**Volume X ◊ Issue 3**

2019

# Logos & Episteme

an international journal
of epistemology

**Romanian Academy**
**Iasi Branch**
◆◆◆
**“Gheorghe Zane” Institute**
**for Economic and Social**
**Research**

**Contact address:**

Institutul de Cercetări
Economice și Sociale „Gh.Zane"

Iaşi, str.T.Codrescu, nr.2, cod 700481

Tel/Fax: 004 0332 408922

Email: logosandepisteme@yahoo.com

http://logos-and-episteme.acadiasi.ro/

https://www.pdcnet.org/pdc/bvdb.nsf/journal?openform&journal=pdc_logos-episteme

ISSN 2069-053

# TABLE OF CONTENTS

RESEARCH ARTICLES

DISCUSSION NOTES/ DEBATE

# RESEARCH ARTICLES

# INNER SPEECH AND METACOGNITION: A DEFENSE OF THE COMMITMENT-BASED APPROACH

Víctor FERNÁNDEZ CASTRO

ABSTRACT: A widespread view in philosophy claims that inner speech is closely tied to human metacognitive capacities. This so-called format view of inner speech considers that talking to oneself allows humans to gain access to their own mental states by forming metarepresentation states through the rehearsal of inner utterances (section 2). The aim of this paper is to present two problems to this view (section 3) and offer an alternative view to the connection between inner speech and metacognition (section 4). According to this alternative, inner speech (meta)cognitive functions derivate from the set of commitments we mobilize in our communicative exchanges. After presenting this commitment-based approach, I address two possible objections (section 5).

KEYWORDS: inner speech, metacognition, commitments

## 1. Introduction: Talking to Oneself

Metacognition or thinking about thinking is a fundamental human cognitive capacity.[1] This capacity is devoted to evaluating, predicting or modifying our cognitive performances, so it endows us with a unique cognitive and behavioral flexibility and adaptability. Several authors have claimed that there is a constitutive connection between these metacognitive capacities and the linguistic ability of talking to oneself,[2] so humans are able to flexibly modify, regulate and access their cognitive processes because they are able to structure their own mental states in a linguistic format through self-directed talk. This so-called *format view of inner speech*[3] claims that capturing our mental states in linguistic format allows

[1] Michael T. Cox, Anita Raja, and Eric Horvitz, eds., *Metareasoning. Thinking about Thinking* (Cambridge, Mass.: MIT Press, 2011); John Dunlosky and Janet Metcalfe, eds. *Metacognition* (Los Angeles: SAGE Publications, 2019).

[2] Daniel C. Dennett, *Consciousness Explained* (London: The Penguin Press, 1991); Ray Jackendoff, *The architecture of the language faculty* (Cambridge, Mass.: MIT Press, 1997); Andy Clark, *Being there* (Cambridge, Mass.: MIT Press, 1997); Jose Luis Bermudez, *Thinking without words* (Oxford: Oxford University Press, 2003).

[3] Fernando Martínez-Manrique and Agustín Vicente, "The activity view of inner speech," *Frontiers in psychology* 6 (2015): 232.

us to acquire the metarepresentational capacities underlying the unique human metacognitive competence of modifying and accessing our own mental states

The aim of this paper is to defend an alternative to the format view as a theory of the connection between inner speech and metacognition. The alternative I put forward is based on a Commitment-Based approach to communication and inner speech according to which the main purpose of communication is to establish commitments and entitlements to coordinate agents; so, the cognitive function of inner speech derivate from this social function of outer speech. The structure of the papers goes as follows: Firstly, I present the format view along with two objections (section 2 and 3). These objections challenge two central ideas of the format view: (1) the notion of metacognition as access, and (2) the idea that metacognition requires metarepresentations. In section 4 and 5, I introduce the commitment-based view and how it can account for the different cognitive functions associated with metacognition. Finally, in section 6, I address two possible objections to the alternative.

## 2. The Format View of Inner Speech

Inner speech is often defined as the phenomenon we experience when talking silently to ourselves. The contemporary interest on the phenomenon starts with the publication in English of the work of the Soviet psychologist Lev Vygotsky who, after realizing that children systematically talk to themselves out loud (private speech), started to study the role of private and inner speech in the development of high cognitive capacities.[4] In contemporary psychology, the research on private and inner speech has resulted into different studies that connect inner speech with different cognitive capacities, including conscious control, working memory and attention.[5]

Besides this empirical evidence, there are different debates on the format, nature, and function of inner speech.[6] The fundamental question underlying those

[4] Lev S. Vygotsky, *Thought and language* (Cambridge, Mass.: MIT Press, 1984, Original work published 1934).

[5] Rafale Diaz and Laura Berk, eds. *Private speech: From social interaction to self-regulation* (Hillsdale, N.J.: L. Erlbaum, 1992); Daniel Gregory “Inner speech, imagined speech, and auditory verbal hallucinations,”*Review of Philosophy and Psychology* 7,3 (2016): 653–673; Adam Winsler, Charles Fernyhough and Ignacio Montero, eds. P*rivate speech, executive functioning, and the development of verbal self-regulation* (Cambridge: Cambridge University Press, 2009).

[6] Martínez-Manrique and Vicente, “What the...! The role of inner speech in conscious thought,”*Journal of Consciousness Studies* 17 (2010): 141–167; Keith Frankish,“Evolving the linguistic mind,”*Linguistic and Philosophical Investigation* 9 (2010): 206–214;Peter Langland-Hassan, “Inner speech and metacognition: in search of a connection,”*Mind & Language*29 (2014):

debates is why do we talk to ourselves? A widespread answer in philosophy of mind maintains that we talk to ourselves in order to display metacognitive abilities, that is, we talk to ourselves to consciously access our own thoughts.[7] This so-called format view of inner speech associates the function of our self-talk to some structural features of the linguistic format. The main thesis is that language, in virtue of these features, is the only representational vehicle that allows codifying mental states in a way that can be objects of further thoughts. In other words, language facilitates what Clark calls second-order dynamics. Language codifies thoughts that can be brought into working memory in a way that attention can be directed to them, and thus, be objects of conscious access. Although these authors share the perspective of inner speech as a metacognitive facilitator, they differ about which properties make language appropriate for such function. In this sense, for instance, Clark argues that the features of language that allows us to recruit it for cognitive purposes are its context-dependency and neutral modality.[8] On the other hand, Bermudez considers that, given that all conscious access must be carried out on perceptual modality, language is the only representational vehicle that allows personal level conscious access and is, at the same time, a structured vehicle. Contrary to other personal vehicles as images, language is structured and compositional. Contrary to other structured vehicles as mentalese inner speech is a vehicle we can consciously access.[9] Thus, inner speech is the only representational format that facilitates second-order dynamics to conceptually structured thoughts.

This picture on inner speech face several problems related with some of its fundamental theses.[10] However, the aim of this paper is to reveal the problems of the view regarding two fundamental assumptions; namely, how the model assigns a central role to metarepresentations in metacognitive capacities, and how metacognition is understood in terms of access to mental states or processes. First, according to the format view, when an agent experiences an episode of inner

511– 533; Peter Langland-Hassan and Agustin Vicente, eds. *Inner Speech: New Voices* (USA: Oxford University Press,2018).

[7] Jose Luis Bermudez, *Thinking without words*; Andy Clark, *Being there*; Daniel C. Dennett, *Consciousness Explained*; Jackendoff, *The architecture of the language faculty*.

[8] Clark, *Being there,* 178.

[9] Jerry Fodor, *The language of thought* (Cambridge, Mass.: Harvard university press, 1975)

[10] Apart from Martínez-Manrique and Vicente, "The activity view of inner speech," the problems of the format view has been emphasized by Marta Jorba and Agustin Vicente, "Cognitive phenomenology, access to contents, and inner speech," *Journal of Consciousness Studies* 21, 9-10 (2014): 74-99; Víctor Fernández Castro, "Inner Speech in Action," *Pragmatics & Cognition* 23, 2 (2016): 238-258; or Bart Geurts, "Making sense of self talk," *Review of Philosophy and Psychology* 9, 2 (2018): 271-285.

speech, for instance when someone utters silently a sentence such as 'the unemployment in Europe have decreased at the expense of worker's rights,' she can access her own mental state because, through the access of this internal episode, she can infer the state that she believes that the unemployment in Europe have decreased at the expense of worker's rights. So, metacognition requires forming representations about that mental state in order to perform other actions as controlling or regulating the state in question. This metacognitive capacity can be understood as a device that takes the content of an inner speech episode as an input and produce a metarepresentational state of the form 'I believe (desire, imagine) that P' as an output. Likewise, inner speech episodes allow us to access to our mental states as far as facilitates the generation of metarepresentations with the form 'S verbs P.' Understanding metacognition in metarepresentational terms is not new. As Proust has shown, considering that metacognitive capacities rely upon the ability to form metarepresentation is widely shared assumption in cognitive sciences and philosophy.[11] The innovation of the format view, then, is connecting these metarepresentational capacities to inner speech and the capacity of putting thoughts in a linguistic format.

Second, the format view is strongly committed to a particular notion of metacognition as access.[12] Again, as Proust argues, most of the philosophical approaches to metacognition in philosophy and cognitive sciences assume that the second-order regulation and control of cognitive processes require the subject to be able to access, either through introspection or inference, to the contents of the first level processes and states. So, humans could not regulate, evaluate and modify their first-order mental processes and states without having access to such processes and states. In the format view, capturing our thoughts through inner episodes allow us

[11] Joëlle Proust has examined this and other aspects the standard view of metacognition (see Joëlle Proust, "Metacognition," *Philosophy Compass* 5, 11 (2010): 989-998; *The philosophy of metacognition: Mental agency and self-awareness* (Oxford UK: Oxford University Press, 2013). She mentions as proponents of such standard view to John Flavell, "Metacognition and Cognitive Monitoring: A New Area of Cognitive- Developmental Inquiry," *American Psychologist* 34 (1979): 906–911; Alan Leslie, "Pretense and Representation: The Origins of Theory of Mind,'' *Psychological Review* 94 (1987): 412–26; Josef Perner, *Understanding the Representational Mind* (Cambridge, Mass.: MIT Press, 1991); Alison Gopnik, "How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality," *Behavioral and Brain Sciences* 16, 1 (1993): 1–15; Peter Carruthers, "Meta-cognition in Animals: A Skeptical Look," *Mind and Language* 23 (2008): 58–89; "How Do We Know Our Own Minds: The Relationship between Mindreading and Metacognition." *Behavioral and Brain Sciences* 32 (2009): 121–82

[12] Joëlle Proust,"Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition?," *Synthese* 159, 2 (2007): 271-295.

to access our mental states and processes because we can self-ascribe such states by forming metarepresentation of the form 'I verb P.' So, inner speech episodes facilitate the second-order access our metacognitive capacities consist in.

