Draft of December 2022—please do not cite without permission.

Eliminativism and Reading One's Own Mind

T. Parent (Nazarbayev University)
nontology@gmail.com

1. Introduction

Since Descartes' *Meditations*, many have been convinced that an individual can know merely by reflection that she is the locus of contentful mental states. For instance, Cartesian reflection might suggest that I am currently judging that water is wet, or that it appears to me that I am now sitting by the fire. The claim, then, is that such reflections make it (at least) likely that I have that judgment or mental appearing, at least at the time. Almost no one thinks that the deliverances of reflection are infallible, indubitable, or incorrigible—yet philosophers are often persuaded that reflection can justify various self-ascriptions of mental representations.

However, if it can be armchair-justified that I am now judging that water is wet, then it is armchair-justified that at least one judgment exists. In which case, it is armchair-justified that eliminativism is false. Some advocates of introspection try to subvert this consequence, for eliminativism is apparently a substantive, empirical matter which could not be discredited merely from the armchair (see Bernecker 1998; Dretske 2003, 2004).[1] In stark contrast, however, some philosophers think we *are* justified in rejecting eliminativism by such means. For example, Lycan (2019) declares "We know there are propositional attitudes because we introspect them in ourselves… [This and other arguments] can be fleshed out into powerful

[1] In my (2017) book, I myself defended a kind of infallibility about self-ascriptions, and yet dodged the anti-eliminativist consequence of this in ch. 8. The tactic there was to defend a view known as mental fictionalism—a view where a true self-ascription would be true merely in the sense of "true according to the fiction." (The truth of a self-ascription would then be much like the "truth" that Sherlock Holmes lives in London.) A true self-ascription would then *not* have a real folk psychological state as its truth-maker; hence, a "true" self-ascription justified by introspection would not falsify eliminativism. However, mental fictionalism is admittedly a minority view with some serious objections to contend with (for an overview, see Demeter et al. 2022.) Even so, at the end, I shall say a word about how the account here naturally dovetails with mental ficitonalism.

defenses of folk psychology" (p. 39). Baker (2013) argues in a similar albeit more programmatic manner, by rejecting "naturalism" on the grounds that it is incompatible with a kind of Cartesian mental life.

My aim here is to oppose such anti-eliminativists. To do so, I shall first review two arguments which could bolster their position, but then use a suggestion from Alex Rosenberg (2011, 2022) on why such arguments are non-demonstrative. Rosenberg's suggestion, however, leaves some questions unanswered. It especially creates a need for an *error theory* on how Cartesians are mislead into confidently self-ascribing various mental representations. The paper then offers such an error theory. The error theory will not only show how eliminativists can explain Cartesian self-ascriptive tendencies, but ultimately, it will also strengthen the eliminativist position against Cartesian arguments.

Let me clarify up front, however, that I am not an eliminativist; I am rather agnostic on whether eliminativism is true. But such agnosticism is sufficient to motivate the issue: If introspection can justify the *bona fide* existence of mental representations, that already forces a commitment against eliminativism. So my kind of agnosticism requires resisting the Cartesian arguments.

2. Preliminary Clarifications.

"Eliminativism" is here understood as the rejection of *mental contents*, and thus, *mental representations*, assuming the latter are essentially "vehicles" of content. This entails the rejection of *propositional attitudes* (beliefs, desires, intentions, etc.) including the components of such attitudes, to wit, concepts, thoughts, and attitudes directed toward those thoughts (the believing attitude, the desiring attitude, etc.). The present eliminativism is hence fairly wide-ranging, but I take to be in line with the classic eliminativist views from P.M.

Churchland (1981, 1989), P.S. Churchland (1983, 1994) and Stich (1983, 1991).[2] Even so, eliminativism as construed here does *not* reject other mentalistic phenomena such as bare sensations, "raw feels," etc. This will be important later.

Again, the main question will be whether armchair reflection can falsify eliminativism; however, it proves useful to first examine briefly whether eliminativism falsifies *itself* (a worry voiced by Baker 1987, ch. 4, and Boghossian 1990a, b, among others). Briefly, one might speak of what an eliminativist believes or claims, etc., yet this may seem odd insofar as beliefs, claims, etc., are usually seen as having representational contents. We therefore need a different way to speak about such things.

Accordingly, I use the term 'upholds' in a quasi-technical way as follows. Let "*p*" be any declarative sentence of the language in which a human organism *S* is competent—that is, a language in which *S* has received sufficient training or conditioning within the relevant linguistic community. Then:

(U) *S* upholds *p* iff *S* is disposed to utter "*p*" under Normal conditions.

A few clarifications. First, the identity- and existence-conditions for "dispositions" remain unclear (see, e.g., Armstrong, Martin, & Place 1996). But I must pass over this issue in what follows. Talk of dispositions is at least precise enough for many scientific purposes, as when talks about salt having a disposition to dissolve when stirred in water.

