This is the Accepted Manuscript of an article published by Taylor & Francis in *Philosophical Explorations* 18/2 (2015), 183-198, on June 11, 2015, available online at http://www.tandfonline.com/doi/full/10.1080/13869795.2015.1032328. Please cite only the published version.

Self-knowledge about attitudes:

rationalism meets interpretation

Franz Knappik*

Institute of Philosophy, Humboldt-Universität zu Berlin, Berlin, Germany

Recently influential 'rationalist' views of self-knowledge about our rational attitudes hold that such self-knowledge is essentially connected to rational agency, and therefore has to be particularly reliable, immediate, and distinct from third-personal access. This approach has been challenged by 'theory theory' or (as I prefer to call them) 'interpretationist' views of self-knowledge: on such views, self-knowledge is based on the interpretation of information about ourselves, and this interpretation involves the same mindreading mechanisms that we use to access other persons' mental states. Interpretationist views are usually dismissed as implausible and unwarranted by advocates of rationalism. In this article, I argue that rationalists should revise their attitude towards interpretationism: they can, and ought to, accept themselves a form of interpretationism. First, I argue that interpretationism is correct at least for a substantive range of cases. These are cases in which we respond to a question about our attitudes by a conscious overt or inner expression of our attitude, and form a self-ascriptive belief on the basis of that expression. Second, I argue that rationalists can adopt interpretationism without abandoning their basic tenets: the assumption that both approaches are incompatible is unfounded.

Keywords: self-knowledge; rational agency; theory theory of self-knowledge; cognitive phenomenology.

1. Introduction

Much of the recent literature on the knowledge that we have of our own present rational attitudes (such as beliefs and intentions) proceeds on the assumption that such self-knowledge must be seen as part of the relation in which we stand to our mental lives qua reason-oriented, self-critical thinkers, or 'rational agents'. As we are the authors of our attitudes, our access to them, it is argued, is distinct from the third-personal access to other persons' rational attitudes. For instance, several authors have advocated so-called 'transparency' views of self-knowledge (e.g., Moran [2001], Boyle [2009], and Byrne [2011]). According to them, the

* Email: franz.knappik@hu-berlin.de.

paradigmatic form of agential self-knowledge issues from situations in which we make up our minds through conscious deliberation. In such cases, it is argued, we answer the question what we believe or intend to do not by looking inside us, or by collecting and evaluating evidence about ourselves – as we would do in the third-personal case –, but by reasoning about first-order questions, for instance (in the case of belief) questions about what the world is like. There are other proposals, too, regarding how we should account for our distinctively agential first-personal access to our rational attitudes (cf., for instance, Shoemaker [1996], Burge [1996], and Bilgrami [2006]).[1] What is common to the various accounts is the idea that our role as rational agents makes an important difference regarding the nature of our self-knowledge.[2]

In addition to a difference regarding the nature of first-personal access, it is often claimed that our role as rational agents gives rise to two particular epistemic features of self-knowledge. First, it is held that our first-personal access to our own rational attitudes needs to be particularly reliable: we could not be self-critical thinkers, it seems, if we were usually ignorant about our attitudes (cf., e.g., Shoemaker [1996]; Burge [1996]; Bilgrami [2006]). Second, such self-knowledge is normally thought to be *immediate*. We would not have a specifically agential, first-personal access to our attitudes if we would have to find out about them by observing ourselves, or by drawing inferences. Rather, it seems that we can normally tell straightway what we believe, intend etc. (e.g., Moran [2001, 124–34]).

Importantly, authors who subscribe to these views about self-knowledge – following Gertler (2011a), we can call them '*rationalists*' – do not claim that our agential access extends to *all* of our rational attitudes. Of course, they do not deny that there are such things as repressed beliefs and wishes, hidden stereotypical opinions and preferences, etc., and that we often can gain self-knowledge regarding *such* attitudes only by gathering and evaluating information about ourselves. But rationalists argue that these are cases in which our rational control and authorship have suffered a break-down – cases in which we hold attitudes that we

do not manage to adjust to our reasons. Correspondingly, our epistemic access to such phenomena is of a third-personal, detached form; and since that form falls short of the characteristics of the normal first-personal access of rational agents, it should count as pathological or *alienated* form of self-knowledge. (This is particularly emphasized by Moran [2001]).

Despite its intuitive appeal, the rationalist approach to self-knowledge has not gone unchallenged. In particular, it has been opposed in recent years by a view on which self-knowledge involves *interpretation* of information about oneself – including both publicly observable behaviour and circumstances, and inner sensory and quasi-sensory evidence (such as episodes of inner speech) (cf., e.g., Gopnik [1993], Churchland [1999, 73–80], and Carruthers [2009] and [2011]). According to this '*interpretationist*'[3] approach, which is historically rooted in the work of Gilbert Ryle and Wilfrid Sellars, such interpretation usually takes place unconsciously, and employs the same, or at least relevantly similar, interpretive mechanisms as those that allow us to detect the mental states of others. (In particular, it has been argued that we gain self-knowledge by applying to ourselves the 'theory of mind' that allows us to interpret other persons' behaviour in terms of rational attitudes and other mental states: e.g., Gopnik [1993]; Carruthers [2011].) The resulting self-knowledge is therefore based on a *mediating* process, rather than being immediate; and rather than having a distinctive nature, it is basically on a par with knowledge of other persons' propositional mental states. Finally, there is no reason for interpretationists to hold that our interpretive access to our propositional mental states yields particularly *reliable* results. Indeed, some of them take empirical data to show that our accounts of them are often, unbeknownst to us, confabulated (see in particular Carruthers [2011, ch. 11]).

Authors on both sides normally agree that rationalism and interpretationism are two mutually exclusive views of self-knowledge – that self-knowledge cannot be agential and interpretive at the same time (cf., e.g., Carruthers [2011, 12–21, 79–108]; Moran [2001, 5–8];

Bilgrami [2006, 12–22]). But rationalists generally do not take interpretationism very seriously as a rival view. Not only are they bound to consider interpretationism a strongly counterintuitive position, as it seems to contradict our commonsensical self-understanding as rational agents. Rationalists also tend to be unimpressed by the arguments that have been advanced in favour of interpretationism. These are mostly empirical arguments;[4] rationalists partly doubt that the existing data enforce an interpretationist conclusion (e.g., Bilgrami [2006, 17]), and partly, they assume that the philosophical and the empirical investigation of self-knowledge occupy distinct explanatory spaces (e.g., Moran [2001, 5–8]).

