absence of any "relevant emotion-direction alternative". After presenting the outline, I will discuss each clause in turn-

At time *t*, subject *S* has introspectively-based *emotion-direction knowledge* that she has an emotional state of type *E* about whatever is conceptually represented by *P* if and only if...

(1) *S* has an emotional state *A* of type *E*.

(2) *S* has some representational mental state *M*.[28]

[434] (3) *A* is *directed at M*'s representational content in virtue of its standing in an *emotion-direction relation R* to *M*.

(4) *S*'s introspective processes reliably redeploy (with translation, if needed) at least part of *M*'s content into *P*'s, unless *P* constitutes *M*.

(5) At least partly in virtue of *A*'s phenomenal properties and/or *M's* representational properties, *S*'s belief production processes accurately identify *A*'s type in terms of some phenomenal and/or functional emotion concept *E*.[29]

(6) *S* comes to believe as a causal result of (1)-(5) that she is feeling *E* about *P*, and

(7) For *S* at *t*, there is no psychological state of affairs (*A*, M*, R**) that is a *relevant emotion-direction alternative* of (*A*, *M*, *R*).

Clauses (1) through (3) spell out the truth conditions of emotion-direction knowledge. They presuppose the view of directed-emotion outlined in section 1. They assume that emotional state *A* has introspectable phenomenal properties, and leave open the question of whether it also has representational properties. Clause (2) distinguishes representation

[28] For simplicity's sake, I am omitting from this analysis a small set of cases in which *M* is a non-representational phenomenal state – for instance, a novel sensation about which one might feel anxious. In such a case, the emotional feeling would be directed at (or be *about*) the instance of the phenomenal quality itself.

[29] Emotion-type concept *E* may require *A* to have a combination of phenomenal and functional properties. The concept may begin as phenomenal (direct, standing or demonstrative), and then evolve to include functional factors after it becomes apparent that *A* is *R*-related to *M*, and if *M* is perceptual, that some aspect of *M's* non-conceptual content is *P*. Finally, when expressed linguistically, *E* might be a community-based concept requiring *A* to have mostly functional properties. Not understanding the contextual flexibility of *E*'s content can lead to needless debate about the necessary conditions of emotion-type concepts.

P from representation *M* because *P,* being part of a belief*,* must be a conceptual representation of what the emotion is about, while *M*, which also represents what the emotion is about (and in many cases triggered the emotion) can have either conceptual or non-conceptual content. That is, as clause (4) will specify, *P* either *redeploys M*'s content (if *M* is conceptual), or else translates into concepts at least part of *M*'s non-conceptual content (if *M* is perceptual).

Clause (3) leaves open the question of how specific the emotion-direction relation *R* must be. However, I would argue that *R* must be more specific than, say, the mere *juxtaposition* of a somatosensory image of bodily conditions with "the mental images that initiated the cycle", as Damasio (1994, 146) describes it. Such a juxtaposition of the two "images" in working memory could be a precondition of one's introspectively recognizing that an emotion-direction relation is instantiated between the two. But juxtaposition (like association) is a symmetrical relation, and emotion-direction obviously is not. Also, the various prepositions that usually appear in directed-emotion ascriptions suggest that the relation can vary with emotion-type. For instance, 'with' in "I am in love *with* you", 'at' in "I am angry at you", and 'of' in "I am fearful *of* you" seem to signal somewhat different relations.[30] Many theorists presuppose that the relation normally "retraces" the causal relation that they take to hold between *M* and *A*, but elsewhere I provide several counterexamples to this view.[31] As I mentioned in section 1, I believe that the main function of emotion-direction is to channel *A*'s motivational properties relative to what *M* represents. Obviously, there is much work yet to be done in this area. But for present purposes, we need only note that having emotion-direction knowledge depends on an ability to reliably recognize that the relevant sort of relation holds.

[435] Clause (4) incorporates Goldman's suggestion that redeployment (and translation, as needed) are necessary for introspectively arriving at a belief about at least part of *M*'s content. However, (4) is broad enough to allow for relatively rare cases in which *S*'s emotion-direction belief *constitutes* – as opposed to *recognizes* – *M*. For instance, my emotion-direction belief that *I feel ashamed that I believe that P* might *make it the case* that I believe that *P*, even if it cannot make it the case that I feel ashamed, or that I feel ashamed about the belief that *P*.[32] So in this limited respect, my analysis is consistent with some rationalist or "constitutivist" approaches to self-knowledge.

