Philosophical Studies (2007) 133:411–424
DOI 10.1007/s11098-005-6327-x

T. PARENT

INFALLIBILISM ABOUT SELF-KNOWLEDGE

ABSTRACT. Descartes held the view that a subject has infallible beliefs about the contents of her thoughts. Here, I first examine a popular contermporary defense of this claim, given by Burge, and find it lacking. I then offer my own defense appealing to a minimal thesis about the compositionality of thoughts. The argument has the virtue of refraining from claims about whether thoughts are "in the head;" thus, it is congenial to both internalists and externalists. The considerations here also illuminate how a subject may have epistemicially priviledged and a priori beliefs about her own thoughts.

1. INTRODUCTION

In contemporary discussions of mind, Descartes is often attributed the striking claim that a subject is *infallible* with respect to the contents of her own thoughts. This Cartesian Claim about self-knowledge is often met with doubt these days, but in my view, this is because it has not been adequately defended to date. It is my aim in this paper to supply such a defense.

The claim I wish to defend is *not* the claim that whenever a subject thinks that p, the subject knows that she thinks p. Cases of Freudian repression make this implausible.[1] On the other hand, the converse of this claim is trivial: If a subject knows that she thinks p, then she thinks p – simply because, in general, knowledge that Φ entails Φ. However, a modified version of the converse claim has the potential to be both true and non-trivial.

(CC) If a subject believes that she thinks p, she does think p.

Since belief that p does not entail p, the threat of triviality is subverted. Whether (CC) is true is another matter. But if it is true, this would amount to our having infallibly true beliefs

about our own thoughts. Does infallibly true belief count as *knowledge*? I do not wish to decide that issue here; nevertheless, I shall continue to call (CC) a claim about "self-knowledge," even though I wish to use this talk of "knowledge" non-committally.

A related Cartesian claim about self-knowledge is that a subject may possess *indubitable* beliefs about the contents of her thoughts. I do not wish to defend this claim; indeed, I'm inclined to believe the range of what can be doubted is limitless. Even so, if my argument for infallibility is sound, this might entail that self-knowledge is *rationally* indubitable, in the sense that there may be no legitimate reason for doubting it. Actually, I think that inference would be too quick, but regardless I won't investigate this particular issue here.

If self-knowledge is infallible, albeit not indubitable, this would still be quite remarkable in itself; one would like an account of it for its own sake. Moreover, there is additional pressure to account for such knowledge, since arguably it is impossible in light of certain semantic considerations by Putnam (1975) and Burge (1979).[2] However, I do not intend to enter into the imbroglio concerning Putnam–Burge semantics. My argument for Cartesian self-knowledge, I hope, will succeed regardless of the semantic views one may hold.

A few words are in order concerning the term 'thinking' and its cognates. Thinking that *p* does not entail belief that *p*. Moreover, thinking that *p* might be confused with entertaining that *p*; however, thinking that *p* is always compatible with believing that *p*, whereas on some occasions of use, entertaining that *p* is not (e.g. entertaining a hypothesis). More broadly, to think *p* is to have a propositional attitude with the content *p*, but it does not entail having any particular attitude toward that content. Accordingly, with respect to (CC), we are concerned only with knowledge of the *contents* of propositional attitudes, and not of the *attitudes* one might have toward these contents.

In addition, terms like 'thinking' and 'believing' in ordinary discourse are ambiguous between an *occurrent* mental state and a *nonoccurrent* mental state. The former type of state is

present at a particular instant (it is ''occurrent'' at time t), whereas the latter is had over time, as with dispositional and/or tacit beliefs. However, in the present discussion I only want to use 'thinking' and 'believing' to denote *occurrent* mental states.

As we said, to have an attitude (of whatever sort) toward *p* entails thinking that *p*. Accordingly, if we wished to make (CC) more general, we might replace the term 'believe' in the antecedent with the term 'think.' Equivalently:

(CC*) If a subject has a second-order thought that she thinks *p*, she has the first-order thought that *p*.

It is this more general claim that I will examine in the bulk of this paper.