In this article, it is my aim to show that rationalists ought to revise their attitude towards interpretationism. In the bulk of the paper (sections 2 and 3), I will offer an argument for interpretationism regarding a substantive range of normal, non-alienated cases of self-knowledge. This argument defends interpretationism on conceptual grounds, and hence cannot be as easily dismissed by rationalists as the existing empirical arguments. However, my focus will not be on the self-knowledge that issues from situations of conscious deliberation, and that many rationalists (as I have mentioned before) treat as paradigmatic – rather, I will leave open here what precise analysis should be given of *these* cases of self-knowledge. Instead, my strategy will be to focus on self-knowledge that issues from situations of a different type – situations in which we gain self-knowledge about rational attitudes that we already hold, and which I call 'expressive episodes'. (I will set out in detail what it takes for a situation to count as an expressive episode in my sense at the beginning of section 2.) In section 2, I will argue that such expressive episodes occur frequently, and that it would be implausible to regard the self-knowledge that arises from them as being per se deficient or 'alienated'. In section 3, I will argue that such expressive episodes nevertheless should be seen as involving interpretation – and more precisely, interpretation that involves mechanisms which are at least relevantly similar to those mechanisms by which we ascribe rational attitudes to other subjects.

At the same time, I submit that one can accept this argument and still retain the basic tenets of rationalism: for I contend that the common assumption according to which both views exclude each other is mistaken, at least as far as self-knowledge that issues from expressive episodes is concerned. I will argue for this additional claim in section 4. Taken together, this will yield the conclusion that it is both *possible* and *mandatory* for rationalists to combine their position with some form of interpretationism.

Some preliminary remarks are in place. First, rather than stipulating a definition of the notoriously problematic notion 'interpretation', I will merely assume for the sake of my discussion that it is *sufficient* for interpretation that ascriptions of propositional mental states are issued on the basis of indicators which do not themselves have propositional content. Importantly, the interpretive process need not be conscious, and it need not be a strictly inferential process: its internal steps need not consist of full-blown *premises* that we can hold true and deliberate upon at the personal level.[5] I assume that this notion captures the spirit of interpretationism, and renders the claim that self-knowledge requires interpretation neither trivially true nor obviously wrong.[6]

Second, I will not attempt here to defend rationalism in its own right. The possibility to reject rationalism for reasons which do *not* presuppose an incompatibility between agential and interpretive self-knowledge will remain unaffected by my argument.

Finally, the positions that I am concerned with here share the assumption that normal, non-alienated self-knowledge has an epistemology, and is a cognitive achievement (even if we can obtain this achievement without conscious effort). I will take this assumption for granted, and my following argument will be conditional upon it.

2. The frequency and non-alienated nature of expressive episodes

As I have already mentioned, I will be concerned in this and the following section with situations in which we gain self-knowledge about already existing rational attitudes. More

precisely, I will aim to show that in an important range of cases, we gain self-knowledge about existing attitudes that is both interpretive and non-alienated in the course of episodes which combine the following structural features:

> (α) I have a rational attitude (e.g., a belief that *p*, an intention to ϕ), without possessing already the second-order belief *that* I have this attitude. (β) At some point, a question regarding this attitude arises (for instance, I am asked whether I believe that *p*, or whether I am going to ϕ; or I ask myself such a question, or encounter it in some other way). (γ) Without deliberating about the subject-matter of my attitude, (δ) I respond to this question by a conscious event that expresses the attitude – e.g., a thought or an utterance with the content '*p*' or 'I shall ϕ'. (ε) On the basis of this event, I form a higher-order belief in which I self-ascribe the first-order attitude (e.g., a belief with the content 'I believe that *p*', or 'I intend to ϕ').

I shall call episodes with this structure '*expressive episodes*' because their central element is a thought that expresses the first-order attitude. I will refer to higher-order beliefs in which we self-ascribe rational attitudes as '*self-ascriptive beliefs*'; to the rational attitudes that are, or can be, self-ascribed in such beliefs, as '*target attitudes*'; and to the conscious expression by which we respond to the initial question within an expressive episode as '*response event*'. – Regarding such expressive episodes, I will argue for the following two claims:

(1) In a substantive range of cases, normal, non-alienated self-knowledge issues from expressive episodes.

(2) The process that leads from the response event within an expressive episode to the self-ascriptive belief includes interpretation. At least in many cases, it employs interpretive capacities that are identical, or at least relevantly similar,[7] to those that we use for the attribution of mental states to other persons.

In this section, I will argue for claim (1).

To begin with, we should notice that it is by no means a trivial question whether there are situations at all that display each of the above features (α)-(ε). I do take it to be uncontroversial that we often experience situations which satisfy the conditions (β) and (δ) for expressive episodes – i.e., situations in which we respond to a question about a rational attitude by a conscious event that expresses the attitude. By contrast, it is a matter of contention whether any of the further conditions (α), (γ), and (ε) is fulfilled, too, in a substantive number of non-pathological cases. For there are several ways of understanding the relation between a target attitude and a self-ascriptive belief which, if correct, would exclude that those conditions regularly obtain:

- *First*, rational attitudes could always, or at least normally, be already accompanied by corresponding self-ascriptive beliefs. In that case, condition (α) for an expressive episode would be never, or at least not normally, be fulfilled.

- *Second*, one could hold that rational attitudes *can* occur without a self-ascriptive belief, but that the formation of warranted self-ascriptive beliefs is due to a subpersonal causal mechanism, rather than being based on a conscious step. In that case, condition (ε) for expressive episodes would normally not be fulfilled; conscious events in response to questions about our rational attitudes would be epistemically idle epiphenomena.

- *Third*, even if it is accepted that we regularly form self-ascriptive beliefs about already existing target attitudes by means of a conscious expression of the target attitude, claim (1) can still be resisted if it is assumed that non-alienated self-knowledge *always* issues from situations of deliberation about the subject-matter of the attitude. In that case, self-knowledge that results from situations in which condition (γ) for expressive episodes – the absence of deliberation – is satisfied would ipso facto be alienated.