Second, "Normal conditions" for an utterance are not necessarily the most common conditions for the utterance. Rather, following Millikan (1984; 2005, etc.), Normal conditions for a linguistic behavior are those which explain the continued proliferation of the behavior within the linguistic community. This is analogous to adaptationist explanations for biological traits. What explains the continued production of sperm is that sperm fertilize eggs,

[2] One occasionally sees these writers using the term 'representation' and other folk psychological terms, apparently in earnest. But I take it these are convenient stand-ins for more complicated, non-intentional descriptions, or they should be interpreted as non-intentional as they stand. Thus, under one construal, a calculator screen has "representations" of numbers, but such pixel-aggregates do not exhibit *original* or *underived* intentionality.

even though statistically speaking, it is far more common for a sperm to not fertilize an egg. Similarly, what explains the continued use of 'A cat is on the mat' (among English speakers) is that it communicates information. This is just to make the unremarkable point that, in the Normality condition, an utterance of 'A cat is on the mat' allows the audience to infer that a cat is on the mat—much like how a thermometer's reading allows one to infer the temperature. (Its Normal use is thus not a lie, nor is it an actor's line in a play, nor is it a performance-error, etc.) For more on such explanations of linguistic usage, I would refer the reader to Millikan—yet unlike Millikan, the eliminativist should of course not regard this as explaining the *meaning* or *content* of the utterance. The eliminativist may co-opt Millikan-like explanations, not as concerning the proliferations of utterances with specific meanings, but rather the proliferation of the utterances as such. This sort of anti-semantical angle on Millikan is detailed further in Hutto & Myin (2013, ch. 4).

For short, let us say that *S affirms* "*p*" iff *S* utters "*p*" under Normal conditions. Take heed that affirming here need not imply the occurrence of a special folk psychological "attitude." We can instead regard an affirming of the sentence "*p*" as indicating that the person is disposed to use "*p*" in inferences of various sorts, where these inferences can be understood as purely syntactical transformations, free of any semantic trappings. On this approach, "affirming" a sentence basically amounts to producing a certain kind of syntactic string, which is then available for use as input into a variety of computational processes. This is akin to a computer "affirming" a piece of code by tokening the code for "consumption" within a purely syntactic engine.

As a final preliminary, observe that the eliminativist as envisioned here operates with a deflationary notion of truth, as opposed to a "thick," metaphysical notion of truth (see, e.g., Field 1994). This accords with Rosenberg's (2022) eliminativism, and it avoids the folk psychological implication that a true sentence has a content that somehow "matches" the

ding-an-sich. Our eliminativist instead claims merely that affirming the truth of a sentence is computationally equivalent to affirming the sentence itself. Nothing more is added. Granted, we might feel the metaphysical urge to say "yes, but what is it that *makes* a sentence true?" But this is an urge that deflationisms resists. The only questions we answer about the use of sentences are questions about their inferential role, understood purely computationally.[3]

3. Two Arguments for Self-Mind Reading

We can now clarify that the primary issue consists in whether armchair reflection justifies the thesis of *self-mindreading:*

(SMR) Some self-ascriptions of contentful mental states are true.

Again, if (SMR) is armchair-justified, then we are armchair-justified in affirming the existence of mental content, contra eliminativism. But why think that (SMR) can be so justified?

Well, remember Descartes. In *Meditation 2*, he claimed the following to be an utterly foundational piece of knowledge:

(D) I doubt.

Speaking not as a Descartes-scholar, Descartes justified (D) by what has been called a "diagonal argument."[4] Briefly, if I doubt (D), then trivially, I am doubting. But my doubting is exactly what (D) claims. So from my doubting (D), it patently follows that (D) is true, indicating that (D) cannot rationally be doubted. Rational certainty about (D) is thereby achieved, meaning that (D) is true. So, (D) is a true self-ascription; hence, (SMR) is true.[5]

---

[3] Similarly, when I spoke of an utterance of "*p*" being used to "communicate information," this was not said in a folk psychological spirit. Rather, it was meant in the sense of allowing an audience to *make an inference* as to whether *p* (i.e., as to whether "*p*" is true, where 'true' should be understood in the deflationary way).

[4] See Slezak (1983).

[5] Perhaps (D) itself is not an ascription of a *propositional* attitude, since there no proposition that is explicitly mentioned as the target of the doubt. But no matter; one could see (D) as shorthand for "I doubt that *p*," for some proposition *p*. The diagonal argument would then show that such a self-ascription must be true, for at least one replacement of '*p*'.

But what the diagonal argument shows, in the first instance, is that *rational doubt* in (D) is impossible. This, moreover, is fully compatible with eliminativism insofar as "rational doubting" is a propositional attitude which is already under ban. It is a non-sequitur, moreover, to infer thereby that there must be *rational certainty* toward (D). For such certitude would also be a propositional attitude that the eliminativist has eliminated.

Basically, the diagonal argument suggests that *if* I begin in doubt regarding (D), I am unavoidably led to certainty. But the eliminativist denies that I can literally enter a state of *doubt*; hence, she can resist that are thereby forced into a state of certainty. Notice, moreover, that the truth of (D) is concluded only after a state of certainty about (D) is alleged. So if the intermediary step of certainty is withheld, the argument does not get us to the truth of (D).

(Aside: An eliminativist can uphold that we do something functionally akin to "doubting." Under some circumstances, we refrain from upholding statements which we would ordinarily uphold. But this is not to be understood as anything folk psychological; it is rather understood by dispositions *not* to affirm specific sentences in specific contexts, ones which might otherwise be affirmed.)