2. BURGE'S WAY OF SELF-KNOWING

One attempt to defend something like (CC*) is given in Burge (1988), who argues that a self-attribution of a thought that *p* is guaranteed to be true. The crucial part of Burge's account here is the ''reflexive'' or ''self-referential'' character of a second-order thought, which necessitates that such a thought have the first-order thought as a proper part. On this, Burge writes:

by its reflexive, self-referential character, the content of the second-order judgment is logically locked (self-referentially) onto the first-order content which it both contains and takes as its subject matter'' (pp. 659–660).

Burge does not say explicitly what it is for thoughts to be ''logically locked,'' but the natural reading would be that if one thought is logically locked to another, then thinking the one thought logically necessitates thinking the other.

Because second-order thoughts are logically locked to first-order thoughts, via self-reference, the upshot for Burge is that second-order thoughts are self-verifying.

[Second-order] thoughts are self-referential and self-verifying. An error based on a gap between one's thoughts and the subject matter is simply not possible in these cases. When I judge: I am thinking that writing

requires concentration, the cognitive content that I am making a judgment about is self-referentially fixed by the judgment itself; and the judgment is self-verifying. (p. 658).

In virtue of the self-referential mechanism in second-order thoughts, to have a second-order thought is always to create conditions which make the content of the thought true.

But what exactly is the model here? A second-order thought, it seems, is typically not self-referential in the sense that such a thought refers to itself. Second-order thoughts may be self-referential on some occasions, e.g. "I am thinking this thought," but they are not characteristically self-referential, e.g. "I am thinking that water is wet."

Perhaps Burge's idea is to understand the relevant second-order thoughts as always having a self-referential device. For example, the thought "I am thinking that water is wet" might be reconstrued as "I am thinking, *with this very thought*, that water is wet" (cf. Burge, 1988, p. 649). However, it unclear whether this would be a satisfying account of these self-attributions. My second-order thoughts, after all, are supposed to be thoughts of *first-order* thoughts I have. But if second-order thoughts are infallible because of a self-referential device, then it seems such an account describes how the thought will infallibly "inherit" the *second*-order thought as part of its content, but not reveal how such a judgment is logically locked to the *first*-order thought. Hence the full explanation of first- and second-order locking can't be given by the fact that the second-order judgment refers to the second-order judgment. This shows that the second-order thought is logically locked to *itself*, but we were wanting an account of how such thoughts are locked to *first*-order thoughts.

A natural response here would be to say that a second-order thought is locked to the first-order thought by the fact that the second-order thought has a mechanism for *referring* to the first-order thought. However, Burge (1996) resists construing the first-order thought only as an object of reference.

Suppose that I think that I am engaging in a thought that there are physical objects. In thinking this, I have to engage the very thought that I am

referring to and ascribing to myself. The reference to the content – expressed in the that-clause – cannot be carried out unless I actually engage in the thought. The intentional content mentioned in the that-clause is not merely an object of reference or cognition; it is part of the cognition itself. (p. 96).

The second-order thought does refer to the first-order thought, but the first-order thought is not *merely* an object of reference. Rather, it is also proper part of the cognition which constitutes the second-order thought. Thus, the idea is that if my first-order thought is a proper part of my second-order thought, then for me to think 'I think that *p*' is, in the very same act, for me to think *p*, since *p* is a proper part of the second-order thought.[3]

Yet it is not clear why we are compelled to accept this last claim, for it seems to be an instance of the fallacy of division, i.e. reasoning from the properties of the whole to the properties of the parts. As a matter of logic, my thinking that I think that *p* does not formally entail that I do think *p*. (After all, in general, a subject thinking Φ does not entail Φ.) Such an entailment may sound plausible in this instance, so perhaps the property of "being thought by me" is a special case here. Perhaps my thinking 'I think *p*' logically necessitates my thinking all the parts of 'I think *p*,' including *p* itself. But if so, this would be a nontrivial fact – a fact that would need to be explained rather than simply assumed. In the remainder of this paper, the primary aim is to account for this fact.

3. CARTESIAN SELF-KNOWLEDGE VINDICATED

The claim in question, recall, is as follows:

(CC*) If a subject has a second-order thought that she thinks *p*, then she has the first-order thought that *p*.