If it is taken for granted that we have some non-alienated self-knowledge at all, and that conditions (β) and (δ) are unproblematic, these options exhaust the possible alternatives

to claim (1). In the following, I will aim to establish claim (1) by discussing those options in turn, and argue that they are not satisfactory.

2.1 Universal self-knowledge?

Consider, to begin with, the first of the above options: the view on which rational attitudes are always or normally already accompanied by a self-ascriptive belief. This 'universalism' about self-knowledge, as we may call it, can be motivated, for instance, by explaining self-knowledge on the basis of a constitutive relation between target attitudes and self-ascriptive beliefs. Thus, on Shoemaker's version of a rationalist account of self-knowledge, it is part of the essence of rational attitudes to be accompanied by corresponding self-ascriptive beliefs (cf. Shoemaker [1996, e.g. 242]).

As it stands, universalism is confronted with a regress problem. Since self-ascriptive beliefs are rational attitudes, too, the view gives immediately rise to an infinite hierarchy of beliefs, which would be psychologically impossible.

In response to this problem, a universalist has to add a qualification that blocks the regress. The most promising possibility to do so is to introduce a distinction between explicit and implicit (or 'tacit') attitudes (cf. Shoemaker [1996, 240–1]). The proposal would be that only our *explicit* attitudes are always or normally already accompanied by an at least *implicit* self-ascriptive belief. – But it can be argued that by modifying the original proposal in this way, the universalist de facto makes room for a substantial range of situations that satisfy condition (α) – at least if a plausible reading of the explicit-implicit distinction is adopted. I shall discuss two relevant readings of that distinction.

To begin with, it has been suggested that an implicit attitude is a disposition to form, without examining further evidence, the corresponding explicit attitude, once one considers its subject-matter (cf., e.g., Gertler [2011b, 130–1] (on belief), as well as the literature cited by Crimmins [1992, 242]). If this is applied to universalism, the resulting view would be that

whenever a subject holds an explicit attitude A, it also has a disposition to form, without examining further evidence, an explicit self-ascriptive belief with A as its target attitude, once it considers the question whether it holds A. But nothing would exclude that this disposition is typically manifested by a formation of a self-ascriptive belief *in the course of an expressive episode*. Hence, this view is entirely compatible with the idea that we frequently gain self-knowledge about our attitudes – *including* our explicit attitudes – in the course of expressive episodes.

According to a further prominent reading of the explicit-implicit distinction, implicit attitudes share the dispositional profiles of the corresponding explicit attitudes, but differ from them in that they do not involve the possession of whatever concrete mental entity it takes to hold the explicit attitude (Crimmins [1992]).[8] It is true that the modification of universalism that results if *this* reading of the explicit-implicit distinction is adopted rules out the existence of situations with feature (α) for the case of *explicit* attitudes. But at the same time, the resulting view is entirely compatible with the possibility that *implicit* attitudes (in the proposed sense) often, if not always, lack a corresponding self-ascriptive belief. And implicit attitudes in *this* particular sense are far from being peripheral for rational agency, or being ipso facto alienated. For on the one hand, it is generally agreed upon that we all hold enormously many implicit attitudes. (Implicit attitudes are often thought to include, for instance, many of the logical consequences of our explicit beliefs, as well as the complex beliefs and intentions that are ascribed to speakers in speech-act theory: cf. Crimmins [1992, 240–1].) On the other hand, if the implicit attitudes really share the dispositional profiles of their explicit counterparts, someone with a given implicit attitude will reason and act just as someone who has the corresponding explicit attitude. Hence, implicit attitudes have the very same significance for rational thought and action as their explicit counterparts. So even if the modified version of universalism is presupposed, and the existence of expressive episodes is restricted to implicit attitudes, self-knowledge that issues from expressive episodes could still

constitute an important form of agential self-knowledge. (For the sake of the following discussion, I will assume that universalism is wrong.[9] For readers who disagree, the view that I have just sketched is my fall-back position.)

2.2 Direct causation?

The assumption that self-ascriptive beliefs about already existing target attitudes are reliably formed by a subpersonal causal mechanism – a view that may be called the '*direct-causation model*' –, promises to offer an attractively simple account of self-knowledge. Its most prominent variant is the 'Monitoring Mechanism' postulated by Nichols and Stich (2003, 160–3) (see also Cassam [2010, 90–3]): this mechanism directly accesses the representations contained in our belief box, desire box etc., adds a corresponding operator 'I believe', 'I desire' etc. to them, and stores the resulting self-ascriptive representations in the belief box. Such a mechanism is, according to Nichols and Stich (2003, 171), 'trivial to implement', and hence offers the best available explanation for our capacity to self-ascribe rational attitudes.

However, as Goldman (2006, 238–9) has argued, Nichols and Stich's Monitoring Mechanism founders on a crucial difficulty: while it offers an explanation of how the *contents* of target attitudes are detected, it fails to account for our knowledge of the attitude *types* – our knowledge of whether a given attitude is a belief, an intention, a hope, etc. On Nichols and Stich's account, what type the Monitoring Mechanism ascribes to a given attitude depends on the 'box' in which that attitude is stored. But since such 'boxes' stand for functional roles, this raises the question of how the Monitoring Mechanism could identify the functional role of the attitude.[10] The need for such an identification could be circumvented if there was a distinct mechanism for each type. But as Goldman remarks (2006, 239), this would lead to an implausible inflation of the number of requisite mechanisms. And if there is only one Monitoring Mechanism, it remains unclear how that mechanism could possibly identify the different functional roles that define the various attitude types.

Since it is generally assumed that attitude types are defined by functional roles, this difficulty applies to direct-causation models of self-knowledge in general. In the absence of an account of how a direct-causation model could deal with attitude types, we should conclude that condition (ε) for expressive episodes is regularly fulfilled, and self-ascriptive beliefs are regularly based on conscious response events.

2.3 Alienation?

This still leaves open the possibility that self-knowledge which issues from expressive episodes is ipso facto alienated, because non-alienated self-knowledge presupposes deliberation about the subject-matter of the self-ascribed attitude (contra condition (γ) for expressive episodes). Thus, one might hold, with Shah and Velleman, that in order to know one's already existing or 'antecedent' beliefs, one needs to 'refrain from any reasoning as to whether *p*, since that reasoning might alter the state of mind that one is trying to assay' (Shah and Velleman 2005, 506). But this would certainly amount to a third-personal, alienated stance towards one's beliefs. It therefore may seem that non-alienated self-knowledge can be gained only if we treat the question whether we believe that *p* as an 'invitation to reasoning' (Shah and Velleman 2005, 507), and consider whether *p* is true. In cases where we already have a belief about the truth of *p*, we would then have to suspend that belief, and re-open the question. (A parallel point applies to intentions.)