## 3. Telepaths and the Young Rich Communist

This section brings into focus two objections of the format view, which lay on the two aforementioned assumptions. That is, the idea that metacognition must be understood in terms of access and the idea that metacognition is carried out in a metarepresentational format. These two objections prepare the ground for defending the commitment-based approach I characterize in the next section.

The first problem to the format view lies on the restricted power of the notion of metacognition as access to account for how inner speech make a difference for explaining the cognitive and behavioral flexibility associated to metacognition. In principle, the explanandum of a theory of this type must be to explain how inner speech, as long as it endows linguistic creatures with certain metacognitive capacities, can account for some of the patterns of actions and mental skills associated with thinking about thinking, for instance, cognitive flexibility or the capacity to evaluate and regulate actions. However, the format view seems to fail to achieve this objective. Although the format view gives a reasonable explanation of how a creature can access to her mental states, it is hard-pressed to explain how this access is translated into certain special cognitive abilities. For instance, why the metacognitive capacities associated with inner speech facilitate the rise of cognitive regulation or flexibility. Part of the obstacle a defender of the format view must address is that, although the position claims that inner speech brings certain mental states into consciousness, it does not explain how this 'bringing mental states into consciousness' plays a role in regulating or evaluating first- order processes. As McGeer argues, having access to our own mental states would play a role analogous to the role of a telepath that could read our mind, seeing our mental states and processes, but could not exercise any type of power to modify or regulate them.[13] If the format theory aims to explain which function the inner speech plays in the acquisition of metacognitive capacities, the theory should not only explain how certain distinctive mental states or processes are produced, but also how accessing those states and processes make a difference for the type of abilities we usually associate with metacognition (control of attention, regulation, cognitive flexibility).[14] In other words, monitoring our

[13] Victoria McGeer, "The Moral Development of First-Person Authority," *European Journal of Philosophy* 16, 1 (2007): 81-108.

[14] See Proust, *The philosophy of metacognition,* 29-78.

mental states is not sufficient for explaining the cognitive and behavioral flexibility associated with metacognition, and thus, the format view must be regarded as incomplete.

A possible way out to this problem may appeal to the notion of metarepresentation. The defender of the format view could argue that the metarepresentational states that inner speech produces could modify certain pattern of behavior and cognition in a flexible way. For instance, image a physicist on the way home to finish an article that the editors of a journal have been waiting for. At the moment, she is entering her house, an utterance crosses her mind 'the dinner!' Suddenly, she remembers she has invited some friends for dinner and the fridge is empty. 'I gotta go to the grocery store.' The physicist changes her route and stops at the grocery store before going home. According to the format view, inner speech episodes could allow the agent to access her mental states (remembering that she has organized a dinner, the belief that the fridge is empty and the belief that she must go to the grocery store) in a way that she can abort her action of going home and trigger the action of walking toward the store.

However, this solution does not solve the problem. Notice that explaining how behavioral and cognitive flexibility derivate from inner speech does not seem to necessarily rely on metarepresentational states. In principle, the physicist's cognitive processes can be carried out by first-order processes. The appropriate behavioral pattern can be triggered by bringing out the appropriate information without a self-ascription of the given mental states; for instance, bringing out the information that she should go to the store and that she has a dinner tonight rather than the self-attribution of such mental states. It is the mental states per se and not the self-attribution of these states what seems to play a role in the realization of the action. As Jorba and Vicenteargue, if the function of inner speech is to put on a propositional content in a format that allows our 'inner eye' to access the content, then the format theory explains how we can produce a metarepresentational state, e.g. 'I believe that P,' from an utterance with the content P.[15] However, if the outcome of the cognitive processes involving inner speech episodes are second-order states, it is difficult to see how they can affect the first order states that, after all, are the producers of the behavior at stake. As Martínez-Manrique and Vicente say:

> [T]he model they propose seems to only be able to explain how IS gives us knowledge of what and how we think. Let's say that by using sentences of our language, we are able to have some kind of object before our minds. What do we

[15] Jorba and Vicente, "Cognitive phenomenology, access to contents, and inner speech;" see also, Martínez-Manrique and Vicente, "The Activity View of Inner Speech."

> gain with that? Presumably, we only gain knowledge about what we are thinking. We "see" the sentence, get its meaning, and reach the conclusion "ok, I'm thinking that p." This knowledge about what and how we are thinking may be very useful, of course, but we would say that this is only a use of IS, among many others. The account, in any case, does not explain how thought-contents are made access-conscious.[16]

That is to say, gaining access to our mental states by producing a self-ascribed metarepresentational state does not account for how our actions or the first-order mechanism are monitored, evaluated or regulated. Furthermore, the format view does not seem to respect the way we experience the inner speech episodes. When our physicist talks to herself 'the dinner!' or 'I should go to the grocery store,' she is encouraging herself to perform the action in the same way she would do it when directing these sentences to someone else. In this sense, the type of experience associated with the inner speech act is analogous to the external speech act but it does not seem to bear any resemblance with our ascriptions of mental states as the emphasis on the metarepresentational aspects suggests. In this sense, the format view does not respect our intuitions regarding how we experience inner speech episodes.

Certainly, the defenders of the format view could exploit other argumentative strategy. For instance, defending that the inner speech episodes that lead to self-ascriptions of the type 'I believe that P' or 'I desire that P' play a decisive role for a special kind of metacognition: future directed self-control. Future directed self-control requires evaluating our mental states and explore the type of genuine actions and processes that derivate from these ascriptions. In this sense, the defender of the format view could attribute to inner speech some kind of cognitive control over the behavioral consequences of their past, present and future mental states. Vierkant has illustrated this move through an example of Parfit where a young communist wins the lottery.[17] The young communist knows that rich people uses to be conservative, so he considers that if he does not get rid of the money (donating), he will become someone who does not want to be in the future, a conservative. So, the young communist is in the difficult position of donating the money and stick her ideals, or enjoying a comfortable life but becoming someone that he now would detest. The kind of mental skills the young communist engages in his considerations require self-ascribing mental states to

[16] Martínez-Manrique and Vicente, "The Activity View of Inner Speech," 4-5.

[17] See Tillman Vierkant, "What metarepresentation is for," in *The foundations of metacognition,* eds. Michael Beran, Johannes Brandl, Josef Perner, and Joëlle Proust (Oxford: Oxford University Press, 2012). The example appears in Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984).

himself and his future self, that is, metarepresentations. In this view, the defenders of the format view can embrace the idea that inner speech, as a producer of metarepresentations, will allow the young communist to attribute mental states to himself and his future self in order to evaluate which pattern of action to follow in the present given his attributions. This and analogous cases, where metacognitive capacities involve self-ascriptions, seem to be a plausible way for resisting the onslaughts against the format view.

This brings me to the second objection. Notice that the rationale for the format view is that inner speech facilitates the detection of underlying mental states that, after being metarepresented, we can manipulate. This idea assumes that our inner speech episodes voice or express the causally efficacious mental states that compose our first-order processes. However, this idea conflicts with empirical evidence regarding the phenomena of confabulation.[18] These studies show that humans are not always aware of the real causes of their actions, and in fact, they systematically provide ad hoc reasons to rationalize them. For instance, in the classic experiments carried out by Nisbett and Wilson, several subjects were asked which pair of panties they prefer and why. The panties were distributed on a table in a way that the subjects chose them by the distribution but they appeal to aspects such as the elasticity and the quality even though the panties were the same. These and other studies speak in favor of the idea that our reasons often are an instance of confabulation. Following this reasoning, it expectable to assume that our inner speech episodes do not necessarily voice our real mental states, and thus, it would be problematic to assume that the mental states the young communist attribute to his present self really reflect his mental states. Likewise, it is not clear how the mental states he ascribes to himself were real descriptions of his current mental states, and thus, played a causal role to modify his behavior for non-ending up being a conservative old person.[19]

Furthermore, even accepting the format view as an accurate explanation of this kind of metacognitive control, the explanatory power of the theory is too

[18] Richard Nisbett and Timothy D. Wilson, "Telling more than we can know: Verbal reports on mental processes," *Psychological review* 84, 3 (1977): 231; Michael Gazzaniga, *The mind's past* (Berkeley: University of California Press, 1998); Timothy D. Wilson, *Strangers to ourselves* (Cambridge: Belknap. 2002); Thalia Wheatley, "Everyday confabulation," In *Confabulation: views from neuroscience, psychiatry, psychology, and philosophy*, ed. William Hirstein (New York: Oxford University Press, 2009).

[19] Admittedly, not all versions of metacognition as access necessarily have troubles for explaining confabulation. An instance of this is Peter Carruthers, *The opacity of mind: an integrative theory of self-knowledge* (New York: Oxford University Press, 2011). However, they still would have to answer the telepath argument. Thanks to Tobias Störzinger for bringing my attention to this.

restricted. Although the cognitive function the young communist exercises could be accurately captured by the format view, the explanatory power of the theory is restricted to the cases involving metarepresentations, leaving aside cases where we directly control our first-order processes and behavior without such metarepresentations. As a conclusion, the format view cannot give a satisfactory explanation of how inner speech, as facilitator of second-order access, provides the acquisition of metacognitive capacities that regulate, modify or evaluate our cognition and behavior.

## 4. A Commitment-Based Approach to Inner Speech

In the previous section, I offered several arguments against the format view of inner speech. The aim of this section is to provide an alternative to the format view. This alternative, known as commitment-based approach, has been recently proposed by Geurts as an appropriate understanding of the cognitive functions of inner speech.[20] For the purpose of this article, I concentrate on how this approach can convincingly account for the role of inner speech in metacognition.

The commitment-based approach starts from the idea that the functions of inner speech derivate from the functions that speech acts play in coordinating agents in social interactions.[21] One way to capture how speech acts facilitate coordination between agents is by attending to how they modify the normative statuses of the speakers and her audience in terms of the commitments, duties and enabling conditions the speaker and audience undertake.[22] For instance, Geurts presents the idea as follows: "Commitment is a sine qua non for action coordination: social agents must rely on each other to act in some ways and refrain from acting in others. Commitments are coordination devices, and the main purpose of communication is to establish commitments."[23] Similarly, Kukla and Lance understand speech acts in terms of pragmatic input and outputs, where the

[20] Geurts, "Making sense of self talk."

[21] The idea that the function of inner speech derivates from the social function of outer speech is often traced back to Lev S. Vygotsky, *Thought and language*. For contemporary versions of these idea see Martínez-Manrique and Vicente, "The activity view of inner speech;" Jorba and Vicente, "Cognitive phenomenology, access to contents, and inner speech,"Fernández Castro, "Inner Speech in Action;" or Geurts, "Making sense of self talk."

[22] Robert Brandom, *Making it explicit: Reasoning, representing, and discursive commitment* (Cambridge: Harvard university press, 1998); Rebecca Kukla and Mark Norris Lance, *'Yo!'and'Lo!': The Pragmatic Topography of the Space of Reasons* (Cambridge: Harvard University Press, 2009), Bart Geurts, "Communication as commitment sharing: speech acts, implicatures, common ground," *Theoretical linguistics*(2019).