My contention is that, given certain minimal assumptions about the language of thought, (CC*) is a logical consequence. The language of thought (LOT) hypothesis, defended by Fodor (1975), is the hypothesis that thoughts are composed of concepts according to specific formation and

transformation rules, i.e. a "grammar." LOT is typically used as a device for modeling cognitive processes, but what interests me here is the epistemological import of such models for second-order thoughts.

Suppose I token the second-order LOT expression I THINK THAT WATER IS WET.[4] In order to token such an expression, according to the LOT hypothesis, I must token WATER IS WET, since this stands as the complement clause of the expression. Thus if I token I THINK THAT WATER IS WET, I token WATER IS WET. This is just to say that if I second-order think that I think water is wet, I do have the thought that water is wet. Hence any second-order thought that I am thinking some proposition *p* is guaranteed to be true; such a thought is infallible.

Let us introduce the two-place *T*hinking predicate T*xy*, the index 'i' which picks out a subject S who tokens sentences in the language of thought,[5] and variables *p* and *q* which range over sentences. (Let us also acknowledge the LOT counterparts to these devices: T*XY*, I, P, and *Q*.) The derivation of (CC*) proceeds as follows:

(1) S thinks Ti*p*. [Assume for conditional proof]
(2) S thinks that *q* iff S tokens *Q*. [Assumption from the LOT hypothesis][6]
(3) S tokens TI*P*. [From (2), (1)]
(4) If *Q* is of the form FA*P*, where A is a name for an individual *a*, and *P* is a complement clause, then if S tokens *Q*, then S tokens *P*. [Minimal Compositionality Assumption]
(5) TI*P* satisfies the antecedent of (4). [Assumption]
(6) So, if S tokens TI*P*, then S tokens *P*. [From (4), (5)]
(7) So, S tokens *P*. [From (6), (3)]
(8) So, S thinks that *p* [From (2), (7)]
(9) So, if S thinks that Ti*p*, then S thinks *p*. [By conditional proof, (1)–(8)]

As should be clear, (9) is equivalent to (CC*). The most contentious premise in the foregoing is (4), a.k.a. the Minimal Compositionality Assumption (MCA). This premise reflects the compositional nature of the language of thought; however,

it is *minimal* in the sense that one could accept (MCA) without accepting that thoughts are thoroughly compositional. That is to say, one might maintain that second-order thoughts have first-order thoughts as complement clauses, without maintaining that thoughts are complexes of multiple constituents, with a full-blown grammar governing every such constituent.[7]

However, it is worth pointing out that if one did accept an additional compositionality thesis, then we could show not only that a subject is infallible about her own thoughts, but also that she is about the *concepts* she possesses. Thus, consider the following Cartesian Claim about concepts:

(CCC) If a subject thinks she has a concept of *b*, then she does have a *b*-concept.

This claim can be shown by reasoning similar to the line above, by exploiting the compositionality assumption below, in place of (MCA):

(MCA2) If a thought Q is of the form FAB, where A is a name for an individual *a*, and B is a concept of *b*, then if S tokens Q, then S tokens a concept of *b*.

It should be noted that although (MCA2) is (arguably) stronger than (MCA), it is still a relatively minimal assumption about compositionality. According to (MCA2), thoughts are just two-place relations, nothing more. Without going into great detail, the argument for (CCC) starts from a subject's thought "I have a concept that represents *b*." By (MCA2), then, it follows that having such a thought requires a tokening of a concept that represents *b*. But if the subject tokens such a concept, then she possesses such a concept. So whenever a subject thinks she possesses a *b*-concept, the thought is true.[8]

Note that one can accept compositionality principles without accepting other, controversial assumptions about LOT. For instance, it is possible to accept (MCA) without supposing that LOT is innate, or that it is universal among creatures capable of thought. For that matter, we might jettison talk of LOT entirely, and simply note that as a surface phenomenon, second-order thoughts appear compositional. For instance,

such thoughts exhibit *systematicity*. One capable of thinking "Mary thinks that John likes Mary" may also think "John thinks that John likes Mary" by ordering the same concepts in different ways. But not just any thought will result from a particular ordering, nor will every ordering result in a thought, e.g. "John thinks Mary John." Thus it seems particular second-order thoughts result from particular orderings in line with certain *rules*. This is just to say that such thoughts display compositionality.