Yet this would hardly be a satisfactory view. It is true that having a rational attitude must leave open the possibility that a relevant change in one's further attitudes leads one to re-open the issue – e.g., if new evidence becomes available which speaks against a belief; or if an unexpected situation forces one to revise an intention. But this possibility only requires a *readiness* to suspend and revise one's attitudes, *once* one becomes aware of such overriding factors. If, by contrast, someone would continuously re-open the question even in the absence of overriding factors, this would show that he actually does not have the belief or intention in

question – rather, he would be uncertain or undecided about the issue. – Hence, the absence of deliberation in expressive episodes (condition (γ)) is not in conflict with rational agency, as long as it comes (pace Shah and Velleman) with a readiness to reconsider one's rational attitudes *if necessary*.

Thus, the three alternatives to claim (1) have turned out to be either implausible, or, after all, compatible with a version of (1). I therefore conclude that a broad range of instances of non-alienated self-knowledge indeed issues from expressive episodes in which all of conditions (α) to (ε) are satisfied.

3. The interpretive character of expressive episodes

In this section, I will turn to the above claim (2), and argue for an interpretationist account of expressive episodes. I will do so by discussing how precisely the relation between the response event and the self-ascriptive belief in expressive episodes is to be understood. Throughout this argument, I will focus on the case in which conscious thoughts, rather than public utterances, figure as response events. At the end of the section, I will indicate how the argument can be seen to apply to the case of public utterances, too.

3.1 The epistemic role of P-consciousness

First, we have to understand better in which sense(s) the response event is *conscious*. There are two common notions of consciousness that are relevant here: namely, in Ned Block's influential terminology, phenomenal consciousness (or 'P-consciousness') and access consciousness (or 'A-consciousness'). According to Block, that a state is P-conscious means that it is part of the subject's experience, or that it contributes to what it is like for the subject (Block 1997, 380). By contrast, a state is A-conscious if it is 'poised for direct control of thought and action' (Block 1997, 382), or 'inferentially promiscuous' (Block 1997, 384).

In normal cases of expressive episodes, it seems that the response event is both P-conscious and A-conscious. It makes a difference to how we experience the episode, and it is available to be uttered in speech (if it is not itself already an utterance), to be used as premise in subsequent reasoning, and – most importantly for our purposes – to be processed by the mechanisms by which we form higher-order beliefs (which we can subsume under the term 'meta-cognitive capacity').

But what precise roles do these forms of consciousness play in the process that leads to warranted self-ascriptive beliefs within expressive episodes? Regarding A-consciousness, what matters for this process is specifically the availability of the response event to the meta-cognitive capacity. But this availability might in principle obtain independently of the availability of the response event for *other* information-processing capacities. So A-consciousness as such (which would require a broader availability) does not seem to be the form of consciousness in virtue of which the response event can figure as basis of the self-ascriptive belief.

Regarding P-consciousness, by contrast, there are in principle two possibilities. The process that leads from the response event to a warranted self-ascriptive belief – or, in other words, the access that our meta-cognitive capacity has to the response event – can either be explanatorily independent of the fact that the response event is P-conscious, or it can explanatorily depend on that fact. The first possibility would amount to a form of *epiphenomenalism*, according to which the P-consciousness of the response event is *explanatorily idle* regarding our self-knowledge. But if we adopted this view, we would be back with an account of self-knowledge in terms of subpersonal mechanisms akin to the direct-causation model that we have discussed earlier. The only difference would be that on the epiphenomenalist view, the mechanism in question assesses our attitudes not directly, but via the response events. In order to detect our attitudes, this mechanism would have to assess not only the content of the response events, but also the types of the attitudes that are

expressed by such events – it would have to be able to detect whether a given response event expresses a belief, or an intention, etc. But this ability remains equally unexplained as in the case of the direct-causation model.[11]

Therefore, we should embrace the alternative to the epiphenomenalist view: an account on which the response event is available to the mechanism which forms the self-ascriptive belief *in virtue of being P-conscious*.[12] – In the next section, I will discuss how precisely the response event can form the basis of a warranted self-ascriptive belief in virtue of being P-conscious.

3.2 Cognitive phenomenology and interpretation

The role that the phenomenal properties of the response event play for the formation of warranted self-ascriptive beliefs can be understood either in internalist or in reliabilist terms. An internalist will hold that (some of) the phenomenal properties that characterize the response event serve as introspectively available *evidence* for the self-ascriptive belief. On that view, expressive episodes have to include an awareness of the response event *as* manifestation or expression of the target attitude.[13] By contrast, a reliabilist will merely assume a causal mechanism such that the occurrence of the phenomenal properties which characterize the response event is regularly followed by the formation of a corresponding self-ascriptive belief.

On either approach, whether expressive episodes require interpretation or not will depend on what precise account is given of the phenomenal properties that characterize the response event. These phenomenal properties are an instance of what is called 'cognitive phenomenology' – that is, the phenomenology that characterizes cognitive states and events, such as conscious occurrent thoughts. Traditionally, cognitive phenomenology has usually been thought to consist in, or to be reducible to, (quasi-)*sensory*[14] phenomenology, such as inner speech, visual imagery, and affective feelings (a view that is sometimes called '*impure*

cognitive phenomenology'). (For defences of impure cognitive phenomenology, see, for instance, the contributions by Carruthers and Veillet, Levine, Prinz, and Tye and Wright in Bayne/Montague [2011].) If this view of cognitive phenomenology is adopted, it follows quite straightforwardly that expressive episodes have to involve interpretation. Take the simplest case, in which the sensory phenomenology of a response event consists in an episode of inner speech – for instance, an inner rehearsing of the English sentence 'It will rain tomorrow'. We can imagine the same piece of inner speech as being rehearsed *without understanding* – e.g., if a person who does not understand English arbitrarily rehearses simulated spoken sounds and by coincidence mentally utters 'It will rain tomorrow' (cf. Pitt [2004, 24]). The episode would then have the same (or roughly the same) sensory phenomenal properties as the original episode.[15] Nevertheless, it hardly could give rise on its own to a warranted self-ascriptive belief. Having the sensory phenomenology in question without understanding would neither amount by itself to having an awareness of the P-conscious episode *as* expression of the target attitude (as the internalist approach would require). Nor would the episode normally have the right causal role to provide reliabilist warrant: it would be an unlikely coincidence if the subject's rehearsing 'It will rain tomorrow' without understanding would cause a correct belief in which the subject self-ascribes her belief that it will rain tomorrow. We therefore have to conclude that on the assumption of impure cognitive phenomenology, the phenomenal properties of the response event can serve as epistemic basis for the self-ascriptive belief only if, in addition, a process of interpretation takes place.[16]