The key point is that the eliminativist is within her rights to accept Descartes' diagonal argument as far as it goes, by emphasizing that it only goes so far: It establishes only the impossibility of rationally doubting (D), an impossibility which is quite consistent with eliminativism. Even so, I suspect that the diagonal argument is not what philosophers have found so compelling in the early part of the *Meditations*. Philosophers like Lycan and Baker do not craft some piece of philosophical argument to justify (SMR); rather, they directly appeal to *introspective appearances* to justify that propositional attitudes exist.

Recall that when doubting the external world, Descartes withheld his judgment that:

(F) I am now sitting by the fire.

Descartes nonetheless was certain that (F) *appears* to be true. The existence of the appearance thus looks justified, and such would be a mental representation—one that represents him as sitting by the fire. And so, if it is justified that the representation exists, then we have an armchair-justification of anti-eliminativism.

However, this looks question-begging, for the argument *starts* with an introspective appearance, and then deduces the falsity of eliminativism. But eliminativists will reject the starting point; there are no mental representations of (F), and that includes introspective appearances of (F). Notwithstanding, one might feel it is undeniable that there is some sort of internal phenomenology at hand. Granted, we can resist construing the phenomenology as *representational*, much less representational of (F). It might instead be seen as a nonconceptual internal sensorium, much like the newborn's "blooming, buzzing confusion" as described by William James. It is just an unassorted array of colors, sounds, etc. without any folk-representational features. (These colors, sounds, etc*., a fortiori* do not represent those very sensory properties. The splashes of colors, sounds, etc., are mere events, without any semantic features.)

Even so, the internal sensorium would still cause dispositions to affirm 'It seems that I am now sitting by the fire'. Certain segments of the phenomenological stream will make one inclined to affirm that kind of syntactic string. Can we take this sensorium-induced inclination as justifying that there is an *appearance*, as a kind of mental representation?

A positive answer would be tendentious for more than one reason.[6] But I want to focus on a line from Rosenberg (op. cits.).[7] One notable aspect of Rosenberg's general platform is that, despite his eliminativism about representational states, he is *not* an

[6] I must at least footnote that a positive answer would flouts Sellars' (1963) dictum that unconceptualized sensoria could not be justifiers. However, I am not at leisure to explore this issue here.

[7] Rosenberg is not responding to a Carteisan argument but rather an argument from Horgan & Tienson (2002). That argument instead emphasizes the introspectable difference between understanding vs. not understanding a sentence like 'Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo'. But this is the same sort of introspective appeal to justify the existence of a folk psychological state; hence, Rosenberg's reply applies straight-forwardly to the Cartesian argument as well.

eliminativist about phenomenology.[8] This makes his view unlike the eliminativism of P.S. Churchland (1994), for example. Rosenberg is therefore in a unique position to argue as follows:

> Eliminativists can admit that we are all subject to the phenomenological illusion that thought has intentional content, that the illusion is powerful, and can at best only be temporarily counteracted or suspended…[Even so,] thought is completely different from what conscious experience led us to suppose. (p. 14)

In an attending footnote, Rosenberg confirms that an "illusion" should not be seen as a representational state, but as more like a (non-representational) state which causes dispositions for various behaviors—including, presumably, self-ascriptive behaviors. Similarly, his talk of "thought" concerns internal processes that do not involve intentionality, much like the serial/parallel computational states in a laptop or a neural net.

Thus, I take his point to be that certain linguistic behaviors result from nonconceptualized internal sensations (Rosenberg's term is 'mental imagery'). Foveating an active fireplace causes sensoria which, in turn, lead to affirmations of 'It appears to me that I am sitting by the fire'. Yet for the eliminativist, this amounts to an "illusion" insofar as the affirmations we are caused to make are false. There is no *appearance* of sitting by the fire, even though some types of internal phenomenological states may cause us to affirm as much.

In response to such "internal world skepticism," there are two rejoinders to consider, although the first will not occupy us as much as the second. The first rejoinder appeals to commonsense or commonsense practices with introspection in order to bolster its credentials. The claim is that, in ordinary circumstances, our internally-prompted affirmations are worthy of trust in much the same way as those prompted by sensations from the outside world. Of course, external-world sensations can mislead us, as when a rectilinear tower in the distance

[8] The question of how to naturalize phenomenology of course remains, but this is bracketed in the present discussion.

causes us to affirm that it is cylindrical. Notwithstanding, we are usually justified in trusting the guidance of external-world sensations, unless one has significant evidence to the contrary (e.g., evidence that one has ingested a hallucinogenic). The suggestion then is that an analogous point holds with introspection. Thus, although phenomenology can give way to illusion, here too our affirmations are justifiably guided by phenomenology, absent any defeaters. For instance, the sensation of pain in the foot is normally (although not invariably) indicative that something is neuro-physiologically awry in the foot. This epistemic analogy between introspection and sensation has been defended previously by philosophers, such as in introspection arguments for the existence of free will (see Lehrer 1960).[9]

Rosenberg (2011), however, responds with a catalogue of cases where introspection indeed misleads us. So it seems that even if we usually trust introspection, perhaps out of practical necessity, it is not always a legitimate practice from a purely epistemic angle. Indeed, Rosenberg uses freewill as a case in point, alluding to Soon et al. (2008), in which an fMRI can predict a subject's "choice" up to 11 seconds before the subject reports the sense of choosing. This suggests that the feeling of choice-making is *not* keyed to the occurrence of a choice (11 seconds is a long time, neurologically speaking). Yet the defender of introspection may reply that this just repeats the earlier point that internal sensations are deceptive in many cases—a point which has already been conceded. And she may reiterate that we nonetheless *do* trust internal sensations to guide our affirmations on many occasions, as when we report to the doctor where it hurts, or when we identify in ourselves a preference for one item on the menu over another. What's more, the fact remains that we often seem correct or at least reasonable in our affirmations, even knowing about the frequency of error.