[23] Geurts, "Making sense of self talk," 8.

inputs are a set of enabling conditions and the outputs are a set of commitments, duties and entitlement the speaker and the audience undertake when recognizing the force and content of the speech act. In these views, the commitments we undertook when performing a speech act can be seen in terms of new possibilities for action. For instance, If I promise to someone that I will go with her to the theater, I am expressing a set of commitments with particular patterns of actions, including being at the theater at the time we stipulate. Certainly, not all speech acts present these direct goal-oriented commitments but even when one performs an assertion, the speaker is exhibiting certain commitment with what is rationally and socially expected from this assertion. For example, if I assert that the ice of the lake is dangerously thin, I am committing myself with future patterns of actions my audience is entitled to expect: that I will not skate on the ice or that I will warn other people of the danger. In other words, asserting something is expressing certain commitments with actions that our audience may expect us to follow.

Notice that carrying out a speech act does not necessarily involve we are in a particular mental state. As Geurts puts it:

> Commitments are obligations, and although they may be underwritten by suitable mental states, it is not necessary that they are. Insincere commitments are as binding as sincere ones, and there are unintended commitments, too. If I raise my hand at an auction, I thereby commit myself to be making a bid for whatever is currently under the hammer, even if I have no intention of doing so. True, I can try to get out of my commitment, for example, by arguing that I was only waving away a fly, but that presupposes there is a commitment to be undone.[24]

The patterns of actions associated with the commitments that follow from a particular speech act do not necessarily rely on the assumption that we are in particular mental state causally connected to these actions. Instead, the theory assumes that certain normative structures (rational and social) police our interactions in a way that connect the content of our commitments with such patterns of actions. For instance, we know what to expect from someone asserting that the ice is dangerously thin because we know what an agent ought to do in such circumstances given the rational and social structures that regulate our actions.

The commitment-based approach can help us to explain the social functions of our speech acts. The main advantage of this view is that it can account of the role of our speech acts in social coordination without reducing them to a mere exchange of information. Given that, the view is better posed to explain the speech acts whose function cannot be explained in terms of the information they provide

[24] Geurts, "Making sense of self talk," 9.

to the audience, that is, speech acts such as commands or promises, whose function does not seem to rely on how the audience gain certain piece of knowledge. Furthermore, the approach gives an automatic explanation of how our speech acts are connected to our social actions, so how they facilitate the coordination between speaker and audience.

This theoretical apparatus allows us to account the cognitive functions of inner speech in terms of the social functions of outer speech. That is, the inner speech episodes play a functional role in our cognitive machinery that is analogous to the role that external speech acts play in our social interactions. When someone asserts internally that ice of the lake is too thin, one is giving rise to private commitments with what is followed from the ice of the lake being too thin. So, she can regulate her actions and align her mental states in accordance with the commitments associated with the content of the assertion. Similarly, when an agent privately commands something to herself go to the store, she gives rise to certain goal-directed commitment to perform the action of going toward the store.

At this point, one may object that there is an important disanaology between outer and inner speech. Notice that, according to the commitment-based approach, the main function of communication is to coordinate agents. However, it is not entirely clear what exactly is the analog to coordination in the case of self-talk. In other words, if the function of inner speech derivate from the coordinating role of outer speech, then there must be a clear analog for coordination in the inner case. In order to address this challenge, one may argue that the function of outer speech for coordinating agents lies on the entitlements and commitments our speech instantiates. Once we learn how outer speech are associated to different patterns of action and cognition via those commitments and entitlement, we can rehearse such episodes in order to trigger the appropriate patterns.[25]

[25] Further, one may argue that, as for the case of intentions, inner speech episodes, as prompters of commitments, can promote intra-personal coordination by aligning volitional attitudes and practical reasoning. For instance, if I say to myself 'I will take the bus earlier tomorrow', this episode can instantiate a commitment that will help me to align my desire-like attitude toward intending to take the bus with the practical reasoning capacities necessary to find the more rational way to perform the action. For such a view regarding intentions see Michael Bratman, *Intention, plans, and practical reason* (Cambridge, MA: Harvard University Press, 1987) and Elisabeth Pacherie, "Conscious Intentions: The Social Creation Myth," *Open MIND* 29 (2015).

## 5. The Metacognitive Functions of Inner Speech

This section aims to account for the metacognitive functions associated with inner speech without postulating second-order access mechanisms o metarepresentational capacities. To see the contrast, notice that the format view appeals to the representational information included in the inner speech episode that produces a metarepresentation of the agent being in certain mental state in order to explain cognitive and behavioral flexibility. As I argued before, the two fundamental problems of this view are that self-ascriptions do not necessarily involve the capacity of modifying our first-order mental processes and actions. So, we can conceive circumstances where an agent ascribes to himself a particular mental state but this ascription does not make any difference. Furthermore, we can conceive several circumstances where agents regulate their actions and mental processes without having access to these states. In other words, intervening our own cognition and action do not require metarepresenting or accessing our mental states.

In order to see how the commitment-based approach can explain the connection between inner speech and metacognition, consider again the example of the physicist explained in section 3. The physicist privately utters the expression 'the dinner' which make her remember she has a dinner that night. Furthermore, she says to herself 'I should go to the grocery store' after considering she did not have food at home. The rationale behind the idea that the action of the physicist exhibits a kind of metacognitive endeavor rely on the fact that she refrains to perform the action she was doing (going back home) and triggers a new action on the light of new considerations. In this sense, she evaluates the situation and regulates her cognitive mechanisms to change her mind and carry out the action of going to the store instead of going home. The problem of the format view is that the outcome of the physicist's chain of reasoning is a self-ascription that in principle does not necessarily involve to regulate her action. Furthermore, it is hard to see how we can understand her regulatory capacities in terms of access to a mental state, especially when her private episode 'I should go to the store' does not seem to be a previous mental state in the physicist cognitive machinery, rather than a conclusion she has arrived from an episode of reasoning considering the situation. Given that, the format view should accept that the mental state represented by the private speech 'I should go to the store' was previously instantiated in the physicist's mind or abandon the idea that this case represents a case of metacognition in terms of access.

In the commitment-based approach, we can account for the case of the physicist in terms of evaluation and regulation. The metacognitive capacities displayed has to do with evaluating an action or mental processes in accordance with certain commitments and regulating first-order mental processes and patterns of action to align them with these commitments. When the physicist brings into consciousness her memory episode through the expression 'the dinner' she evaluates her current actions in terms of the commitments the utterance expresses. Thus, she refrains to go back home when considering her utterance gives rise to certain commitments her current action is not instantiating. In other words, her current action was not conforming the expected patterns given the restrictions

imposed by the commitments of having a dinner that night. On the other hand, when she concludes that she should go to the store, she is privately committing herself with the appropriate pattern of action, and thus, she can regulate her actions in accordance with such commitments. In this sense, the inner utterance expresses the same set of commitment with actions that the sentence will express when used in a conversation with the purpose of coordinating with another person.

This position differs from the format view in two fundamental aspects. Firstly, metacognition is associated with the notions of evaluation and conformation, rather than to the notion of access. When we assert P privately, we express a set of commitments that draw a cognitive trajectory we tend to conform in order to perform what these commitments prescribe us to do, that is, self-imposed constraints to our actions. In this sense, the commitment-based approach allows us to account for the metacognitive function of inner speech in terms of evaluation and conformation rather than in terms of access. Following Proust's idea, the type of cognitive and behavioral flexibility associated with metacognition does not require the agent to access her own mental states. In my view, rather revealing our previous mental states, our metacognitive capacities shape our cognition and action by triggering different prospective patterns we are inclined to follow given the commitments that the private episodes of inner speech generate.[26]

Secondly, respecting our intuitions, the metacognitive function of inner speech is not related to the notion of metarepresentation. Modifying our cognitive capacities in a flexible way does not require being able to self-ascribe mental states. In several occasions, the regulation or evaluation of our cognition and action do not require engaging in metarepresentational thinking. In fact, we often engage in reasoning chains that lead us to a private judgment that we do not hold before, and thus, do not represent previous mental states. When we arrive at these judgments we can modify or regulate our actions in the light of the commitments these judgments without the necessity of self-ascribe any particular mental state. In

[26] Proust, *The philosophy of metacognition,* 53-78.

other words, the effective power of the inner sentence to instantiate the appropriate pattern of action does not require the person to be in the state associated with the sentence, and far less, to represent such mental states.

## 6. Objections

In the previous sections, I have offered a theoretical model of inner speech that account for some metacognitive functions without appealing to metarepresentations or taking for granted that metacognitive capacities require accessing mental states. This move enables us to get around the concern of the format view of inner speech. However, one may wonder whether or not embracing the commitment-based approach could give rise to another type of problems. In principle, there are two main objections one may envisage for the commitment-based approach. Firstly, one may argue that future directed self-control (see section 3) fall out of the explanatory reach of the commitment-based approach. Secondly, one may consider that the notion of speech acts in terms of commitments is problematic or, at least, unnecessary for explaining the function of inner speech. This section is devoted to addressing these two objections.

For addressing the first problem, consider again the case of the young rich communist. As we have seen, Vierkant argues this case exemplify a kind of metacognitive capacity that cannot be performed without the metarepresentations and access required by the format view. Given that, one may wonder whether this kind of metacognitive control is a feasible counterexample against the commitment- based approach. After all, the young rich communist case exhibits the features of metacognitive control the commitment-based approach casts into question as necessary for the display of the metacognitive function of inner speech. Now, it must be clarified that the commitment-based approach is compatible with the fact that we can display mental concepts (belief, desire, fear) in our reasoning or inner speech episodes. In fact, we often self-attribute mental states (avowals) putting those mental concepts into work. However, this does not mean such self-ascriptions endow us with a particular mental access to our own psychological states.