It could seem that this conclusion can be avoided if it is assumed instead that cognitive states like thoughts have phenomenal properties which cannot be reduced to sensory phenomenology, but have a distinctive, non-sensory phenomenology of their own (a view called '*pure cognitive phenomenology*'). For at least on some versions of this view, the phenomenal properties of thoughts are not only individuative of the thoughts' intentional

properties (i.e., their properties of having a determinate intentional *content*, and of taking up a determinate *attitude* towards that content): they are even *constitutive* of these intentional properties (see Horgan and Tienson [2002]; Pitt [2011]; Kriegel [2011]). On such a view, we think a thought with a particular intentional property (e.g., a thought that assents to *p*) *by* experiencing an occurrence of the corresponding phenomenal property. Accordingly, it might be thought that on this understanding of cognitive phenomenology, the response event would have phenomenal properties that are intrinsically meaningful, and could therefore provide direct access to one's rational attitudes.[17]

Consider, first, how this proposal fares if the epistemic role of the response event is given an internalist reading. On this approach, we would have to consider the phenomenal properties of the response event as evidence for the target attitude, and expressive episodes would involve an accompanying thought or awareness that the response event expresses the target attitude. Now while on the strong version of pure cognitive phenomenology, an occurrence of a cognitive phenomenal property P *is* ipso facto an occurrence of a thought with the correlated intentional property Q, it does not follow that an *awareness* of the occurrence of P is ipso facto an awareness of the occurrence of a thought with Q. For despite being connected by a constitutive relation, phenomenal and intentional properties are at least intensionally distinct: they correspond to two different modes of presentation. It follows that in addition to the occurrence of the phenomenal property, the accompanying awareness assumed by the internalist approach requires an additional step – a step in which we (at least unconsciously) detect the intentional properties of the current thought on the basis of its phenomenal properties. This detection presupposes an ability to map phenomenal onto intentional properties. And since the phenomenal properties are not *themselves* intentional properties, this ability is an interpretive one.

It should be noted that the assumption of such a detection mechanism would be quite implausible. In the case of sensory phenomenal properties, we possess phenomenal concepts

by which we are able to discriminate, and to communicate, phenomenal properties (e.g., perceived colours) independently of an antecedent interpretation in terms of represented objects etc. By contrast, we do not seem to possess analogous capacities regarding pure, non-sensory cognitive phenomenal properties (if there is such a thing). Rather, when advocates of pure cognitive phenomenology try to convey what the phenomenology of a thought consists in, they first identify the thought in terms of its intentional properties and *then* draw the attention to the difference that this thought makes in terms of what-it's-like-ness (see, e.g., Pitt [2004, 26–9]). This strongly suggests that we are actually not able to detect intentional on the basis of phenomenal properties, as the account in question would have to claim.

This last point speaks against a reliabilist version of the pure cognitive phenomenology approach, too. On such a view, the occurrence of the phenomenal properties in virtue of which the response event has its intentional properties reliably causes the right self-ascriptive belief. This approach would not face the first problem of the internalist variant. Nevertheless, if the view is to assign phenomenology a role at all, it would have to assume that the causal mechanism in question consists in an ability to discriminate pure cognitive phenomenal properties. And as we have seen above, it is just very implausible that we possess such a discriminatory ability independently of antecedent intentional interpretations.

Thus, expressive episodes should be seen as involving interpretation, no matter what precise account of cognitive phenomenology is adopted.

3.3 Features of the interpretive process

What consequences has the argument in the previous sub-section with regard to the specific features of the interpretive process that is required, as we have seen, for the formation of self-knowledge in expressive episodes?

First, the problem of discriminatory abilities that has emerged in the last part of our discussion speaks against the idea that the interpretation process could take non-sensory, pure

cognitive phenomenal properties as input. Regardless what precise view of cognitive phenomenology and its relation to intentionality is held, the interpretation process should be seen as being based on *sensory* data.

Second, such sensory data are not sufficient on their own as interpretive basis, either. Consider, for instance, an expressive episode in which the phenomenology of the response event consists in a simulation of 'Yes', or even a mere feeling of confidence, in response to the initial question: without taking into account the *context* of the response event within the expressive episode, no adequate interpretation of this phenomenology will be possible.

Third, it can be argued that very often, the interpretive process employs mindreading abilities that are at least very similar to those that we use in order to ascribe intentional states to other persons. It is uncontroversial that the sensory phenomenology of thought often consists of inner speech (cf. also Heavey and Hurlburt [2008]). In that case, self-ascriptive beliefs that we form in the course of expressive episodes have to be based, according to the view that I have been arguing for in this section, on interpretation of that inner speech. For instance, if I ask myself whether I believe that tomorrow will be Wednesday, and I find myself responding by rehearsing in inner speech 'Yes', I will have to (unconsciously) interpret that episode of inner speech as expression of my underlying belief that tomorrow will be Wednesday. So apart from the fact that the input data of the first-personal interpretation process are only *imitations* of phonological phenomenology, this case of interpretation is entirely parallel to one in which I interpret someone else's utterance 'Yes' in response to the question 'Will tomorrow be Wednesday?' as an expression of her underlying corresponding belief. As a consequence, parsimony recommends to assume that both forms of interpretation are carried out by the same mechanism, or at least by two closely related mechanisms.