On the other hand, it is one thing to trust introspection on a mundane question of where it hurts, and another thing to trust introspection on the truth of a substantive

[9] An important difference, however, is that libertarians tend to speak of the veridicality of perceptual and introspective *appearances* (understood as mental representations). But again, here we are speaking of unconceptualized sensations that lead to (dispositions to) affirming various sentences.

philosophical thesis, such as the existence of free will or the existence of propositional attitudes. Introspection is evolutionarily adapted to discern the truth on matters of day-to-day survival and health—that would explain the reliability of introspectively-based affirmations of 'It hurts here' or 'I am hungry'. But there is no similar adaptive reason to say that introspection is honed to spot the truth on the abstract philosophical questions. Given that, the appeal to commonsense tendencies may ring hollow—and we might hope for more from the Cartesian anti-eliminativist.

4. An Error Theory for Cartesians

However, a true Cartesian would not have allowed Rosenberg to push this far the possibility of introspective illusion. The basic intuition is that, if it *appears* that there is an appearance of sitting by the fire, then there really is such an appearance. There is no such thing as a merely apparent appearance that *p*, for an apparent appearance that *p* just is an appearance that *p*.[10] Or so the Cartesian claims.

As formulated, however, the Cartesian is simply helping herself to the folk psychological notion of appearing. So the point is again question-begging. Nonetheless, I respect that the Cartesian intuition is real—it is very tempting to affirm the existence of appearances, understood as mental representations. So if this Cartesian intuition is mistaken, it would be nice if we had an *error theory* for how nonconceptualized phenomenology causes one to uphold mistakenly the existence of mental representations. Providing such an error theory is the aim of this section.

The basic phenomenon to be explained, if we describe it in non-question begging terms, is as follows.

[10] Márton & Tőzsér (2013) champion this point in their defense of folk psychology.

> *Explanandum*. Certain segments of the (unconceptualized) sensorium lead to confident self-ascriptions of mental appearances, e.g., affirmations of 'It appears to me that I am now sitting by the fire'.

A desideratum on an adequate eliminativist explanation, of course, is one that does not assume the reality of any mental representations, including appearances.

The explanandum is a certain kind of linguistic behavior (albeit one with a phenomenological cause)—and when it comes to linguistic behavior, we can hardly do better than start with neural net or "connectionist" models for language use. Indeed, the suggestion shall be that affirming 'It appears to me as if I am sitting by the fire' is explained in large part by one's linguistic training with the first-order statement 'I am sitting by the fire.' And currently, neural nets provide arguably the best explanation of such a thing.

Since these models are widely discussed, I shall be brief in my description, relying on other sources to supply the details.[11] Essentially, a network can be trained to output specific linguistic items on specific kinds of inputs, where the inputs can be seen as analogous to the inputs we receive from our sensory organs. Thus, when fed as inputs 'meowing', 'purring', and 'long whiskers', the network can be trained to output 'cat' by adjusting weights on connections between input- and output-nodes, often with the help of a back-propagation algorithm. Suppose, for instance, that when given 'purring' and 'long whiskers' as input, the network initially outputs 'dogs' rather than 'cats'. Back-propagation would flag the error, and then readjust the weights of the connections to make this input-output pair less likely in future cycles. The result is increased accuracy.

---

[11] The *locus classicus* is the two-volume Rumelhart & McClelland (1986), but usuful summaries of the evidence are found in Christiansen & Charter (1999) as well as in Rohde & Plaut (2004). Some eliminativists also present good overviews of the issues, e.g., P.S. Churchland & Sejnowski (1989) and Ramsey (2007). Also, the papers on connectionism in Haugeland's (2000) classic anthology remain illuminating (including a contribution by Rumelhart). More recently, Clark (2013) provides an accessible introduction to developments concerning predictive coding. Finally, Buckner (forthcoming) introduces to a philosophical audience "deep learning" as exemplified by AlphaZero (cf. Silver et al. 2018).

The explanation of linguistic behavior, based solely on neural network models, seems quite congenial to eliminativism. Paul Churchland (2012) makes this point when he writes:

> The cognitive achievement here portrayed—a particular form of visuo-motor coordination—is quite evidently an acquired *skill*. And on that account, there may seem little that is distinctly *cognitive* about it, in the sense that is likely to interest an epistemologist. (p. 49).

Churchland bolsters this further by stressing "the poverty of any sharp or fundamental distinction between knowledge *how* and knowledge *that*…The major fault lines that divide our neurobiological [states] lie quite elsewhere, and largely cross-classify those embodied in Folk Psychology" (ibid.). But while Churchland's remark is well-taken, it might be better to say that neural nets suggest that *all* knowledge is a knowledge-how which does not rest on any folk psychological states.