Even someone like Dennett (1975) would seem to agree to this much. Compositionality would be, in the language of Dennett (1991), a "real pattern." What Dennett questions is whether this compositionality is due to a parallel compositionality of *a physically realized code* in the brain. But if the compositionality of thought is nevertheless real, then in particular the composition of second-order thoughts seems real, by the same kinds of systematicity considerations. So we may acknowledge that some thoughts are composed of certain first-order thoughts and certain concepts, without suggesting there is literally "brain writing." Nevertheless, it is worth emphasizing that the infallibility of such thoughts is a direct consequence of (arguably) our best theory of cognition available.

Note that the argument here does not make use of any premise concerning whether mental content is "in the head." All that the arguments rest upon is a *syntactic* feature of thoughts, to wit, that an occurrence of a thought about a thought (or concept) requires an occurrence of that thought (or concept). Beyond that, I have left open how mental content is individuated; thus, the argument is congenial to both externalists and internalists about content.

Some may protest that this leaves a few important issues about self-knowledge unresolved. If, as Burge (1979) would have it, mental content is individuated in part by the social and physical environment, then a thought like WATER IS WET will be a thought about H_2O. But if so, then there is a sense that the corresponding second-order thought will not provide the subject *knowledge* that she is thinking about H_2O. For in many cases the subject may lack all ability to *distinguish* her

H_2O-thoughts from thoughts about any qualitatively similar substance, hypothetical or real.[9]

If the subject cannot distinguish her H_2O-thoughts from thoughts about qualitatively similar substances, it is not clear she *ipso facto* lacks knowledge of her H_2O-thoughts – especially if there are no qualitatively similar substances to be reckoned with. Regardless, the point would hold that whatever substance a first-order thought is about, the second-order thought will infallibly concern a first-order thought of that substance. Whether these infallible second-order thoughts constitute *knowledge* of first-order thoughts is another matter. But as I mentioned, the expression 'knowledge' is used here noncommittally: Although (CC*) might be called a claim about "self-knowledge," I am interested in defending (CC*) itself, and not the additional thesis that (CC*) *alone* can account for knowledge of one's own thoughts.

4. CLOSING REMARKS

Nevertheless, I do not mean to suggest that (CC*) is of no epistemological import. Since thinking is required for any propositional attitude, (CC*) has the following important epistemological consequence:

(CC) If a subject believes that she thinks *p*, she does think *p*.

Second-order *beliefs* about one's thoughts are infallible on this view. This is a striking fact about our position as epistemic agents. Accordingly, (CC*) can function to explain why such beliefs are epistemically privileged, as compared to the fallible store of empirical beliefs the subject possesses. And what's more, the arguments here do not depend on the subject engaging in any cognitive activity beyond the having of second-order beliefs. Thus, (CC*) may explain how second-order beliefs are epistemically privileged, even though the subject has not undertaken any of the usual actions to garner epistemic status for her beliefs, e.g., empirical observation, inference, etc.

Even though (CC*) has significant epistemological import, it is limited in other respects, since it does not entail other

theses which also seem important in an account of self-knowledge. In particular, (CC*) does not entail:

(CCB) If a subject has a second-order belief that she believes p, then she has the first-order belief that p.

(CCQ) If a subject believes she has a quale of kind q, then she does have a quale of kind q.

I'm inclined to these other claims as well, but I cannot argue for them on this occasion.

Although I call the infallibility of self-knowledge a "Cartesian" thesis, the defense of this thesis is distinctly un-Cartesian in some respects. For one, there is no invocation of a mysterious "Cartesian ego" which makes the infallibility of second-order judgments possible. Given the compositionality of thought, we can defend (CC) on purely naturalistic grounds. Secondly, unlike Descartes, we argued (CC) on *empirical* grounds instead of *a priori* ones. After all, compositionality is an *empirical* thesis about thought, at least in the sense that it is falsifiable by experience. Even so, it may also be possible to introspect the compositionality of one's thoughts and thereby come to know (CC) on somewhat *a priori* grounds. But even if this is correct, I take the empirical defense to lend further credence to (CC) beyond what mere *a priori* conjecture could bestow. Finally, unlike Descartes, I will not attempt to establish second-order judgments as part of an epistemic "foundation" by which we may come to justify other beliefs. Indeed, if Bonjour (1978) is correct, the argument for (CC) possibly *precludes* second-order judgments from being foundational, since foundational beliefs would seem to be beliefs with no argument backing them.