But doesn't the first-personal case differ from the third-personal case insofar as when *I* say something (in overt or covert speech), I know *ex ante* what I want to say, and hence how

my utterance should be interpreted, whereas in the third-personal case, I normally only know *ex post* – by interpreting the utterance – what the speaker wants to say? I agree that in cases of deliberate speech, there is normally such an asymmetry. But in the episodes of inner speech that we are concerned with, we spontaneously respond to a given question, rather than deliberately rehearsing a piece of speech. And *if* such an episode of inner speech really constitutes the response event within an expressive episode in my terminological sense, our knowledge regarding its meaning (that is, about the fact that it expresses a particular attitude) *must* be of the *ex-post* type: otherwise, we would have to know in advance of the response event what our answer to the initial question is. This would either already amount to a self-ascriptive belief, or provide at least sufficient basis for such a belief. In either case, the self-ascriptive belief would not be based anymore on the response event; hence, the situation would not count as an expressive episode at all. – I therefore conclude that expressive episodes not only involve *interpretation*, but that this interpretation also uses capacities that are very similar to those that we employ in third-personal interpretation – at least in the (frequent) case in which the phenomenology of the response event consists in an episode of inner speech.

Throughout this section, I have focused on expressive episodes in which conscious *thoughts* figure as response events. However, the same line of argument applies to expressive episodes with response events that are public utterances. Such utterances can be treated in two ways. Either it is assumed that the utterance is preceded by a response event in thought, and expresses this thought (or a self-ascriptive belief based on it); or the utterance is understood as public equivalent to a response event in thought. In either case, expressive episodes that involve a public expression of the target attitude are merely a special case of expressive episodes with a conscious thought as response event, and the argument in this section equally applies to this case, too.

4. Self-knowledge as both interpretive and agential

If the foregoing argument is sound, the form of interpretationism that consists in the conjunction of claims (1) and (2) is mandatory even for rationalists. However, the argument in the previous sections leaves open whether the epistemology of situations in which we adopt new attitudes as the result of deliberation is interpretive or not. So the rationalist could respond to our argument by restricting the range of truly agential self-knowledge to situations of deliberation. But such a restrictive strategy is hardly satisfactory: as we have seen in section 2.3, it would be implausible to hold that in expressive episodes, our stance towards our attitudes is necessarily alienated or detached.

In this concluding section, I wish to argue that the restrictive strategy is not compulsory for rationalists: for as I hope to show, there is, contrary to what is usually assumed by both rationalists and interpretationist, no good reason to see a conflict between interpretive self-knowledge of the kind we have discussed in the last section, and the perspective of rational agency. In order to argue for this compatibility claim, I shall briefly point out how interpretationism leaves room for the most salient features that rationalists ascribe to self-knowledge. I begin with the aforementioned three features of high reliability, immediacy, and distinctiveness, and then turn to the more general issue of activity vs. receptivity.

Consider, first, *reliability*. A high success rate for our self-ascriptions of rational attitudes is entirely compatible with the idea that the knowledge in question is interpretive: directness and reliability are two logically independent features of epistemic access (as is rightly emphasized by Byrne [2012] against Carruthers [2011]). Regarding the self-knowledge that we acquire in expressive episodes, all that an interpretationist has to grant in order to account for the requisite reliability is the following: for something to be a rational attitude A of a rational agent, it must have a causal role which includes a disposition to cause, in the course of expressive episodes, a phenomenology (inner speech, mental imagery,

possibly irreducible cognitive phenomenology) that is very likely to be interpreted as expression of A by the interpretation mechanism; and those of our self-ascriptive beliefs that issue from expressive episodes are normally formed by virtue of this mechanism. Parallel conditions can be postulated if other types of situations in which we acquire self-ascriptive beliefs are given an interpretationist account – such as situations of deliberation.

It is possible, of course, that as a matter of fact, *our* rational attitudes do not actually satisfy such conditions, and that therefore, the idea that we normally relate to our attitudes as rational agents is illusory. Thus, if it turned out that the majority of our self-ascriptive beliefs issue from confabulation, we would not have the self-knowledge required for rational agency – our self-ascriptive beliefs would normally be wrong. But even if the extant literature on confabulation is taken to show that the ways in which we form self-ascriptive beliefs are less reliable than we may tend to think, it is an open question whether we are really unreliable about our current rational attitudes to a degree that would suffice to undermine our rational agency.[18]

Nor is interpretationism incompatible with the condition of *immediacy*, as long as this condition is understood in terms of *phenomenal* immediacy: as interpretationists hold that the interpretation process normally takes place unconsciously (including P-unconsciousness), our access to our attitudes can still be phenomenally immediate (cf. Gopnik [1993, 11]; Cassam [2010, 91–3]). In particular, it has been suggested that the interpretive access in question can take the form of an inner aspect-perception.[19] If this proposal is applied to our above account of expressive episodes, it follows that we directly experience response events *as* expressions of the target attitudes.[20] Such an experience would be interpretive and phenomenally immediate at the same time.

Next, I want to argue that interpretationism even leaves space for a significant structural difference between first- and third-personal access, and can insofar also take into account the *distinctive nature* of self-knowledge. In a third-personal case, when an ascription

of a rational attitude that is based on an earlier episode of interpretation conflicts with the way the other person presently behaves, there are three possible ways to react. We can (1) regard the present behaviour as a mistake (e.g., a case of weak will, or of absent-mindedness); we can (2) assume that the person has changed or abandoned the attitude in question in the meanwhile; and we can (3) revise our previous interpretation, assuming that we have misinterpreted the earlier evidence about the person's rational attitude. What option we go for will depend on details of the situation and background beliefs; but normally, all three options are in principle available in such situations.

By contrast, imagine an analogous first-personal case in which I have come to self-ascribe a particular attitude in an expressive episode, and now find myself behaving or thinking in a way that conflicts with this attitude. In this case, there seem to be normally only two options, which roughly correspond to the first two options in the third-personal case. I can (1) consider my present piece of behaviour as a mistake (and try to correct it). And I can (2) conclude that I have changed my attitude in the meanwhile (and correspondingly revise my self-ascriptive belief regarding my present attitude). But the third option from the third-personal case does not seem to have an equivalent here: it is *not* normally an option for me to doubt the original interpretation that I had adopted in the expressive episode, and to assume that I actually never have had the attitude that I had self-ascribed. So whereas in the third-personal case we regard our interpretations as open to revision, we normally treat our self-interpretations and our corresponding self-ascriptive beliefs as *authoritative* – we do not take into account the possibility that we may have misinterpreted our own mental states.