In fact, when we pay closer attention to the social role of self-ascriptions, we realize that in conversational contexts we often use the first-personal ascription with pragmatic purposes.[27] For instance, the phrase 'I think' is frequently presented

[27] James O. Urmson, "Parenthetical verbs,"*Mind* 6, 244 (1952): 480–496; Karin Aijmer, "I think: an English modal particle," in *Modality in Germanic Language: Historical and Comparative Perspectives*, eds. Toril Swan and Olaf Westik (De Gruyter Mouton, 1997); Anna Wierzbicka, *English: Meaning and culture* (Oxford: Oxford University Press, 2006); Mandy

as having the function to mitigate the degree of commitment to the sentence it ranges. Wierzbicka provides a deep analysis of parenthetical uses of 'believe,' 'think' and other mental verbs. She claims that the verb 'think' conveys the meaning of disclaiming knowledge "not by saying "I don't know" but by saying "I don't say: I know it."[28] In other words, 'I think P' expresses a certain degree of caution. Similarly, the verb 'believe' (in contrast to 'I think' for instance) seems to play an indicative function. As Aijmer claims: "I believe does not only express a subjective attitude. It also conveys that the speaker has some evidence for what he says."[29] We can see the contrast between 'I think' and 'I believe' in the incompatibility of 'I believe' with phrases like 'I'm not sure.' While 'I think that Riga is the capital of Latvia, but I'm not sure' is idiomatic, 'I believe that Riga is the capital of Latvia but I'm not sure' is not. This difference between the level of reliability that 'think' and 'believe' convey must not divert our attention away from the fact they share their basic function: they are devices for canceling or altering the speaker's commitments. The verbs 'believe' and 'think' seem to be mitigators of the force of the claim. Of course, parenthetical uses are not restricted to these types of indications involving mitigations. Verbs as 'rejoice' or 'regret' indicate emotional orientation, others as 'wish' or 'desire' indicate the preference toward the commitments of the statement. What these parenthetical uses of propositional attitude verbs share is its function for providing indications or prescriptions to the hearer about how to evaluate the commitments of the proposition associated with the mental verb. As a conclusion, mental verbs in self-ascriptions seem to have the pragmatic function of signaling certain attitudes or indications toward the commitments expressed by the statement under the scope of the mental verb.

Taking this inside on board, when the young rich communist is evaluating what to do in the light of his future belief 'I will believe social justice does not matter,' he is considering the commitments he will give rise in the future given the content of his future belief. Furthermore, he assesses the type of actions he must carry out in the present in order to avoid his future commitments with the assertion that social justice does not matter. In this sense, we can recruit the same kind of commitment-based explanation without bringing out any type of access-like explanation. Although this kind of explanation seems to necessitate certain notion of metarepresentation that allows the young rich communist to perceive

Simons, "Observations on embedding verbs, evidentiality, and presupposition," *Lingua 117*, 6 (2007), 1034-1056.

[28] Wierzbicka, *English*, 38

[29] Aijmer, "I think," 17

himself as a minded creature, it does not commit us with understanding self-ascriptions as descriptions of inner processes or psychological states, rather than expressions that make explicit the commitments with the present and future actions associated with the content of the proposition under the scope of the mental verb. Thus, the commitment-based approach could also give a plausible explanation of the metacognitive capacities the future directed self-control requires.

A second objection against the commitment-based approach may cast into question its plausibility as a theory of the social function of speech acts. One may argue, for instance, that a neo-Gricean model of communicationprovides a better understanding of communication, and subsequently, for the cognitive function of inner speech.[30] In the neo-Gricean model, a hearer expects certain patterns of actions from a speaker because her speech acts express certain mental states that are causally connected with the given action. For instance, when a speaker asserts P, the hearer can infer through different pragmatic mechanisms that he is expressing a belief that P, and thus, the hearer can expect from the speaker a range of patterns of actions causally connected with such belief. The neo-Gricean approaches to communication exhibit certain problems whose consideration is beyond the purpose of this paper.[31] However, for the purpose of this article, it is sufficient to notice that such position requires our speech act to voice certain underlying mental states, which again brings out the problem of confabulation. Considering that inner speech requires putting to work pragmatic mechanisms that infer the mental states of the agent implies that the agent must be in a particular mental state that is causally connected to the private episode. However, as the empirical evidence considered in section 3 emphasizes, it is problematic to assume that our reasons, and thus our inner speech episodes always reflect an underlying mental state.

On the contrary, this is not problematic for the commitment-based approach. As Strijbos& de Bruin argue, our confabulatory reasons can have two

[30] For two well-known neo-Gricean Models of communication see Kent Bach and Robert Harnish, *Linguistic Communication and Speech Acts* (Cambridge, Mass.: MIT Press, 1979); and Dan Sperber and Deidre Wilson, *Relevance: Communication and Cognition*, (Oxford: Blackwell, 1986).

[31] For instance, these approaches are usually committed with the idea that communication requires the instantiation of mindreading mechanisms that, as Tadeusz Zawidzki has emphasized, make mental state attribution computationally intractable (see Tadeusz Zawidzki "The function of folk psychology: Mindreading or mindshaping?" *Philosophical Explorations* 11, 3 (2008): 193-210).

purposes.[32] Firstly, they can help to give coherence to our previous actions by providing us with a narrative. Secondly, they can have a prospective function, generating commitments that we are inclined to conform, and thus, that regulate our behavior and cognitive mechanisms. In this sense, the commitment-based approach can help us to elucidate the regulatory function of inner speech while avoiding the problem of confabulation. That is, our inner speech episodes do not necessarily reflect our underlying mental states, rather than it help us to give coherence and regulate our actions by giving rise to the commitments with certain patterns of actions.

## 7. Conclusion

The aim of this paper was to present several concerns regarding the format view of the metacognitive capacities of inner speech and to advocate an alternative. The problems associated with the format view rely on the role that the model assigns to metarepresentations and the notion of access. The solution I have offered respects our intuitions concerning inner speech episodes and accounts for the metacognitive capacities of regulating and evaluation our cognition and action. This position offers an alternative that does not require postulating metarepresentations or considering thinking about thinking in terms of access. Furthermore, the theory can avoid two possible objections. On the one hand, it can account for the cases where our metacognitive capacities require self-ascriptions. On the other hand, the theory can avoid certain challenges that other views of communication that have enjoyed a greater popularity cannot avoid.[33,34]

[32] Derek Strijbos and Leon de Bruin, “Self-interpretation as first-person mindshaping: implications for confabulation research,” *Ethical Theory and Moral Practice* 18, 2 (2005): 297-30.

[33] This article was written thanks to the funding provided by the project “Inner speech, Metacognition, and the Narrative View of Identity” (FFI2015-65953-P)and “Contemporary Expressivism and the Indispensability of Normative Vocabulary” (FFI2016-80088-P) funded by Ministerio de Economía y Competitividad; and the postdoctoral research contract “Puente,” funded by the University of Granada.

[34] Thanks to all members of the Department of Philosophy at University of Granada (AKA Granada Gang) for helpful comments on previous versions of these paper. I would like to also to acknowledge the influence and support of Fernando Martínez-Manrique and Agustin Vicente.

# ACCURACY AND THE IMPS

James M. JOYCE, Brian WEATHERSON

ABSTRACT: Recently several authors have argued that accuracy-first epistemology ends up licensing problematic epistemic bribes. They charge that it is better, given the accuracy-first approach, to deliberately form one false belief if this will lead to forming many other true beliefs. We argue that this is not a consequence of the accuracy-first view. If one forms one false belief and a number of other true beliefs, then one is committed to many other false propositions, e.g., the conjunction of that false belief with any of the true beliefs. Once we properly account for all the falsehoods that are adopted by the person who takes the bribe, it turns out that the bribe does not increase accuracy.

KEYWORDS: accuracy, epistemic consequentialism, scoring rules

## 1.Accuracy, Bribes and Scoring Rules

Belief aims at the truth.[1] So at least in some sense, an agent is doing better at believing the closer they are to the truth. When applied to individual beliefs, this generates epistemic advice that is literally platitudinous: if you know that a change in your attitude towards *p* will make your attitude towards *p* more accurate, make that change! When applied to collective bodies of belief though, the advice turns out to be more contentious. Call **epistemic consequentialism** the view that if an agent knows that a change in their overall belief state will make their belief state more accurate, they should make that change, if they have the power to do so.

Hilary Greaves has recently argued that epistemic consequentialism is false because it licences certain epistemic 'bribes', and these should not be licenced.[2] We'll argue that the best forms of epistemic consequentialism do not licence some of these bribes after all.[3] Here is the key case Greaves uses.[4]

[1] Thanks to Alejandro Pérez Carballo, Richard Pettigrew, and the participants in the Arché Epistemology Seminar for helpful comments.

[2] Hilary Greaves, "Epistemic Decision Theory," *Mind* 122 (2013): 915–952, https://doi:10.1093/mind/fzt090.

[3] Though they do licence others; see section 2.4 for more discussion.

[4] Greaves has four other cases, but the Imps case is the only one that is a problem for all forms of consequentialism she discusses. Similar cases have suggested by Selim Berker and C. S. Jenkins, but we'll focus on Greaves's discussion since she engages more fully with the literature on scoring rules. We'll return briefly to Berker's discussion in section 2. Berker's version is in his "Epistemic Teleology and the Separateness of Propositions," *Philosophical Review* 122 (2013):

> Emily is taking a walk through the Garden of Epistemic Imps. A child plays on the grass in front of her. In a nearby summerhouse are $n$ further children, each of whom may or may not come out to play in a minute. They are able to read Emily's mind, and their algorithm for deciding whether to play outdoors is as follows. If she forms degree of belief 0 that there is now a child before her, they will come out to play. If she forms degree of belief 1 that there is a child before her, they will roll a fair die, and come out to play iff the outcome is an even number. More generally, the summerhouse children will play with chance $(1 - \frac{q(C_0)}{2})$, where $q(C_0)$ is the degree of belief Emily adopts in the proposition $C_0$ that there is now a child before her. Emily's epistemic decision is the choice of credences in the proposition $C_0$ that there is now a child before her, and, for each $j = 1, \ldots, n$ the proposition $C_j$ that the *j*th summerhouse child will be outdoors in a few minutes' time.
>
> ...if Emily can just persuade herself to ignore her evidence for $C_0$, and adopt (at the other extreme) credence 0 in $C_0$, then, by adopting degree of belief 1 in each $C_j (j = 1, \ldots, 10)$, she can guarantee a perfect match to the remaining truths. Is it epistemically rational to accept this 'epistemic bribe'?[5]

The epistemic consequentialist says that it is best to have credences that are as accurate as possible. We will focus on believers who assign probabilistically coherent credences (degrees of belief) to the propositions in some "target set" $\mathcal{X}$, and we will think of the "degree of fit" between her beliefs and the truth as being measured by a strictly proper scoring rule. This is a function $\mathbf{I}_{\mathcal{X}}$ which associates each pair $\langle \mathbf{cred}, @ \rangle$ consisting of a credence function **cred** whose domain includes $\mathcal{X}$ and a consistent truth-value assignment @ for elements of $\mathcal{X}$ with a non-negative real number $\mathbf{I}_{\mathcal{X}}(@, \mathbf{cred})$. Intuitively, $\mathbf{I}_{\mathcal{X}}$ measures the inaccuracy of the credences that cred assigns to the propositions in $\mathcal{X}$ when their truth-values are as described by @. Note that higher $\mathbf{I}_{\mathcal{X}}$-values indicate higher levels of epistemic disutility, so that lower is better from a consequentialist perspective. One popular scoring rule is the Brier score, which identifies inaccuracy with the average squared distance between credences and truth-values. (Greaves calls this the 'quadratic scoring rule', which is a useful description too.) More formally, we have:

$$\mathbf{Brier}_{\mathcal{X}}(@, \mathbf{cred}) = \frac{1}{|\mathcal{X}|} \sum_{X \in \mathcal{X}} (\mathbf{cred}(X) - @(X))^2$$

337–393, http://doi.org/10.1215/00318108-2087645, and "The Rejection of Epistemic Consequentialism," *Philosophical Issues* 23 (2013): 363–387. Jenkins's version is in her "Entitlement and Rationality," *Synthese* 157 (2007): 25–45, http://doi.org/10.1007/s11229-006-0012-2.