Admittedly, many balk at neural nets as modeling human neurological processing. It is common to doubt that anything like backpropagation is neuro-biologically realized. The reason is that "this seems to require the rapid transmission of information backwards along the axon, that is, antidromically from each of its synapses. It seems highly unlikely that this happens in the brain" (Crick 1989, p. 130). One natural reply is that perhaps the brain receives feedback on its successes and errors is from the wider linguistic community: The community can be seen as flagging successes and errors as such, feeding this information back into one's wetware—this in turn, might cause the system to readjust the weights of connections and the biases of nodes.

In fact, adding social cognition to the mix is something I would embrace.[12] But it does not address the basic problem with backpropagation. Backprop becomes neurologically implausible not because of how error is identified. Rather, the problem lies rather in how an

[12] One of the foundational works here is Hutchins (1995). A more contemporary discussion is Huebner (2014). An excellent article-length review of the history and issues is Theiner (2014).

artificial network *uses* this information to readjust the weights and biases. The network basically "reverse engineers" a more accurate output by determining what state, in the immediately preceding hidden layer of nodes, would more likely yield the desired output. And then, it determines what would more likely produce that hidden state by determining what the hidden layer prior to *that* should look like. And so on. This action of working backwards through the net is what requires the bidirectionality of the connections and is what makes it different from our neural anatomy.

I have hope, however, that a process *functionally like* backprop is neuro-biologically plausible. The weights and biases (or some functional equivalent[13]) might be readjusted by some process which somehow uses the error-data in a comparably effective manner. Indeed, there are more recent attempts to find neurologically plausible means for achieving what backprop does, including O'Reilly's (1996) algorithm that computes an error value not based on the error value of each individual node. Other examples include Lillicrap et al. (2016) and Bengio et al. (2017). These alternatives also differ significantly from backpropogation *per se*, but they fit the functional description of "backprop;" indeed, these authors basically advertise their alternatives in this manner. And achieving the function of backprop is all that we need to hope for.

Suppose, then, that neural nets explain to a respectable approximation our first-order affirming-behaviors—suppose that something functionally like a trained neural network explains why foveating an active fireplace causes a disposition to affirm 'I am sitting near the fire', and why foveating rainfall outside does not generate such a disposition. Then, the hypothesis would be that the network is also trained to conform to the following rule:

---

[13] Another complication is that the activation of a node in an artificial network is continuous, whereas neuronal activation is binary (either a neuron fires or it does not). Here too, I have hope that something functionally akin to continuous values for artificial nodes might be biologically realized, e.g., perhaps the number of times that a neuron fires (and/or the rate at which it fires) might perform the job of weighted activation in a neural net. (Notwithstanding, there are further objections to neural nets as models of actual biological brains, and naturally, I cannot address all of them here.)

(Rule) Given as input a state which disposes me to affirm "*p*," output a state that disposes me to affirm "It appears to me that *p*".

(Unless otherwise specified, "states" are henceforth internal states of the network, not external states-of-affairs that the network perceives or the like.) This is a second-order rule—a rule that generates a self-ascriptive-disposition when given a state realizing a first-order affirming-disposition. Conforming to this rule requires the capacity for the network to detect its own dispositional states; consequently, the theory posits that the network has some kind of self-scanning mechanism. This, by the way, is sometimes proposed as just what constitutes introspection (cf. Armstrong 1994; Lycan 1996, ch. 2).[14] But naturally, "detecting" is not here understood as a folk psychological representational process—it is instead seen as a purely causal, covariational process, much like how a metal detector works. A signal can co-vary with the presence of a linguistic dispositional state, and it is that sort of signal which is then be fed into the network, in order to effect conformity with (Rule).

It should not be thought that (Rule) exhausts the use of appearance-talk in English; in particular, there are cases where talk of an "appearance that *p*" is not keyed to any introspective act, but rather simply expresses uncertainty about *p*. (There are likely other uses of appearance-talk as well.) But (Rule) at least indicates one use of appearance-talk which is especially relevant, insofar as it bears on (SMR).

Note well that detection of a dispositional state will not mean detection of that state *as* a "dispositional state" (under that description or what have you). One could instead suppose that a state disposing one to affirm 'I am now sitting by the fire' comes with a distinctive sort of phenomenology. Indeed, one could assume that a state disposing one to affirm "*p*" has exactly the sort of phenomenology which Horgan & Tienson (2002) claim for the mental

[14] I would also suggest that the human organism can detect its disposition to affirm "*p*" by actually witnessing itself affirm "*p*". (Clearly, if it affirms "*p*," then it has a disposition to do so.) This alternate form of detection does not depend on introspection as much as "extrospection;" cf. [author] (2017) chs. 10 and 11.

representation that *p*.[15] The difference, of course, would be that the phenomenology indicates *only* a state realizing a linguistic disposition and not some folk psychological state.

Alternatively, perhaps introspection detects a disposition by detecting some more basic, sub-personal state-of-affairs. In which case, perhaps the phenomenology "attaches" not to the dispositional state itself, but rather to the *introspecting* of the state. Or, perhaps the dispositional state and the introspective act each enjoy their own unique phenomenologies.