The resulting asymmetry is not a matter of the way in which we *arrive* at our own self-ascriptions, but of how we *treat* them – a matter of granting them authority by standardly relying upon them. It therefore can obtain even where the methods or mechanisms through which we arrive at our interpretations are essentially the same in the first- and the third-

personal case. Hence, interpretationism can allow for a significant first-/third-personal asymmetry, too.[21]

Still, one might wonder if the resulting form of interpretationism is really able to do justice to our role as rational agents – if it can allow, despite of the receptive character of the self-knowledge that we have discussed, for an *active* stance towards our attitudes. In particular, one may have the following worry.[22] As I have mentioned in the introduction, rationalists understand that active stance in terms of rational control over our attitudes, or the ability to adjust them to one's reasons. Therefore, a version of interpretationism that is meant to leave room for rational agency has to grant that if we find the attitudes which we learn about through expressive episodes unsupported by reasons, we normally revise them. But this revision makes the self-knowledge that we have gained in the expressive episode obsolete. So it can seem that insofar as interpretationism allows for rational control, it becomes unclear what contribution interpretation can make to self-knowledge at all. Instead, one might argue that what attitudes we have *beyond* the immediate context of expressive episodes depends on what our reasons are; so self-knowledge regarding these attitudes seems to require a method (such as the method of 'transparency') that takes into account those reasons.

This objection can be accommodated by looking more closely at the case in which I find an attitude unsupported upon self-ascribing it in an expressive attitude. To begin with, it is very plausible to assume that in such a case, the conscious revision of the unsupported attitude becomes possible *because* the attitude has become the object of a self-ascriptive belief. *Knowing* that one holds a particular attitude seems to be a presupposition for consciously assessing it in critical thought (cf. Shoemaker [1996, 240]). Yet precisely because the attitude is unsupported by one's reasons, an assessment of those reasons cannot suffice as a basis for knowledge that one has the attitude. Rather, it seems to take a *receptive* mode of access, such as interpretation within expressive episodes, in order to detect the attitude, and to subsequently adjust it to one's reasons. So in this case, interpretation contributes to self-

knowledge at least insofar as it provides the self-knowledge that is needed to get the process of conscious rational revision started.

But in addition, there is also reason to believe that interpretation is needed for the further course of rational revision, too. For becoming aware that one of my attitudes is unsupported by reasons does not always automatically make this attitude vanish. The attitude may initially continue to exist, and it may take some more reflection for me to convince myself entirely that things are not the way (or that I should not act the way, etc.) in which this attitude presents it. But if this is so, it is crucial for my exercise of rational control that I have a way of controlling the progress of that exercise – that I can test, for instance, whether I actually have abandoned the attitude in question. Once again, the requisite knowledge cannot be had merely on the basis of knowledge of what one's reasons are – for what we need to know in this case is to what extent we actually live up to the demand of those reasons. Instead, we seem to need once more a receptive form of knowledge here, and interpretation of the kind that we have been discussing seems apt again to provide that knowledge. So interpretive self-knowledge is crucial not only for getting started with, but also for carrying out, a central form of rational control. It should therefore be seen as making a crucial contribution to the self-knowledge that we have as rational agents.

I thus conclude that interpretive self-knowledge, despite its receptive character, is not incompatible with the active stance of agency: rather, it should be seen as an integral element of rational agency. As a consequence, there is no good reason to hold that rationalism and interpretationism as such are mutually exclusive: on the contrary, rationalists *can* and *should* accept interpretationism.[23]

---

1. Other philosophers (e.g., McGeer [1996]) have developed the basic idea that self-knowledge is
2. Some authors claim that the first-/third-personal asymmetry requires an epistemic difference in the *method* that we use to acquire self-knowledge (e.g. Byrne 2011). I will not treat this as essential part of the view in question.

3. It is more common to refer to the positions in question as 'theory theory' or 'inferentialist' accounts of self-knowledge. I prefer 'interpretationism' because this term leaves open which precise theory of mindreading is adopted, and whether the process leading to self-knowledge is, strictly speaking, inferential in nature or not (see note 5).
4. E.g., from developmental psychology: Gopnik (1993); from cognitive architecture and evolutionary biology: Carruthers (2011).
5. Similarly, the perception of facial expressions as expressive of emotions can be said to involve interpretation, although we normally would not be able to spell out the input data for the interpretation process in terms of explicit premises. (Without this qualification, interpretationism would founder on the objections levelled by Bilgrami [2006, 19–20]).
6. Carruthers (2009, 123) adds the further condition that information about context (such as the subject's circumstances and public behaviour) is accessed. The form of interpretation that I will argue for in section 3 fulfils this additional condition, too.
7. I shall remain non-committal here about the precise identity-conditions of mechanisms.
8. This view is adopted by Shoemaker (1996, 241) in his response to the regress problem.
9. For one thing, I do not believe that Shoemaker's arguments succeed in establishing a constitutive account (unless it is understood in the sense of the first form of modified universalism that I have discussed above). For a critique of Shoemaker's arguments and his positive account, cf., e.g., Peacocke (2008, 268–75).
10. It would not have to achieve such identification if there was a distinct mechanism for each type. But as Goldman remarks (2006, 239), this would lead to an implausible inflation of the number of requisite mechanisms.
11. Cf. also Goldman (2006, 240–1) for a similar critique of an analogous position.
12. For the sake of my discussion, I will assume that this is true for all aspects of the response event that are available to the meta-cognitive capacity. Alternatively, it might be held that the *content* of the response event is independently available to that capacity, e.g. through 'activation' in working memory. But this view will have to allow for some form of interpretationism, too. For normally, there will be further activated contents at the same time (e.g., of beliefs and intentions that are relevant to what one is presently doing). Hence, the meta-cognitive capacity would need a mechanism that identifies *which* of the various simultaneously activated contents belongs to the thought that presently occupies P-consciousness. This mechanism would have to *interpret* the phenomenal properties of the response event in terms of various simultaneously activated contents, and identify the content that best fits the phenomenology and the context. (Thanks to Tobias Rosefeldt for pressing me on this point.)
13. This awareness might consist, for example, in an accompanying, P-unconscious thought (cf. Rosenthal [2005, 14–5, 126]).
14. For the sake of our discussion, we can neglect the contrast between 'sensory' and 'quasi-sensory'.
15. On a rich notion of the sensory, one might hold that hearing something *as* bearer of a determinate linguistic meaning *is* a sensory phenomenon. In that case, there would be a strong sensory difference between both scenarios; but such hearing-as would itself require previous interpretation of sensory data.
16. It might be objected that the mere requirement of understanding inner speech (and similar phenomenology) is too 'easy' an interpretive achievement to yield a controversial position. But what is at stake is not whether the interpretation that self-knowledge is based on is a difficult task or not, but whether self-knowledge *is* based on interpretation or not (cf. Carruthers' reply to a similar objection by Petty and Briñol in his [2009, 169]).
17. At least if it is assumed in addition that the fact that a thought expresses a target attitude is part of the intentional properties that define the thought.
18. Cf., e.g., Goldman (2006, 231–4) for a critical discussion regarding the implications of confabulation findings for self-knowledge.
19. For the idea of inner aspect-perception, cf. Gopnik (1993, 10–12). Carruthers (2011, 87) seems to allow for a similar possibility regarding inner speech ('global broadcast' of interpretation that is 'bound' into sensory content; cf. ibid., 50–1).