Clause (5) refers to a complex set of possible introspective sub-processes. Which ones actually occur depends on the case. Sometimes the relationship between *A* and *M* is easily recognized and conceptualized, as when an experienced arachnophobic sees a large spider hanging a few inches from her head. In such a case, *A*'s phenomenal properties and *M*'s representational (or even early sensory) properties are likely contained in the subject's working memory almost simultaneously, monopolizing attention and instigating flight behavior. In such a case, conceptualizing *A* as *fear* and the content of *M* as *that spider* might be relatively easy, given the already strong links between the two. However, in other cases *A* might accurately be conceptualized as being of type *E* prior to *S*'s becoming aware of *M*. This might be fairly common in cases of anxiety or frustration, where conceptualizing *A* as *E* via its phenomenal properties might help to narrow down candidates for *M*, assuming that *E*'s paradigm scenarios are well-known to the subject. Similarly, when likely

[30] Of course, sometimes the relation is merely implied, as in "I love you".
[31] Herzberg (2009).
[32] I explore such cases in detail in Herzberg (2008).

candidates for *M* include percepts, conceptualizing *A* as *E* on the basis of its phenomenal properties might narrow down the aspects of *M*'s non-conceptual content that are to be translated into *P*'s conceptual content. For instance, conceptualizing *A* as *surprise* might signal to the belief formation process that it should focus on aspects of *M* that run contrary to expectations. In other sorts of case, *M* might already be conceptual but be accompanied by a pre-conceptual awareness of *A* via its phenomenal properties. For instance, upon learning that one has been laid off one's job, one might be aware of experiencing *some* type of emotion, but then have to consider a host of factors before being able to accurately conceptualize *A* as, say, a feeling of *relief* rather than of *resignation* or even of *sadness*. In sum, the conceptualizations of *A* as *E* and (at least part of) *M* as *P* can proceed either independently or interdependently, and there may be no particular order in which they must occur. But when they occur *interdependently*, the potential epistemic benefits (and dangers) of a directed-emotion's complexity become most evident.

Clause (6) focuses on the belief formation stage at which *S* becomes disposed *as a causal result of (1)-(5)* to judge that she is feeling *E about P*. It might seem redundant, but it is meant to insure that the causal process leading to the subject's judgment or self-report that she is feeling *E* about *P* is not hijacked by a mischievous neurologist or the like.

Questions of reliability come to the fore in clause (7), which formalizes the condition that emotion-direction knowledge requires the ability to discriminate one's actual directed emotion from relevant alternative triplets. My characterization of a *relevant emotion-direction alternative* (or *REDA* for short) begins by assuming that all other conditions of emotion-direction knowledge have been satisfied, and then sets out conditions that would *defeat* a claim to emotion-direction knowledge-

[436] If *S* at time *t* forms the true belief that she is feeling *E* about *P* on the basis of her introspective awareness of *A, M*, and *R*, then (*A*, M*, R**) is a *relevant emotion-direction alternative* of (*A, M, R*) for *S* at *t* if and only if

(a) *A** is *R**-related to *M**,

(b) *A**'s being *R**-related to *M** would tend to cause *S* to believe – or sustain her in believing – that she is feeling *E* about *P*, and

(c) *S*'s belief that she is feeling *E* about *P* would be *false*.

(*A*, M*, R**) might differ from (*A, M, R*) with respect to one or more of its elements, and in particularly dysfunctional cases one or more of its elements might be null-valued. Firstly, despite its happening to have produced a true belief on this occasion, *S*'s belief-forming process might be incapable of *reliably* discriminating an emotion of type *E* from other types of emotion, or even from a non-emotional type of state. For instance, it might be disposed to mistake mere irritations for angers, or feelings of fatigue for feelings of sadness.[33] An even more dysfunctional process in this regard might tend to produce beliefs

[33] For ease of exposition, I am referring here to a *general* incapacity to conceptually discriminate between emotion types on the basis of introspection. However, given Chalmers' distinction between direct, demonstrative, standing, individual-based and community-based phenomenal concepts, it seems clear that different cases may involve different (and sometimes multiple) potential incapacities.

that one feels, say, ashamed of a behavior represented by *M* when in fact one merely *conceptually* represents that behavior as shameful while emotionally feeling no shame at all. Call these sorts of cases *A-based REDA*s.