The eliminativist need not be wedded to any of these particulars. The point is just that introspection can detect the state disposing one to affirm "*p*," albeit not under that "mode of presentation." Put differently, the process would not usually cause one to affirm something so philosophically loaded as 'I am in a state disposing me to affirm "*p*".' Rather, when introspection detects a state disposing one to affirm 'I am sitting by the fire', the network enters a state disposing one to affirm "It appears (/to me) that I am sitting by the fire."[16]

For the eliminativist, moreover, this is the key to why we are mislead into embracing mental appearances. Being trained in conformity with (Rule) means that introspecting certain states leads us to affirm sentences with appearing-vocabulary. Yet what actually triggers such affirmations is not any mental appearings, but rather just a disposition for the first-order affirmation. *We thus mistake a mere disposition to affirm "p" for a mental appearing of the state-of-affairs that p.*

5. Further Development of the Error Theory

The question may now arise: What difference is there really between a state which disposes one to affirm "*p*" and a mental appearance that *p*? Granted, one might have an

[15] See also discussions of "cognitive phenomenology" such as Pitt (2009) and Siewart (2011). But again, our eliminativist would not endorse that such phenomenology is *cognitive* in the folk psychological sense. Rather, they would just endorse that the kind of phenomenology discussed by these authors is actually associated with a mere affirmative-disposition rather than any kind of mental representation.

[16] Henceforth, I drop 'to me' in a sentence "It appears to me that *p*," but it should be remembered that 'it appears that *p*' is still an ascription *to oneself* of an appearance, in a way that makes (SMR) relevant.

affirming-disposition without a *sensory* appearance that *p*—I might be disposed to affirm '7+5=12' without any visual or other sensory phenomenology. But one could still speak of an "intellectual" phenomenology in that case. So the question remains: What difference is there between a state disposing one to affirm "*p*" and a mental appearance that *p*?

My answer, although I cannot defend this here, would be that a mental appearing comes with *built-in norms of correctness*; there is something inherent to an appearing which makes it accurate or inaccurate vis-à-vis the world. Not so with a disposition—it is either triggered or not. A triggered affirmation, of course, can be corrected by one's linguistic community, but that too is something which "just happens," and may or may not be correct *as* a correction from the intuitive point of view. A correction itself can be corrected as well, but this also is something that just happens according our eliminativist. And so on, indefinitely. (For more on the indefinite regress of corrections, see [author] 2017, section 0.8.)

Although this is sketchy, I suspect folk psychologists would agree with the essential point, viz., that a state realizing a linguistic disposition is not ipso facto a *mental appearing*. However, there is a nearby issue worth debating. Briefly, if speakers are trained in conformity with (Rule), then it suggests that it is *correct* in English to affirm the sentence "It appears that *p*" in the relevant circumstances. That is, (Rule) suggests that it is a principle of English usage that:

(*) "It appears that *p*" is true if the speaker detects an internal state disposing her to affirm "*p*".

If English speakers are trained by other English speakers to use the 'it appears that' operator in the circumstance specified by (Rule), then intuitively, it is right to do so. That is, detecting a state disposing me to affirm "*p*" is a condition on which "It appears that *p*" is true, as determined by English linguistic conventions. This is not to say that the dispositional state is

the *same* as a mental appearance; it is rather to say that detecting the state is a sufficient condition for the truth of the appearance-ascription.

If (*) is correct, then the account which invokes (Rule) to explain ascription-behavior would not amount to an *error* theory. Rather, it would be a theory that portrays affirmations of "It appears that *p*" as true when produced in conformity with (Rule).

In reply, the eliminativist might grant that (*) is believed (or rather, "upheld") by the community; nonetheless, this does not imply that (*) is factually correct (cf. Churchland 1989, ch. 1). One might compare this situation with a (hypothetical) scenario where ancient Greeks upheld the following:

(Z) "Zeus is near" is true if thunder is audible.

If the Greek community subscribed to (Z), then where thunder is audible, there would be wide agreement that Zeus is near. But this would not mean Zeus was ever near!

This can be puzzling from one angle, since one would think that Greeks would be free to assign truth-conditions to their sentences *ad libitum*, as long as they are all agreed on which sentence has which truth-condition. From this perspective, a principle like (Z) then may seem to be "true by convention" when affirmed upon hearing thunder. But of course, it would not be true, given the non-existence of Zeus.

In the same way, a principle like (*) may seem to have an air of "truth by convention" about it. We get to say how an English sentence is used, and so, community-wide agreement on (*) may seem to simply *legislate* a truth-condition for a sentence of the form "It appears that *p*." But in light of examples like (Z), this is not at all obvious. Community-wide agreement on certain principles may simply reflect a community-wide mistake. So even though the eliminativist agrees that English speakers uphold (*), she does not have to grant that (*) is at all true.

There is one further matter which needs addressing. Consider the case where '*p*' in (Rule) is replaced with a sentence that already uses the 'It appears that' operator. For example, consider the following instance of (Rule):

(Rule′) Given as input a state disposing me to affirm "It appears that I am sitting by the fire," output a state disposing me to affirm "It appears that it appears that I am sitting by the fire".

How plausible is it that our linguistic dispositions conform to this? Outside of the epistemology room, one never hears speakers iterating the 'it appears that' operator in succession. Perhaps the influence of (Rule′) explains why epistemologists speak this way, but it does not seem very faithful to folk usage.