[5] Greaves, "Epistemic Decision Theory," 918.

where$|\mathcal{X}|$ is the number of propositions in $\mathcal{X}$ and $@(X)$ is either zero or one depending upon whether X is true or false.

Another common score is the logarithmic rule, which defines inaccuracy as:

$$\mathbf{Log}_{\mathcal{X}}(@, \mathbf{cred}) = \frac{1}{|\mathcal{X}|}\sum_{X \in \mathcal{X}} -\log(\mathbf{cred}(X)) \cdot @(X)$$

For now we will follow Greaves in assuming that our epistemic consequentialist uses the Brier score to measure epistemic disutility, but we will relax that assumption in a little while.

Now let's think about the 'bribe' that Greaves offers, from the point of view of the epistemic consequentialist. The choices are to have one of two credal states, which we'll call **cred1** and **cred2**. We'll say **cred1** is the one that best tracks the initial evidence, so $\mathbf{cred1}(C_0) = 1$, and $\mathbf{cred1}(C_i) = 0.5$ for $i \in 1,\ldots,10$. And **cred2** is the credence Emily adopts if she accepts the bribe, so $\mathbf{cred2}(C_0) = 0$, while $\mathbf{cred2}(C_i) = 1$ for $i \in 1,\ldots,10$. Which state is better?

Thinking like an epistemic consequentialist, you might ask which state is more accurate? It seems like that would be **cred2**. While **cred1** gets $C_0$ exactly right it does not do very well on the other propositions. In contrast, while **cred2** gets $C_0$ exactly wrong, it is perfect on the other ten propositions. So overall, **cred2** looks to have better epistemic consequences: when compared to being right about one proposition and off by 0.5 on ten others, being right on ten is surely worth one false belief. The Brier score seems to bear this out. If we let $\mathcal{X}$, the target set, consist of $C_0, C_1, \ldots, C_{10}$, then we have

$$\begin{aligned}
\mathbf{Brier}_{\mathcal{X}}(\mathbf{cred1}, @) &= \frac{1}{11}[(1 - \mathbf{cred1}(C_0))^2 + \sum_{i=1}^{10}(\,@(C_i) - \frac{1}{2})^2] = \frac{10}{44} \\
\mathbf{Brier}_{\mathcal{X}}(\mathbf{cred2}, @) &= \frac{1}{11}[(1 - \mathbf{cred2}(C_0))^2 + \sum_{i=1}^{10}(\,@(C_i) - cred(C_i))^2] = \frac{1}{11}
\end{aligned}$$

So, it seems that a good epistemic consequentialist will take the bribe. But, doesn't that seem like the height of epistemic irresponsibility? It means choosing to believe that $C_0$ is certainly false when you have conclusive evidence for thinking that it is true. If you see the child on the lawn in front of you, how can you sanction believing she is not there?

As Greaves admits, intuitions are divided here. Some consequentialists might think that "epistemic bribes" are at least sometimes worth taking, while those of a more deontological bent will always find such trade-offs "beyond the pale."[6] We

[6] Berker, "Epistemic Teleology," 363.

will largely sidestep these contentious issues here, though our argument will offer comfort to epistemic consequentialists who feel queasy about accepting the bribe offered in Imps. We contend that, when inaccuracy is measured properly, the consequences of adopting the **cred2** credences are strictly worse than the consequences of adopting **cred1**.

The basic problem is that Imps cherry-picks propositions in a way no consequentialist should condone. Its persuasive force rests on the assumption that, for purposes of epistemic evaluation, nothing matters except the accuracies of the credences assigned to propositions in the target set $\mathcal{X}$. But $\mathcal{X}$ is the wrong target! By confining attention to it Greaves ignores the many other credences to which Emily becomes committed as a consequence of adopting **cred1** or **cred2**. Any (coherent) agent who invests credence zero in $C_0$ must also invest credence zero in any proposition $C_0 \wedge Y$, where $Y$ is any conjunction or disjunction of elements from $\mathcal{X}$. Likewise, anyone who invests credence one in $C_n$ must invest credence one in any proposition $C_n \vee Y$, where $Y$ is any conjunction or disjunction from $\mathcal{X}$. In the current context (where the probabilities of the various $C_i$ are independent), when Emily adopts a credence function over $\mathcal{X}$ she commits to having a credence for (i) every atomic proposition $\pm C_0 \wedge \pm\ C_1 \wedge \pm C_2 \wedge \ldots \wedge \pm C_{10}$, where '±' can be either an affirmation or a negation, and (ii) every disjunction of these atomic propositions. In short, she commits to having credences over the whole Boolean algebra $\mathcal{A}_{\mathcal{X}}$ generated by $\mathcal{X}$. Since each event of a child coming out is independent, adopting **cred1** will commit her to setting **cred1**$(\pm C_0 \wedge \pm\ C_1 \wedge \pm C_2 \wedge \ldots \wedge \pm C_{10}) = \frac{1}{1024}$ when $C_0$ is affirmed, and 0 when it is negated. While adopting **cred2** commits her to setting **cred2**$(\pm C_0 \wedge \pm\ C_1 \wedge \pm C_2 \wedge \ldots \wedge \pm C_{10})$ equal to 1 when $C_0$ is negated and the rest of the $C_i$ are affirmed, and equal to 0 otherwise. In this way, each of these probability assignments over the 2048 atoms determine a definite probability for every one of the $2^{2048}$ propositions in $\mathcal{A}_{\mathcal{X}}$.

It is our view that consequentialists should reject any assessment of epistemic utility that fails to take the accuracies of *all* these credences into account. All are consequences of adopting **cred1** or **cred2**, and so all should be part of any consequentialist evaluation of the quality of those credal states. The right "target set" to use when computing epistemic disutility is not $\mathcal{X}$but $\mathcal{A}_{\mathcal{X}}$. If we don't do that, we ignore most of the ways in which **cred1** and **cred2** differ in accuracy. If Emily takes the bribe, she goes from having credence 0.5 in $C_0 \leftrightarrow C_1$ to having credence 0 in it. And that's unfortunate, because the chance of $C_0 \leftrightarrow C_1$ goes from 0.5 to 1. This is another proposition, as well as $C_0$, that Emily acquires a false belief in by taking the bribe. Of course, there are other propositions not counted that go the other way. Originally, Emily has a credence of 0.25 in $C_1 \wedge C_2$, and its chance is also 0.25. After taking the bribe, this has a chance of 1, and her credence in it is 1. That's an improvement in accuracy. So there are a host of both improvements and deteriorations that are as yet unaccounted for. We should account for them, and making the target set be $\mathcal{A}_{\mathcal{X}}$ does that.

When seen from this broader perspective, it turns out the seeming superiority of **cred2** over **cred1** evaporates. The rest of this section (and the appendix) is dedicated to demonstrating this. We'll make the calculations a little easier on ourselves by relying on a theorem concerning Brier scores for coherent agents. Assume, as is the case here, that Emily's credences are defined over an atomic Boolean alegbra of propositions. The atoms are the 'worlds', or states that are maximially specific with respect to the puzzle at hand. In this case there are 2048 states, which we'll label $s_0$through $s_{2047}$. In $s_k$, the first child is on the lawn iff $k \leq 1023$, and summerhouse child $i$ comes out iff the $(i + 1)$th digit in the binary expansion of $k$ is 1. Let $\mathcal{S}_{\mathcal{X}}$ be the set of all these states. That's not a terrible target set; as long as Emily is probabilistically coherent it is comprehensive. The theorem in question says that for any credence function **cred** defined over a partition of states $\mathcal{S}$, and over the algebra $\mathcal{A}$ generated by those states,

**Theorem-1**

$$\mathbf{Brier}_{\mathcal{A}}(\mathbf{cred}, @) = \frac{|\mathcal{S}|}{4}\mathbf{Brier}_{\mathcal{S}}(\mathbf{cred}, @)$$

(The proof of this is in the appendix.) So whichever credence function is more accurate with respect to $\mathcal{S}_{\mathcal{X}}$ will be more accurate with respect to $\mathcal{A}_{\mathcal{X}}$. So let's just work out $\mathbf{Brier}_{\mathcal{S}_{\mathcal{X}}}$ for **cred1** and **cred2** at the actual world.

First, **cred1** will appropriately assign credence 0 to each $s_k$ ($k \in 0, \ldots, 1023$). Then it assigns credence $\frac{1}{1024}$ to every other $s_k$. For 1023 of these, that is off by $\frac{1}{1024}$, contributing $\frac{1}{2^{20}}$ to the Brier score. And for 1 of them, namely @, it is off by $\frac{1023}{1024}$, contributing $\frac{1023^2}{2^{20}}$. So we get:

$$\begin{aligned}
\mathbf{Brier}_{\mathcal{S}_{\mathcal{X}}}(\mathbf{cred1}, @) &= \frac{1}{2048}[1024 \cdot 0 + 1023 \cdot \frac{1}{2^{20}} + \frac{1023^2}{2^{20}}] \\
&= \frac{1}{2048} \cdot \frac{1023 + 1023^2}{2^{20}} \\
&= \frac{1}{2048} \cdot \frac{1023 \cdot 1024}{2^{20}} \\
&= \frac{1}{2048} \cdot \frac{1023}{1024} \\
&= \frac{2^{10} - 1}{2^{21}}
\end{aligned}$$

It's a bit easier to work out $\mathbf{Brier}_{\mathcal{S}_\mathcal{X}}(\mathbf{cred2}, s_{2047})$. (We only need to work out the Brier score for that state, because by the setup of the problem, Emily knows that's the state she'll be in if she adopts **cred2**). There are 2048 elements in $\mathcal{S}_\mathcal{X}$. And **cred2** assigns the perfectly accurate credence to 2046 of them, and is perfectly inaccurate on 2, namely $s_{1023}$, which it assigns credence 1, and $s_{2047}$ which it assigns credence 0. So we have

$$\begin{aligned}\mathbf{Brier}_{\mathcal{S}_\mathcal{X}}(\mathbf{cred2}, s_{2047}) &= \frac{1}{2048}(2046 \cdot 0 + 1 + 1) \\ &= \frac{1}{1024} \\ &= \frac{2^{11}}{2^{21}}\end{aligned}$$

In fact, it isn't even close. If Emily adopts **cred2** she becomes a little more than significantly more inaccurate.