However, like Descartes, some of the motivations here are epistemological. I take the infallibility of second-order judgments to be of interest in its own right; however, I suspect many of us defend some form of privileged self-knowledge because of a certain epistemological anxiety. Roughly, the worry is that the process of *first-person epistemic reflection* won't make rational sense if we lack some kind of secure access to our own thoughts. In reflection, if I cannot reliably

judge what my first-order beliefs *are*, then so much the worse for my aspirations to *evaluate* them. This is especially so if such evaluation involves considering a (putative) first-order belief in relation to other (putative) first-order beliefs.

Even so, must our access to our own beliefs be *infallible*? I cannot address the issue adequately here. Briefly, however, infallible access may be necessary; otherwise, the evaluation of first-order beliefs might become more and more skewed as reflection continues onward. If access to my beliefs is fallible, then as reflection progress, my mistakes about what I believe will presumably increase in number. But suppose, again, that an evaluation of a (putative) belief consists at least in weighing it against my other (putative) first-order beliefs. Then, as I misrepresent my own first-order beliefs more and more, I will misunderstand more and more how one belief stands up to my actual belief-set. And so as reflection continues onward, I will increasingly misunderstand the value of a belief, relative to my other beliefs. Reflection then would take me farther, rather than nearer, to a constructive first-personal assessment of my own beliefs.

Nevertheless, (CC) alone does not insure us against wayward reflection; something stronger would be needed, along the lines of (CCB). Even so, (CC) guarantees second-order judgments of a certain kind, namely, about the *content* of thoughts. So even though we have not discussed how we can discern our *believing* that p (as opposed to wishing, doubting, disbelieving p), we have shown how we infallibly track that it is $\boldsymbol{p}$ we are thinking of, as opposed to some other proposition p^*. (E.g. thoughts about water, vs. thoughts about twin-water). In this respect, we have a partial solution to the problem I have gestured at, the problem of wayward reflection.

ACKNOWLEDGEMENTS

This paper represents a revision of my M.A. thesis "Introspection and its Epistemology" written at the University of North Carolina, Chapel Hill, in the spring of 2002. I thank my committee – Dorit Bar-On, Doug Long, and especially

William Lycan – for valuable comments on earlier drafts of the paper. I am grateful, in addition, for helpful feedback I received from Larry Coulter, Sarah Sawyer, two anonymous referees, as well as from audiences at the 2003 Chapel Hill Philosophy Retreat, the 2004 MidSouth Philosophy Conference, and the 2004 Eastern Division Meeting of the American Philosophical Association.

NOTES

[1] See also Nisbett and Wilson (1977) and Nisbett and Ross (1980). This conditional, as well as the related conditional "if the subject thinks that *p* then she *believes* that she thinks that *p*," would seem to hold in many cases. But to the extent that these conditionals do hold, this would seem to be explained by some introspective faculty of the mind, of the sort discussed by Armstrong (1968) and Lycan (1996). However, unlike Armstrong and Lycan, I think it is an open question as to whether the introspective faculty is more or less reliable than ordinary sense-perception. The intuition that there is something especially reliable about self-knowledge might just be explained by the infallibility claim I defend here, not by introspection *per se*. [Cf. Bar-On (2004) who suggests that reliability alone would not be sufficient for an adequate introspectionist account of self-knowledge; see esp. p. 183]

[2] For a defense of this thesis, see Boghossian (1989).

[3] This also seems to be the view in Heil (1988), where Heil talks of second-order thoughts *including* the content of first-order thoughts (see, e.g., p. 224). Presumably, this is just another way of saying that first-order thoughts are *part* of second-order thoughts; accordingly, what I say about Burge's view may be applied to Heil's view as well.