Secondly, the process might be incapable of reliably discriminating emotion-direction from other sorts of relation. For instance, it might be disposed to mistake mere simultaneity of occurrence for emotion-direction. In particularly dysfunctional cases, it might tend to conceptualize *A*s and *M*s of various types as standing in relations of emotion-direction when they are in fact unrelated in *any* relevant way. Call these *R-based REDA*s.

Finally, the process might be incapable of reliably redeploying *M*'s content into *P*'s. This seems more likely to occur in cases requiring translation between non-conceptual and conceptual content. For instance, a musically unsophisticated subject might tend to believe that she hates the *tempo* of a style of music, when in fact she hates the *meter*. In extremely dysfunctional cases, the process might "spontaneously generate" *P*'s content on no grounds whatsoever. For instance, one might tend to falsely *believe* that one is angry at one's spouse, *whatever* one happens to be angry about. Call these *M-based REDA*s.

It is, of course, an empirical question whether any of these types of *REDA*s occur frequently enough to undermine the confidence we ordinarily place in a subject's judgments or self-ascriptions of her directed emotions, but making explicit the various ways in which a *true* emotion-direction belief may yet fail to count as knowledge on process-reliabilist grounds may suggest avenues for future research into the underlying processes.

[437] 4 Concluding remarks

This process-reliabilist analysis of emotion-direction knowledge raises many questions that can be answered only by further research, both conceptual and empirical. There are of course questions related to the underlying "hybrid" view of self-awareness discussed in section 2. If phenomenal realism is the best theory of self-awareness of one's sensations, emotional feelings, and other states with salient phenomenal qualities, might it also be extensible to states with less salient phenomenal qualities, such as desires, intentions, and even some beliefs? Or is it more plausible to analyze introspection of most propositional attitude types in terms of Goldman's quasi-perceptual view? Might there not be roles for *both* types of self-awareness, operating independently or in concert, perhaps along with a small role for rational constitutivism in cases of meta-beliefs about some first-order beliefs?

The epistemological analysis outlined in section 3 is quite schematic as it stands, and obviously needs to be filled in by considering how well it applies to different sorts of possible cases. This is primarily an area for conceptual analysis. But there are also several issues that call out for empirical investigation. One concerns the nature and variability of the emotion-direction relation. Do the different prepositions used in ascriptions of various sorts of directed emotion signal different sorts of emotion-direction relation? How do different languages compare in this regard? Might the relation sub-types correspond to the distinctive motivational tendencies that are cross-culturally associated with particular emotion types? These questions bear on a related but independently significant question concerning the relative frequencies of the various kinds of relevant emotion-direction alternative. For if there are different sub-types of emotion-direction relation, and one's

belief-forming process must be able to discriminate one sub-type from another in order to reliably form a true emotion-direction belief, then *R*-based REDAs might be more prevalent than one might suppose. Finally, an even more central empirical question concerns the relative reliabilities of the different routes to forming an emotion-direction belief, assuming that the sub-processes can occur in any order. For instance, carefully constructed experiments might reveal whether it is easier to form a true belief that one has an emotion of type *E* about *P* when *A* is conceptualized prior to *M*'s being conceptualized or *vice versa*. My hope is that the current analysis – or some descendant of it – will raise many more questions worth investigating.

Acknowledgements- I wish to thank Brie Gertler for her comments on an early version of this paper, presented at the Winter 2013 meeting of the American Philosophical Association, Central Division. I also wish to thank an anonymous referee from *Acta Analytica*, who made several helpful suggestions. Completion of this paper was facilitated by a sabbatical research grant from the Office of Grants and Faculty Development at the University of Wisconsin, Oshkosh.

References

Bilgrami, A. 1998. "Self-Knowledge and Resentment." In *Knowing Our Own Minds*, edited by Crispen Wright et al., 207-242. Oxford: Clarendon Press.

Bonjour, L. 1976. "The Coherence Theory of Empirical Knowledge." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Vol. 30, No. 5: 281-312

Burge, T. 1996. "Our Entitlement to Self-Knowledge", *Proceedings of the Aristotelian Society*, 96: 91-116.

Burge, T. 2010. *The Origins of Objectivity*. Oxford: Clarendon.