It is tedious to calculate $\mathbf{Brier}_{\mathcal{A}_\mathcal{X}}(\mathbf{cred1}, @)$ directly, but it is enlightening to work through the calculation of $\mathbf{Brier}_{\mathcal{A}_\mathcal{X}}(\mathbf{cred2}, s_{2047})$. Note that there are two crucial states out of the 2048: $s_{2047}$, the actual state where all children come out, and state $s_{1023}$ where child 0 does not come out, but the other 10 children all do. There are $2^{2^{11}-2}$ propositions in each of the following four sets:

1. $\{p: s_{2047} \vDash p \text{ and } s_{1023} \vDash p\}$
2. $\{p: s_{2047} \vDash p \text{ and } s_{1023} \nvDash p\}$
3. $\{p: s_{2047} \nvDash p \text{ and } s_{1023} \vDash p\}$
4. $\{p: s_{2047} \nvDash p \text{ and } s_{1023} \nvDash p\}$

If Emily takes the bribe, she will have perfect accuracy with respect to all the propositions in class 1 (which are correctly believed to be true), and all the propositions in class 4 (which are correctly believed to be false). But she will be perfectly inaccurate with respect to all the propositions in class 2 (which are incorrectly believed to be false), and all the propositions in class 3 (which are incorrectly believed to be true). So she is perfectly accurate on half the propositions, and perfectly inaccurate on half of them, so her average inaccuracy is $0.5 \cdot 0 + 0.5 \cdot 1 = 0.5$. And that's an enormous inaccuracy. It is, in fact, as inaccurate as one can possibly be while maintaining probabilistic coherence.

> **Theorem-2**: When inaccuracy over $\mathcal{A}$ is measured using the Brier score, the least accurate credal states are those which assign credence 1 to some false atom of $\mathcal{A}$.

(The proof is in the appendix.) So taking the bribe is not a good deal, even by consequentialist lights. And that isn't too surprising; taking the bribe makes Emily have maximally inaccurate credences on half of the possible propositions about the children.

So far we have followed Greaves in assuming that inaccuracy is measured by the quadratic, or Brier, rule. It turns out that we can drop that assumption. We actually only need some very weak conditions on accuracy rules to get the result that Greaves style bribes are bad deals, though the proof of this becomes a trifle more complicated.

Let $\mathcal{A}$ be an algebra of propositions generated by a partition of $2N$atoms $a_1,\ldots,a_{2N}$. Suppose $a_1$ is the truth, and consider two probability functions, $P$ and $Q$ defined in $\mathcal{A}$. $P$assigns all its mass to the first $N$ atoms, so that $P(a_k) = 0$ for all $k > N$. We also assume that $P$ assigns some positive probability to the true atom $a_1$. $Q$assigns all its mass to the false atom $a_{2N}$. Note that this will be a good model of any case where an agent is offered a bribe of the form: drop the positive confidence you have in proposition $p_0$, instead assign it credence 0, and you'll be guaranteed a maximally accurate credence in $j$ other logically independent propositions $p_1,\ldots,p_j$. The only other assumptions needed to get the model to work are that $p_0$ is actually true, and $N = 2^j$.

Imagine that the accuracy of a probability function $\pi$ over $\mathcal{A}$ is measured by a proper scoring rule of the form

$$\mathbf{I}(a_n,\pi) = 2^{-2N} \sum_{X\in\mathcal{A}} \mathbf{i}\,(v_n(X),\pi(X))$$

where $v_n(X)$ is $X$s truth value when $a_n$ is the true atom, and $\mathbf{i}$ is a score that gives the accuracy of $\pi(X)$ in the event that $X$s truth value is $v_n(X)$. We shall assume that this score has the following properties.

**Truth Directedness**

The value of $\mathbf{i}(1,p)$ decreases monotonically as $p$ increases. The value of $\mathbf{i}(0,p)$ increases monotonically as $p$ decreases.

**Extensionality**

$\mathbf{i}(v_n(X),\pi(X))$ is a function only of the truth-value and the probability; the identity of the proposition does not matter.

**Negation Symmetry**

$\mathbf{i}(v_n(\neg X),\pi(\neg X)) = \mathbf{i}(v_n(X),\pi(X))$ for all $x,n,\pi$.

**Theorem-3**: Given these assumptions, $P$'s accuracy strictly exceeds $Q$'s.

Again, the proof is in the appendix.

Theorem-3 ensures that taking the deal that Greaves offers in Imps will reduce Emily's accuracy relative to any proper scoring rule satisfying Truth Directedness, Extensionality and Negation Symmetry. To see why, think of Emily's credences as being defined over an algebra generated by the atoms $\pm C_0 \wedge \pm C_1 \wedge \pm C_2 \wedge \ldots \wedge \pm C_{10}$, where it is understood that some $C_0$ atom is true and all the $\neg C_0$ atoms are false. Since Emily is convinced of $C_0$ and believes that every other $C_n$ has some chance of occurring, and since the various $C_n$ are independent of one another, her credence function **cred1** will assigns a positive probability to each $C_0$ atom, including the true atom (whichever that might be). Now, let $Q$ be a credence function that places all its weight on some false atom $\neg C_0 \wedge \pm C_1 \wedge \pm C_2 \wedge \ldots \wedge \pm C_{10}$. Theorem-3 tells us that Emily's **cred1** is more accurate than $Q$, and that this is true no matter which $C_0$ atom is true or which $\neg C_0$ atom $Q$ regards as certain. By taking the bribe Emily will guarantee the truth of $C_0 \wedge C_1 \wedge \ldots \wedge C_{10}$, but the cost will be that she must adopt the **cred2** credences, which assign probability one to the false atom $\neg C_0 \wedge C_1 \wedge \ldots \wedge C_{10}$. Extensionality ensures that any two credence functions that assign probability one to a false atom will have the same inaccuracy score, and that this score will not depend on which atom happens to be the true one. The upshot is that **cred2** will have the same inaccuracy when Emily accepts the bribe as $Q$ does when she rejects it. Thus, since **cred1** is more accurate than $Q$, it is also more accurate than **cred2**, which means that Emily should reject the bribe in order to promote credal accuracy.

We do not want to oversell this conclusion. Strictly speaking, we have only shown that consequentialists should reject epistemic bribes when doing so requires them to go from being confident in a truth to being certain of some maximally specific falsehood. This is a rather special situation, and there are nearby cases to which our results do not apply, and in which consequentialists may sanction bribe-taking. For example, if Emily only has to cut her credence for $C_0$ in half, say from $\frac{1}{2}$to $\frac{1}{4}$, to secure knowledge of $C_1 \wedge \ldots \wedge C_{10}$, then Theorem-3 offers us no useful advice. Indeed, depending on the scoring rule and the nature of the bribe, we suspect that believers will often be able to improve accuracy by changing their credences in ways not supported by their evidence, especially when these changes affect the truth-values of believed propositions. The only thing we insist upon is that, in all such cases, credal accuracy should be measured over all relevant propositions, not just over a select salient few. But that's something that is independently plausible. Perhaps it might be pragmatically justified to become more accurate on salient propositions at the expense of becoming very inaccurate over hard to state compounds of those propositions, but it is never epistemically justified.

## 2. Four Caveats

### 2.1 Greaves's Imps Argument May Work Against Some Forms of Consequentialism

We said above that no consequentialist should accept Greaves's setup of the Imps puzzle, since they should not accept an inaccuracy measure that ignores some kind of introduced inaccuracy. That means that, for all we have said, Greaves's argument works against those consequentialists who do not agree with us over the suitability of target sets that are neither algebras or partitions. And, at least outside philosophy, some theorists do seem to disagree with us.

For instance, it is common in meteorology to find theorists who measure the accuracy of rain forecasts over an $n$ day period by just looking at the square of the distance between the probability of rain and the truth about rain on each day. To pick an example almost literally at random, Mark Roulston defends the use of the Brier score, calculated just this way, as a measure of forecast accuracy.[7] So Greaves's target, while not including all consequentialists, does include many real theorists.

That said, it seems there are more mundane reasons to not like this approach to measuring the accuracy of weather forecasts. Consider this simple case. Ankita and Bojan are issuing forecasts for the week that include probabilities of rain. They each think that there is a 0% chance of rain most days. But Ankita thinks there will be one short storm come through during the week, while Bojan issues a 0% chance of rain forecast for each day. Ankita thinks the storm is 75% likely to come on Wednesday, so there's a 75% chance of rain that day, and 25% likely to come Thursday, so there's a 25% chance of rain that day.

As it happens, the storm comes on Thursday. So over the course of the week, Bojan's forecast is more accurate than Ankita's. Bojan is perfectly accurate on 6 days, and off by 1 on Thursday. Ankita is perfectly accurate on 5 days, and gets an inaccuracy score of $0.75^2 = 0.5625$ on Wednesday and Thursday, which adds up to more than Bojan's inaccuracy. But this feels wrong. There is a crucial question that Ankita was right about and Bojan was wrong about, namely will there be a storm in the middle of the week. Ankita's forecast only looks less accurate because we aren't measuring accuracy with respect to this question. So even when we aren't concerned with magical cases like Greaves's, there is a good reason to measure accuracy comprehensively, i.e., with respect to an algebra or a partition.

[7] Mark S. Roulston, "Performance Targets and the Brier Score," *Meterological Applications* 14 (2007): 185–194, http://doi.org/10.1002/met.21.

## 2.2 Separateness of Propositions

There is a stronger version of the intuition behind the Imps case that we simply reject. The intuition is well expressed by Selim Berker.

> The more general point is this: when determining the epistemic status of a belief in a given proposition, it is epistemically irrelevant whether or not that belief conduces (either directly or indirectly) toward the promotion of true belief and the avoidance of false belief in *other* propositions beyond the one in question.[8]

Let's put that to the test by developing the Ankita and Bojan story a little further. They have decided to include, in the next week's forecast, a judgment on the credibility of rain. Bojan thinks the evidence is rather patchy. And he has been reading Glenn Shafer, and thinks that when the evidence is patchy, credences in propositions and their negations need not add to one.[9] So if $p$ is the proposition *It will rain next week*, Bojan has a credence of 0.4 in both $p$and $\neg p$.

Ankita thinks that's crazy, and suggests that there must be something deeply wrong with the Shafer-based theory that Bojan is using. But Bojan is able to easily show that the common arguments against Shafer's theory are blatantly question begging.[10] So Ankita tries a new tack. She has been reading Joyce, from which she got the following idea.[11] She argues that Bojan will be better off from the point of view of accuracy in having credence 0.5 in each of $p$ and $\neg p$ than in having credence 0.4 in each. As it stands, one of Bojan's credences will be off by 0.4, and the other by 0.6, for a Brier score of $(0.4^2 + 0.6^2)/2 = 0.26$, whereas switching would give him a Brier score of $(0.5^2 + 0.5^2)/2 = 0.25$.

But Bojan resists. He offers two arguments in reply.