20. Cf. also notes 15 and 19.
21. By contrast, versions of rationalism that require, in addition, that the first-/third-person asymmetry consists in an epistemic difference between methods of gaining self-knowledge (cf. note 2) *are* incompatible with interpretationism.
22. Thanks to an anonymous referee for drawing my attention to this problem.
23. This move could seem to be precluded by my above account of expressive episodes: why should, on that account, the phenomenal properties of response events be more than 'the testimony of an alien voice, whose rational significance for my thinking now was an open question' (as Boyle [2009, 160] writes in a related context)? Put very briefly, I would reply: because it is part of the phenomenology of P-conscious thoughts figuring as response events that I experience them as *my own responses* to the initial questions (e.g., through 'context integration', see Martin and Pacherie [2013, 115–7]).

Notes on contributor

Franz Knappik has received his doctorate from Ludwig-Maximilians-Universität, Munich, in 2011. Since the same year, he has been Lecturer in philosophy at the Humboldt-Universität zu Berlin. His main research interests are in the philosophy of mind and the history of philosophy (German Idealism).

References

Bayne, T., and Montague, M., eds. 2011. *Cognitive Phenomenology*. Oxford: Oxford University Press.
Bilgrami, A. 2006. *Self-Knowledge and Resentment*. Cambridge, MA: Harvard University Press.
Block, N. 1997. On a Confusion about a Function of Conscience. In *The Nature of Consciousness. Philosophical Debates*, ed. N. Block et al., 375–415. Cambridge, MA: MIT Press.
Boyle, M. 2009. Two Kinds of Self-Knowledge. *Philosophy and Phenomenological Research* 78 (1): 133–64.
Burge, T. 1996. Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society* 96: 91–116.
Byrne, A. 2011. Transparency, Belief, Intention. *Proceedings of the Aristotelian Society Supplementary Volume* 85 (1): 201–21.
Byrne, A. 2012. Review of *The Opacity of Mind*, by Peter Carruthers. *Notre Dame Philosophical Reviews* 2012.05.11, URL: https://ndpr.nd.edu/news/30799-the-opacity-of-mind-an-integrative-theory-of-self-knowledge/, accessed 9/11/2014.
Carruthers, P. 2009. How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition. *Behavioral and Brain Sciences* 32 (2), 121–182.
Carruthers, P. 2011. *The Opacity of Mind. An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.
Cassam, Q. 2010. Judging, Believing and Thinking. *Philosophical Issues* 20 (1): 80–95.
Churchland, P. 1999. *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind*. Revised edition. Cambridge, MA: MIT Press.
Crimmins, M. 1992. Tacitness and Virtual Beliefs. *Mind and Language* 7 (3): 240–63.
Gertler, B. 2011a. *Self-Knowledge*. Abingdon, New York: Routledge.
Gertler, B. 2011b. Self-Knowledge and the Transparency of Belief. In *Self-Knowledge*, ed. A. Hatzimoysis, Oxford: Oxford University Press, 125–45.
Goldman, A. 2006. *Simulating Minds. The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.
Gopnik, A. 1993. How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality. *Behavioral and Brain Sciences* 60 (1): 1–14.
Heavey, C., and Hurlburt, R. 2008. The Phenomena of Inner Experience. *Consciousness and Cognition* 17 (3): 798–810.
Horgan, T., and Tienson, J. 2002. The Intentionality of Phenomenology and the Phenomenology of Intentionality. In *Philosophy of Mind. Classical and Contemporary Readings*, ed. D. Chalmers, 520–33. New York / Oxford: Oxford University Press.
Kriegel, U. 2011. *The Sources of Intentionality*. Oxford: Oxford University Press.

Martin, J.-R., and Pacherie, E. 2013. Out of Nowhere: Thought Insertion, Ownership and Context Integration. *Consciousness and Cognition* 22 (1): 111–22.
McGeer, V. 1996. Is 'Self-Knowledge' An Empirical Problem? Renegotiating the Space of Philosophical Explanation. *The Journal of Philosophy* 93 (10): 483–515.
Moran, R. 2001. *Authority and Estrangement. An Essay on Self-Knowledge*. Princeton NJ: Princeton University Press.
Nichols, S., and Stich, S. 2003. *Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Clarendon Press.
Peacocke, C. 2008. *Truly Understood*. Oxford: Oxford University Press.
Pitt, D. 2004. The Phenomenology of Cognition. Or: What Is It Like to Think That P? *Philosophy and Phenomenological Research* 69 (1): 1–36.
Pitt, D. 2011. Introspection, Phenomenality, and the Availability of Intentional Content. In Bayne and Montague 2011, 141–73.
Rosenthal, D. 2005. *Consciousness and Mind*. Oxford: Oxford University Press.
Shah, N., and Velleman, J. D. 2005. Doxastic Deliberation. *The Philosophical Review* 114 (4): 497–534.
Shoemaker, S. 1996. Self-Knowledge and 'Inner Sense'. In *The First-Person Perspective and Other Essays*, Cambridge: Cambridge University Press, 201–68.