[4] Following established conventions, English expressions in SMALL CAPS are the names for lexemes in LOT with the same content as the English expression.

[5] Here, the index 'i' is what Perry (1979) would call an *essential* indexical, since in what follows 'i' is not substitutable *salva veritate* for a name of the person for which it indexes on a given occasion.

[6] Given the definition of 'think,' it follows that *all* propositional attitudes require a tokening of an LOT sentence. A reviewer has expressed doubts about this claim; however, it is important to note that my argument here can be restated without reference to a language of thought at all (see pp. 6–7). Nevertheless, I would also refer would-be skeptics to Fodor's defense of this particular claim in his (1978).

[7] Thanks to William Lycan for suggesting this minimal sort of compositionality for LOT.

[8] The following has been suggested to me as a counterexample to (CCC): Suppose an undergraduate hears the word 'supervenience,' and starts using the word himself, thinking that he has the concept of supervenience, when in fact he is merely miming the use of the word. Then, contra (CCC), he thinks that he has the concept of supervenience, but doesn't actually have it. One might argue that the student *does* have a concept of supervenience, albeit an impoverished one. But supposing this is not so, we might ask what concept would be used in his second-order judgment if not the concept of supervenience? It may be a concept 'sloopervenience,' and/or a concept of ''whatever competent speakers designate by the word 'supervenience.'" Regardless, if we call this (possibly complex) concept *c*, then his second-order judgment will be the judgment that he has the concept *c*. Yet in such a judgment, *c* is used; thus his judgment about having *c* is infallibly true. His mistake would be in thinking that his concept *c* is *comperable to* the concept competent speakers express by the word 'supervenience.' (It is important to note that, assuming that concepts are ''hyperintentional,'' the concept of supervenience and the concept *c* may be coreferring, yet nevertheless be nonidentical, as when *c* is a concept of ''whatever competent speakers designate by the word 'supervenience.'")

[9] This is, I take it, the distinctly epistemological problem with self-knowledge that arises from Burge's (1988) slow-switching case.

REFERENCES

Armstrong, D.M. (1968): *A Materialist Theory of the Mind*, London: Routledge and Kegan Paul.

Bar-On, D. (2004): *Speaking My Mind: Expression and Self-Knowledge*, Oxford: Oxford UP.

Boghossian, P. (1989): Content and Self-Knowledge, *Philosophical Topics* 17, 5–26.

Bonjour, L. (1978): Can Empirical Knowledge Have a Foundation?, *American Philosophical Quarterly* 15, 1–13.

Burge, T. (1979): Individualism and the Mental, *Midwest Studies in Philosophy* 4, 73–121.

Burge, T. (1988): Individualism and Self-Knowledge, *The Journal of Philosophy* 85(1), 649–663.

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

Dennett, D. (1975): True Believers: The Intentional Strategy and Why It Works, in A.F. Heath (ed.), *Scientific Explanations: Papers based on*

Herbert Spencer Lectures Given in the University of Oxford, Oxford: Oxford UP (pp. 53–75).
Dennett, D. (1991): Real Patterns, *Journal of Philosophy* 88.1, 21–51.
Fodor, J. (1975): *The Language of Thought*, Cambridge, MA: Harvard University Press.
Fodor, J. (1978): Propositional Attitudes, *The Monist* 61(4), 501–523.
Heil, J. (1988): Privileged Access, *Mind* 42(386), 238–251.
Lycan, W.G. (1996): *Consciousness and Experience*, Cambridge: MIT Press.
Nisbett, R.E. and Ross, L. (1980): *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, NJ: Prentice-Hall.
Nisbett, R.E. and Wilson, T.D. (1977): Telling More than We Can Know: Verbal Reports on Mental Processes, *Psychological Review* 8, 231–259.
Perry, J. (1979): The Problem of the Essential Indexical, *Nous* 13, 3–21.
Putnam, H. (1975): The Meaning of 'Meaning', *Mind, Language, and Reality*, Cambridge: Cambridge University Press (pp. 215–271).

Department of Philosophy
University of North Carolina
Caldwell Hall, CB# 3125
Chapel Hill, NC 27510
USA
E-mail: tparent@email.unc.edu