First, he says, for all Ankita knows, one of his credences might be best responsive to the evidence. And it is wrong, always and everywhere, to change a credence away from one that is best supported by the evidence in order to facilitate an improvement in global accuracy. That, says Bojan, is a violation of the "separateness of propositions".[12]

[8] Berker, "Epistemic Teleology," 365, emphasis in original.

[9] Glenn Shafer, *A Mathematical Theory of Evidence* (Princeton: Princeton University Press, 2016).

[10] Patrick Maher, "Depragmatised Dutch Book Arguments," *Philosophy of Science* 64 (1997): 291–305, http://doi.org/10.1086/392552; Brian Weatherson, "Begging the Question and Bayesians," *Studies in the History and Philosophy of Science* Part A 30 (1999): 687–697.

[11] James M. Joyce, "A Non-Pragmatic Vindication of Probabilism," *Philosophy of Science* 65 (1998): 575–603.

[12] Berker, "Epistemic Teleology."

Second, he says, even by Ankita's accuracy-based lights, this is a bad idea. After all, he will be making one of his credences less accurate in order to make an improvement in global accuracy. And that's again a violation of the separateness of propositions. It's true that he won't be making himself more inaccurate in one respect so as to secure accuracy in another, as in the bribes case. But he will be following advice that is motivated by the aim of becoming, in total, more accurate, at the expense of accuracy for some beliefs.

We want to make two points in response. First, if the general point that Berker offers is correct, then these are perfectly sound replies by Bojan. Although Bojan is not literally in a bribe case, like Emily, he is being advised to change some credences because the change will make his overall credal state better, even if it makes it locally worse in one place. It does not seem to matter whether he can identify which credence gets made worse. Berker argues that the trade-offs that epistemic consequentialism makes the same mistake ethical consequentialism makes; it authorises inappropriate trade-offs. But in the ethical case, it doesn't matter whether the agent can identify who is harmed by the trade-off. If it is wrong to harm an identifiable person for the greater good, it is wrong to harm whoever satisfies some description in order to produce the greater good.

So if the analogy with anti-consequentialism in ethics goes through, Bojan is justified in rejecting Ankita's advice. After all there is, according to Berker, a rule against making oneself doxastically worse in one spot for the gain of an overall improvement. And that's what Bojan would do if he took Ankita's advice. But, we say, Bojan is not justified in rejecting Ankita's advice. In fact, Ankita's advice is sound advice, and Bojan would do well to take it. So Berker's general point is wrong.

Our second point is a little more contentious. We suspect that if Bojan has a good reason to resist this move of Ankita's, he has good reason to resist all attacks on his Shafer-based position. So if Berker's general point is right, it means there is nothing wrong with Bojan's anti-probabilist position. Now we haven't argued for this; to do so would require going through all the arguments for probabilism and seeing whether they can be made consistent with Berker's general point. But our suspicion is that none of them can be, since they are all arguments that turn on undesirable properties of global features of non-probabilistic credal states. So if Berker is right, probabilism is wrong, and we think it is not wrong.

### 2.3 Is this Consequentialism?

So far we've acquiesed with the general idea that Greaves's and Berker's target should be called *consequentialism*. But there are reasons to be unhappy with this label. In general, a consequentialist theory allows agents to make things worse in the here and now, in return for future gains. A consequentialist about prudential decision making, in the sense of Hammond, will recommend exercise and medicine taking.[13] And they won't be moved by the fact that the exercise hurts and the medicine is foul-tasting. It is worth sacrificing the welfare of the present self for the greater welfare of later selves.

Nothing like that is endorsed, as far as we can tell, by any of the existing 'epistemic consequentialists'. Certainly the argument that Ankita offers Bojan does not rely on this kind of reasoning. In particular, epistemic consequentialists do not say that it is better to make oneself doxastically worse off now in exchange for greater goods later. Something like that deal is offered to the reader of Descartes's *Meditations*, but it isn't as popular nowadays.

Rather, the rule that is endorsed is *Right now, have the credences that best track the truth!* This isn't clearly a form of consequentialism, since it really doesn't care about the *consequences* of one's beliefs. It does say that it is fine to make parts of one's doxastic state worse in order to make the whole better. That's what would happen if Bojan accepted Ankita's advice. But that's very different from doing painful exercise, or drinking unpleasant medicine. (Or, for that matter, to withdrawing belief in any number of truths.)

When Greaves tries to flesh out epistemic consequentialism, she compares it to evidential and causal versions of prudential decision theory. But it seems like the right comparison might be to something we could call *constitutive* decision theory. The core rule, remember, is that agents should form credences that constitute being maximally accurate, not that cause them to be maximally accurate.

The key point here is not the terminological one about who should be called consequentialist. Rather, it is that the distinction between causation and constitution is very significant here, and comparing epistemic utility theory to prudential utility theory can easily cause it to be lost. Put another way, we have no interest in defending someone who wants to defend a causal version of epistemic utility theory, and hence thinks it could be epistemically rational to be deliberately

[13] Peter J. Hammond, "Consequentialist Foundations for Expected Utility," *Theory and Decision* 25 (1988): 25–78, http://doi.org/10.1007/BF00129168.

inaccurate now in order to be much more accurate tomorrow. We do want to defend the view that overall accuracy right now is a prime epistemic goal.[14]

## 2.4 Other Bribes

As already noted, we have not offered a general purpose response to bribery based objections to epistemic consequentialism. All we've shown is that some popular examples of this form of objection misfire, because they offer bribes that are bad by the consequentialists' own lights. But there could be bribes that are immune to our objection.

For example, imagine that Ankita has, right now, with credence 0.9 in $D_0$, and 0.5 in $D_1$. These are good credences to have, since she knows those are the chances of $D_0$and $D_1$. She's then offered an epistemic bribe. If she changes her credence in $D_0$ to 0.91, the chance of $D_1$ will become 1, and she can have credence 1 in $D_1$. Taking this bribe will increase her accuracy.

We could imagine the anti-consequentialist arguing as follows.

1. If epistemic consequentialism is true, Ankita is epistemically justified in accepting this bribe.
2. Ankita is not epistemically justified in accepting this bribe.
3. So, epistemic consequentialism is not true.

We're not going to offer a reply to this argument here; that is a task for a much longer paper. There are some reasons to resist premise one. It isn't clear that it is conceptually possible to accept the bribe. (It really isn't clear that it is practically possible, but we're not sure whether that's a good reply on the part of the consequentialist.) And it isn't clear that the argument for premise one properly respects the distinction between causation and constitution we described in the last section.

Even if those arguments fail, the intuitive force of premise two is not as strong as the intuition behind Greaves's, or Berker's, anti-bribery intuitions. And that's one of the main upshots of this paper. It's commonly thought that for the consequentialist, in any field, everything has its price. The result we proved at the end of section one shows this isn't true. It turns out that no good epistemic consequentialist should accept a bribe that leads them to believing an atomic proposition they have conclusive evidence is false, no matter how strong the

[14] For further discussion of epistemic consequentialism, see James M. Joyce, "Accuracy, Ratification, and the Scope of Epistemic Consequentialism," In *Epistemic Consequentialism*, eds. H. Kristoffer Ahlstrom-Vij and Jeffrey Dunn (Oxford: Oxford University Press, 2018), 240-266. https://doi.org/10.1093/oso/9780198779681.003.0011

inducements. Maybe one day there will be a convincing bribery based case that epistemic consequentialism is unacceptably corrupting of the epistemic soul. But that case hasn't been made yet, because we've shown a limit on how corrupt the consequentialist can be.

## Appendix: Proofs of Theorems 1, 2, 3

**Theorem-1**: $\text{Brier}_{\mathcal{A}}(\mathbf{c}, @) = \frac{N}{4}\text{Brier}_{\mathcal{S}}(\mathbf{c}, @)$ where

$$\text{Brier}_{\mathcal{S}}(\mathbf{c}, @) \quad = \frac{\sum_{s \in \mathcal{S}} (\, @(s) - c(s))^2}{N}$$

To prove this we rely on a series of lemmas.[15]

Let $\mathcal{A}$ be the algebra generated by a finite partition of states $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$. @ is a truth-value assignment for propositions in $\mathcal{A}$. For simplicity, assume $s_1$ is the true state, so that $@(s_1) = 1$ and $@(s_n) = 0$ for $n > 1$. The credence function $\mathbf{c}$ assigns values of $c_1, c_2, \dots, c_{N-1}, c_N$ to the elements of $\mathcal{S}$, where $\sum_{n=1}^{N} c_n = 1$ in virtue of coherence.

It will be convenient to start by partitioning $\mathcal{A}$ into four "quadrants". Let $B$ range over all disjunctions with disjunctions drawn from $\mathcal{B} = \{s_2, s_3, \dots, s_{N-1}\}$ (including the empty disjunction, i.e., the logical contradition $\perp$). Then, $\mathcal{A}$ can be split into four disjoint parts:

$\mathcal{A}_1 = \{B \vee s_1 \vee s_N : B$ is a disjunction of the elements of $\mathcal{B}\}$

$\mathcal{A}_2 = \{B \vee s_1 : B$ is a disjunction of the elements of $\mathcal{B}\}$

$\mathcal{A}_3 = \{B \vee s_N : B$ is a disjunction of the elements of $\mathcal{B}\}$

$\mathcal{A}_4 = \{B : B$ is a disjunction of the elements of $\mathcal{B}\}$

Notice that:

(i) $\mathcal{A}_1 \cup \mathcal{A}_2$contains all and only the true propositions in $\mathcal{A}$.

(ii) $\mathcal{A}_3 \cup \mathcal{A}_4$contains all and only the false propositions in $\mathcal{A}$.

(iii) $\mathcal{A}_1$ and $\mathcal{A}_4$ are *complementary* sets, i.e., all elements of $\mathcal{A}_4$ are negations of elements of $\mathcal{A}_1$, and conversely.

(iv) $\mathcal{A}_2$ and $\mathcal{A}_3$ are also complementary.

(v) $\mathcal{A}_1 \cup \mathcal{A}_4$ is the subalgebra of $\mathcal{A}$ generated by $\{s_1 \vee s_N, s_2, s_3, \dots, s_{N-1}\}$.

(vi) All four quadrants have the same cardinality of $2^{N-2}$.

[15] Alejandro Pérez Carballo gives a more direct and elegant proof of this result in a recent manuscript. We have kept our inefficient proof since its structure provides a guide for the proof of Theorem-3.

For an additive scoring rule $\mathbf{I}(\mathbf{c}, @) = \sum_{A \in \mathcal{A}} \mathbf{i}\,(\mathbf{c}(A), @(A))$ and $j = 1,2,3,4$, define $\mathbf{I}_j = \sum_{A \in \mathcal{A}_j} \mathbf{i}\,(\mathbf{c}(A), @(A))$, and note that $\mathbf{I}(\mathbf{c}, @) = 2^{-N}(\mathbf{I}_1 + \mathbf{I}_2 + \mathbf{I}_3 + \mathbf{I}_4)$.