[438] Chalmers, D. 2003. "The Content and Epistemology of Phenomenal Belief." In *Consciousness: new philosophical perspectives*, edited by Quentin Smith and Aleksandar Jokic, 220-272. Oxford: Clarendon Press.

Chisholm, R. 1957. Perceiving: A Philosophical Study. Ithaca: Cornell University Press.

Costall, A. 2013. "Introspection and the Myth of Methodological Behaviorism." In *Self-Observation in the Social Sciences*, edited by Joshua W. Clegg, 67-80. New Brunswick, N.J: Transaction Publishers.

Dretske, F. 1971. "Conclusive Reasons," *Australasian Journal of Philosophy*, 49: 1–22.

Ekman, P. 1999. "Basic Emotions." In *Handbook of Cognition and Emotion,* edited by T. Dalgleish and M. Power, 45-60. Sussex, U.K.: John Wiley & Sons, Ltd.

Damasio, A. 1994. *Descartes' error: Emotion, reason and the human brain*. New York: Avon Books.

Gertler, B. 2001. "Introspecting Phenomenal States", *Philosophy and Phenomenological Research*, Vol. LXIII, No. 2: 305-328.

Gertler, B. 2011. *Self-Knowledge*. London: Routledge.

Goldman, A. 1976. "Discrimination and Perceptual Knowledge," *Journal of Philosophy*, 73: 771–791.

Goldman, A. 1979. "What Is Justified Belief?" in G. Pappas (ed.), *Justification and Knowledge*, Dordrecht: Reidel. Reprinted in A. Goldman, *Liaisons: Philosophy Meets the Cognitive and Social Sciences*, Cambridge, MA: MIT Press (1992).

Goldman, A. 1986. *Epistemology and Cognition*, Cambridge: Harvard University Press.

Goldman, A. 2006. *Simulating Minds*, New York: Oxford University Press.

Goldman, A. "Reliabilism", *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/reliabilism/>.

Herzberg, L. (under review). "Emotion, Perception, and Significant Organism-Environment Relations".

Herzberg, L. 2009. "Direction, Causation, and Appraisal Theories of Emotion", *Philosophical Psychology*, Vol. 22, No. 2.

Herzberg, L. 2008. "Constitutivism, Belief, and Emotion", *dialectica*, Vol. 52 Fasc. 4.

James, W. 1918, 1981. *The Principles of Psychology, Vol. II*, Cambridge: Harvard University Press.

Lazarus, R. S. 1991. *Emotion and adaptation*. New York: Oxford University Press.

Lazarus, R. S. 1999. "The Cognition-Emotion Debate: A Bit of History." In *Handbook of Cognition and Emotion,* edited by T. Dalgleish and M. Power, 3-20. Sussex, U.K.: John Wiley & Sons, Ltd.

Ledoux, J. 1996. *The Emotional Brain*, New York : Simon & Schuster.

Prinz, J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*, New York: Oxford University Press.

Nichols, S. and Stich, S. 2001. "How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness." In *Consciousness: new philosophical perspectives*, edited by Quentin Smith and Aleksandar Jokic, 157-200. Oxford: Clarendon Press.

Nisbett, R. E., & Wilson, T. D. 1977. "Telling more than we can know: Verbal reports on mental processes." *Psychological Review*, 84, 231-259.

Nozick, R. 1981. *Philosophical Explanations*, Cambridge, MA: Harvard University Press.

Russell, B. 1910. "Knowledge by Acquaintance and Knowledge by Description." *Proceedings of the Aristotelian Society*, 11: 108-128.

Shoemaker, S. 1975. "Functionalism and Qualia." *Philosophical Studies*, 27: 291-315.

Shoemaker, S. 1996. *The First-Person Perspective and Other Essays*. New York: Cambridge University Press.

Solomon, R. 1984. "Emotions and Choice." In *What Is An Emotion*, edited by C. Calhoun

and R. Solomon, 305-326. New York: Oxford University Press.

Sosa, E. 1996. "Postscript to 'Proper Functionalism and Virtue Epistemology'," in J. L. Kvanvig (ed.), *Warrant in Contemporary Epistemology*, Lanham, MD: Rowman & Littlefield.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Wright, C., Smith, B., and Macdonald, C. eds., 1998. *Knowing Our Own Minds*, Oxford: Clarendon Press.