As a fix, the eliminativist could suggest that (Rule) is usually restricted to cases where '*p*' is replaceable only by sentences that lack the 'it appears that' operator. And for the cases that remain, the suggestion might be that the folk are not trained to (Rule) but rather to:

(Rule2) Given as input a state disposing me to affirm "It appears that *p*," output a state disposing me to affirm "It appears that *p*".

This is in fact an "identity" operation—the inputted state and the outputted state are the same state. So the hypothesis is that when introspection detects the disposition for the appearance-ascription, our training in the language determines that the very same appearance-ascription would be appropriate.

Why think that the folk are trained to (Rule2)? Part of the motivation is the idea that the folk do not bother to distinguish talk of appearances from talk of apparent appearances. They just talk of appearances. But another motivator is that (Rule2) allows the eliminativist to explain the intuition that one enjoys a kind of Cartesian certainty about one's own mental appearances. Consider: Just as (Rule) suggests that the folk uphold (*), (Rule2) suggests that the folk uphold (**):

(**) "It appears that *p*" is true if the speaker detects an internal state disposing her to affirm "It appears that *p*".

Again, the fact that the folk generally uphold such a thing does not imply that it is true. But their upholding (**) can explain why Cartesians have such *confidence* in self-ascribing mental appearances. Consider that if you affirm "It appears that *p*," then it is self-evident that you have a disposition to affirm that. But if (**) is assumed, it then *follows* that the affirmation is true. Affirming an appearance-ascription is then *unavoidably* correct; (**) deems it is true given that the affirmative-disposition patently exists.

Again, the eliminativist rejects (**), and she balks at the idea that appearance-ascriptions are correct (much less infallibly so). At the same time, the folk upholding of (**) due to (Rule2) proves beneficial, for these are what might allow her to explain the confidence with which we make appearance-ascriptions. But as before, she takes this to amount to an illusion. Although when (Rule2) is operative, what is mistaken for a mental appearance is not a disposition for affirming simply "*p*" (as was the case with (Rule)), but rather a disposition for affirming the folk-psychologically loaded sentence "It appears that *p*.".

6. Closing Remarks

I have argued that the Cartesian case against eliminativism is non-demonstrative. The diagonal argument as well as the argument from introspection beg the question at various points. Further, there is a viable error theory on why we have the Cartesian intuition that mental appearings are real, even granting that the intuition is very powerful.

The error theory, moreover, can be extended in a fairly straightforward way to explain Cartesian intuitions about second-order *judgments*, as distinct from appearings. Consider that Cartesians are apt to claim that, if I judge that I judge that *p*, then I really am judging that *p*. Again, the Cartesian cannot start with the supposition that we make second-order judgments,

at least not without begging the question. But we can explain the Cartesian intuitions with reference to higher-order rules akin to (Rule) and (Rule2), where an introspected state disposing me to affirm "*p*" or "I judge that *p*" yields a disposition for affirming "I judge that *p*." Introspection in such cases again leads us in to implying the existence of a judgment when really there is only a disposition for a kind of linguistic behavior. There still may be communal-wide upholding of the claim that such dispositions are sufficient for a judgment—which like before, could be used to explain Cartesian certainty. But the eliminativist is under no pressure to grant that the community-wide view is true rather than just a pervasive fiction.

This leads to a final thought. The eliminativist, I have suggested, does well to incorporate our common-place acceptance of (*) and (**) into her error theory for Cartesian intuitions. But since she regards these as false, it indeed suggests that we are in the grip of a widespread fiction. However, while calling something a "fiction" often suggests that it is false, it does not strictly entail this. It is possible for a novel to contain factual elements, as in historical fictions such as Tolstoy's *War and Peace*. Similarly, if we regard (*) and (**) as fiction, we can still leave open the possibility that they are true, although of course we would not be committed to that. Even so, one could still take seriously the error theory proposed by our eliminativist without actually being an eliminativist. I suppose we would need to rebrand the error theory as a theory of "possible error," to avoid implying that (*) and (**) really are false. But taking a neutral stance on the question of eliminativism may seem more appropriate for a variety of reasons. And yet, we can still take an important lesson from our eliminativist: For all Descartes has shown, introspection misidentifies mere dispositional states for mental representations.

Bibliography

Armstrong, D.M. (1994). Introspection. In Q. Cassam (Ed.), *Self-Knowledge* (pp. 109-117). Oxford: Oxford University Press.

Armstrong, D.M., Martin, C.B., & Place, U.T. (1996). *Dispositions: A Debate*. London: Routledge.

Baker, L.R. (1987). The threat of cognitive suicide. In her *Saving Belief* (pp. 134–148). Princeton: Princeton University Press.

Baker, L.R. (2013). *Naturalism and the First-Person Perspective*. Oxford: Oxford University Press.

Bengio, Y., Mesnard, T., Fischer, A., Zhang, S. & Wu, Y. (2017). STDP-compatible approximation of backpropagation in an energy-based model. *Neural Computation* 29: 555-577.

Bernecker, S. (1998). Self-knowledge and closure. In P. Ludlow & N. Martin (Eds.) *Externalism and Self-Knowledge* (pp. 333-350). Stanford: CSLI Press.

Boghossian, P. (1990a). The status of content. *Philosophical Review* 99: 157–184.

____. (1990b). The status of content revisited. *Pacific Philosophical Quarterly* 71: 264–278.

Buckner, C. (forthcoming). Deep learning: a philosophical introduction. *Philosophy Compass*. Preprint available at https://philsci-archive.pitt.edu/id/eprint/16326/contents.

Christiansen, M.H. & Charter, N. (1999). Connectionist natural language processing: the state of the art. *Cognitive Science* 23: 417–437.

Churchland, P.M. (1981). Eliminativism and the propositional attitudes. *Journal of Philosophy* 78(2): 67-90.

____. (1989). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press.

____. (2012). *Plato's Camera: How the Physical Brain Captures a Landscape of Universals*. Cambridge, MA: MIT Press.

Churchland, P.S. (1983). *Neurophilosophy: Toward a Unified Science of Mind and Brain*. Cambridge, MA: MIT Press.

____. (1994). Can neuroscience teach us anything about consciousness? *Proceedings and Addresses of the American Philosophical Association* 67(4): 23-40.

Churchland, P.S. and Sejnowski, T.J. (1989). Neural representation and neural computation. In L. Nadel, L. Cooper, P. Culicover, & R.M. Harnish (Eds.). *Neural Connections, Mental Computations*. Cambridge, MA.: MIT Press. Reprinted in (1990) in *Philosophical Perspectives*, 4: 343–382.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36: 181–204.

Crick, F. (1989). The recent excitement about neural networks. *Nature* 337: 129–132.

Demeter, T., Parent, T., & Toon, A. (2022). What is mental fictionalism? In their edited volume *Mental Fictionalism: Philosophical Explorations* (pp. 1-24). New York: Routledge.

Dretske, F. (2003). How do you know you are not a zombie? In B. Gertler (Ed.), *Privileged Access: Philosophical Accounts of Self-Knowledge* (pp. 1-14). Cambridge, MA: MIT Press.

____. (2004). Knowing what you think vs. knowing that you think it. In R. Schantz (Ed.), *The Externalist Challenge*. Berlin: Walter de Gruyter.

Field, H. (1994). Deflationist views of meaning and content. *Mind* 103: 249–285.

Haugeland, J. (2000). *Mind Design II: Philosophy, Psychology, and Artificial Intelligence*, 2nd edition. Cambridge, MA: MIT Press.

Horgan, T. & Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In D. Chalmers (Ed.), *Philosophy of Mind: Classic and Contemporary Readings* (pp. 520–533). Oxford: Oxford University Press.

Huebner, B. (2014). *Macrocognition: A Theory of Distributed Minds and Collective Intentionality*. Oxford: Oxford University Press.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Hutto, D. and Myin, E. (2013), *Radicalizing Enactivism: Basic Minds without Content*. Cambridge, MA.: MIT Press.

Lehrer, K. (1960). Can we know that we have free will by introspection? *Journal of Philosophy* 57: 145-156.

Lillicrap, T., Cownden, D., Tweed, D., & Akerman, J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications* 7: 13276.

Lycan, W.G. (1996). *Consciousness and Experience.* Cambridge, MA.: MIT Press.

____. (2019). *On Evidence in Philosophy*. Oxford: Oxford University Press.

Márton, M. & Tőzsér, J. (2013). Mental fictionalism as an undermotivated theory. *The Monist* 96: 622–638.

Millikan, R.G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.

____. (2005). *Language: A Biological Model.* Oxford: Oxford University Press.

O'Reilly, R.C. (1996). Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Computation* 8: 895–938.

Pitt, D. (2009). Intentional psychologism. *Philosophical Studies* 146: 117-138.

Ramsey, W. (2007). *Representation Reconsidered.* Cambridge: Cambridge University Press.

Rosenberg, A. (2011). *The Atheist's Guide to Reality: Enjoying Life without Illusions.* New York: W.W. Norton & Co.

____. (2022). How to be an eliminativist. *Philosophical Aspects of Origin* 19: 133-163.

Rohde, D.L. & Plaut, D.C. (2003). Connectionist models of language processing. *Cognitive Studies* [Japan] 10: 10–28.

Rumelhart, D.E. & McClelland, J.L. (1996). *Parallel Distributed Processing*, 2 vols. Cambridge, MA: MIT Press.

Sellars, W. (1963). Empiricism and the philosophy of mind. In his *Science, Perception, and Reality* (pp. 127-196). Atascadero, CA: Ridgefield Publishing.

Siewart, C. (2011). Phenomenal thought. In T. Bayne and M. Montague (Eds.), *Cognitive Phenomenology* (pp. 236-267). Oxford: Oxford University Press.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., et al., (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* 362(6419): 1140–1144.

Slezak, P. (1983). Descartes' diagonal deduction. *British Journal for the Philosophy of Science* 34: 13-36.

Soon, C., Brass, M., & Heinze, H.J. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience* 11: 543-545.

Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press.

____. (1991). Narrow content meets fat syntax. In B. Loewer & G. Rey (eds.), *Meaning in Mind: Fodor and His Critics* (pp. 239-254). Malden, MA: Blackwell.

Theiner, G. (2014). A beginner's guide to group minds. In J. Kallestrup & M. Sprevak (Eds.), *New Waves in Philosophy of Mind* (pp. 301-322). Basingstoke, UK: Palgrave.