**Lemma-1.1**: If **I** is negation symmetric, i.e., if $\mathbf{i}(\mathbf{c}(\neg A), @(\neg A)) = \mathbf{i}(\mathbf{c}(A), @(A))$ for all $A$, then $\mathbf{I}_1 = \mathbf{I}_4$ and $\mathbf{I}_2 = \mathbf{I}_3$, and $\mathbf{I}(\mathbf{c}, @) = 2^{1-N}(\mathbf{I}_2 + \mathbf{I}_4)$.

**Proof**: This is a direct consequence of the fact that $\mathcal{A}_1$ is complementary to $\mathcal{A}_4$ and that $\mathcal{A}_2$ is complementary to $\mathcal{A}_3$ since this allows us to write

$$\begin{aligned}
\mathbf{I}_1(\mathbf{c}, @) &= \sum_{A \in \mathcal{A}_1} \mathbf{i}\,(\mathbf{c}(A), @(A)) = \sum_{A \in \mathcal{A}_1} \mathbf{i}\,(\mathbf{c}(\neg A), @(\neg A)) = \mathbf{I}_4(\mathbf{c}, @). \\
\mathbf{I}_3(\mathbf{c}, @) &= \sum_{A \in \mathcal{A}_3} \mathbf{i}\,(\mathbf{c}(A), @(A)) = \sum_{A \in \mathcal{A}_3} \mathbf{i}\,(\mathbf{c}(\neg A), @(\neg A)) = \mathbf{I}_2(\mathbf{c}, @). \text{QED}
\end{aligned}$$

Applying Lemma 1.1 with **I** = **Brier** we get

$$\begin{aligned}
(\#)\quad \mathbf{Brier}_{\mathcal{A}}(\mathbf{c}, @) &= 2^{1-N} \sum_{A \in \mathcal{A}} (\,@(A) - c(A))^2 \\
&= 2^{1-N} \sum_{B} [\,(1 - c_1)^2 - 2(1 - c_1)\mathbf{c}(B) + \mathbf{c}(B)^2]
\end{aligned}$$

since

$$\begin{aligned}
\mathbf{Brier}_2 &= \sum_B [\,1 - \mathbf{c}(B \vee s_1)]^2 &&= \sum_B [\,(1 - c_1) - \mathbf{c}(B)]^2 \\
& &&= \sum_B [\,(1 - c_1)^2 - 2(1 - c_1)\mathbf{c}(B) + \mathbf{c}(B)^2] \\
\mathbf{Brier}_4 &= \sum_B \mathbf{c}\,(B)^2
\end{aligned}$$

**Lemma-1.2**

$$(\sum_{n=2}^{N-1} c_n)^2 \quad = \sum_{n=2}^{N-1} c_n{}^2 + 2 \sum_{n=2}^{N-2} \sum_{j>n}^{N-1} c_n\, c_j$$

Proof by induction. Easy.

**Lemma-1.3**

$$\mathbf{Brier}_{\mathcal{S}}(\mathbf{c}, @) \quad = \frac{2}{N}[(1 - c_1)^2 + \sum_{n=2}^{N-1} c_n{}^2 - (1 - c_1)(\sum_{n=2}^{N-1} c_n) + \sum_{n=2}^{N-2} \sum_{j>n}^{N-1} c_n\, c_j]$$

**Proof**: Using the definition of the Brier score and the fact that $s_1$ is true, we have

$$\begin{aligned}\mathbf{Brier}_S(\mathbf{c},@) &= \frac{1}{N}[(1-c_1)^2 + \sum_{n=2}^{N-1} c_n{}^2 + (1 - \sum_{n=1}^{N-1} c_n)^2] \\ &= \frac{1}{N}[(1-c_1)^2 + \sum_{n=2}^{N-1} c_n{}^2 + ((1-c_1) - \sum_{n=2}^{N-1} c_n)^2] \\ &= \frac{1}{N}[(1-c_1)^2 + \sum_{n=2}^{N-1} c_n{}^2 + (1-c_1)^2 - 2(1-c_1)\sum_{n=2}^{N-1} c_n + (\sum_{n=2}^{N-1} c_n)^2] \\ &= \frac{1}{N}[(1-c_1)^2 + \sum_{n=2}^{N-1} c_n{}^2 + (1-c_1)^2 - 2(1-c_1)\sum_{n=2}^{N-1} c_n \\ &\quad + \sum_{n=2}^{N-1} c_n{}^2 + 2\sum_{n=2}^{N-2}\sum_{j>n}^{N-1} c_n\, c_j] \quad (\text{Lemma} - 1.2)\end{aligned}$$

Then grouping like terms and factoring out 2 yields the desired result. QED

**Lemma-1.4**

$$\sum_{n=2}^{N-1} c_n \quad = 2^{3-N} \sum_{B\in\mathcal{B}} \mathbf{c}\,(B)$$

**Proof**: For each $n = 2,3,\dots,N-1$, each $s_n$ appears in half of the $2^{N-2}$ disjunctions with disjuncts drawn from $\mathcal{B}$. As a result, each $c_n$ appears as a summand $2^{N-3}$ times among the sums that express the various $\mathbf{c}(B)$. So $\sum_{B\in\mathcal{B}} \mathbf{c}\,(B) = 2^{N-3}\sum_{n=2}^{N-1} c_n$. QED

**Lemma-1.5**

$$\sum_{B\in\mathcal{B}} \mathbf{c}\,(B)^2 \quad = 2^{N-3}[\sum_{n=2}^{N-1} c_n{}^2 + \sum_{n=2}^{N-2}\sum_{j>n}^{N-1} c_n\, c_j]$$

**Proof**: We proceed by induction starting with the first meaningful case of $N = 4$, where calculation shows $\sum_B \mathbf{c}\,(B)^2 = (c_2 + c_3)^2 + c_2{}^2 + c_3{}^2 = 2[c_2{}^2 + c_3{}^2 + c_2 c_3]$. Now, assume the identity holds for disjunctions $B$ of elements of $\mathcal{B}$ and show that it holds for disjunctions $A$ of elements of $\mathcal{B} \cup \{s_N\}$.

$$\begin{aligned}
\sum_A \mathbf{c}\,(A)^2 &= \sum_B \mathbf{c}\,(B)^2 + \sum_B \mathbf{c}\,(B \vee s_N)^2 \\
&= \sum_B \mathbf{c}\,(B)^2 + \sum_B \left(\,\mathbf{c}(B)^2 + 2c_N\mathbf{c}(B) + c_N{}^2\right) \\
&= 2\sum_B \mathbf{c}\,(B)^2 + 2c_N \sum_B \mathbf{c}\,(B) + \sum_B c_N{}^2 \\
&= 2 \cdot 2^{N-3}\left[\sum_{n=2}^{N-1} c_n{}^2 + \sum_{n=2}^{N-2}\sum_{j>n}^{N-1} c_n\, c_j\right] + 2c_N \sum_B \mathbf{c}\,(B) + \sum_B c_N{}^2 && \binom{\text{Induction}}{\text{Hypothesis}} \\
&= 2^{N-2}\left[\sum_{n=2}^{N-1} c_n{}^2 + \sum_{n=2}^{N-2}\sum_{j>n}^{N-1} c_n\, c_j\right] + 2^{N-2}c_N \sum_{n=2}^{N-1} c_n + \sum_B c_N{}^2 && (\text{Lemma} - 1.4) \\
&= 2^{N-2}\left[\sum_{n=2}^{N-1} c_n{}^2 + \sum_{n=2}^{N-2}\sum_{j>n}^{N-1} c_n\, c_j\right] + 2^{N-2}c_N \sum_{n=2}^{N-1} c_n + 2^{N-2}c_N{}^2 && \text{Since } |\mathcal{B}| = 2^{N-2} \\
&= 2^{N-2}\left[\sum_{n=2}^{N} c_n{}^2 + \sum_{n=2}^{N-1}\sum_{j>n}^{N} c_n\, c_j\right] && \text{QED}
\end{aligned}$$

Plugging the results of the last two lemmas into Lemma-1.3 produces a result of

$$\begin{aligned}
\mathbf{Brier}_{\mathcal{S}}(\mathbf{c}, @) &= \frac{2}{N}[(1-c_1)^2 + 2^{3-N}\sum_{B\in\mathcal{B}} \mathbf{c}\,(B)^2 - 2^{3-N}(1-c_1)\sum_{B\in\mathcal{B}} \mathbf{c}\,(B)] \\
&= \frac{2}{N}\sum_{B\in\mathcal{B}} [\,2^{2-N}(1-c_1)^2 + 2^{3-N}\mathbf{c}(B)^2 - 2^{3-N}(1-c_1)\mathbf{c}(B)] \\
&= \frac{2^{3-N}}{N}\sum_{B\in\mathcal{B}} [\,(1-c_1)^2 + 2\mathbf{c}(B)^2 - 2(1-c_1)\mathbf{c}(B)]
\end{aligned}$$

Comparing this to (#) we see that it is just $\frac{N}{4}$ times $\text{Brier}_{\mathcal{S}}(\mathbf{c}, @)$, as we aimed to prove. QED.

**Theorem-2**. When inaccuracy over $\mathcal{A}$ is measured using the Brier score, the least accurate credal states are those which assign credence 1 to some false atom of $\mathcal{A}$.

**Proof**: As before, suppose that $@(s_1) = 1$, and let $\mathbf{c}$ be a credence function that assigns credence 1 to some false atom $s_2, s_3, \ldots, s_N$ of $\mathcal{A}$. In light of Theorem-1 it suffices to show that $\mathbf{Brier}_{\mathcal{S}}(\mathbf{c}, @) > \mathbf{Brier}_{\mathcal{S}}(\mathbf{b}, @)$ where $\mathbf{b}$ does not assign credence 1 to any false atom. Start by noting that for any credence function $\pi$ defined on the atoms of $\mathcal{A}$ one has

$$\begin{aligned}\mathbf{Brier}_S(\pi,@) &= \frac{1}{N}[(1-\pi_1)^2 + \sum_{n=2}^{N-1} \pi_n{}^2 + (1 - \sum_{n=1}^{N-1} \pi_n)^2] \\ &= \frac{1}{N}[1 - 2\pi_1 + \sum_{n=1}^{N-1} \pi_n{}^2 + (1 - \sum_{n=1}^{N-1} \pi_n)^2]\end{aligned}$$

But, since each $\pi_n \in [0,1]$ is non-negative, it follows that $\pi_1 \geq \pi_1{}^2, \pi_2 \geq \pi_2{}^2, \ldots, \pi_N \geq \pi_N{}^2$ with the inequality strict in each case unless $\pi_n$ is either 1 or 0.

This means that the sum $\sum_{n=1}^{N-1} \pi_n{}^2 + (1 - \sum_{n=1}^{N-1} \pi_n)^2$ is less than or equal to 1, with equality if and only if exactly one of the atoms $s_n$ is assigned probability 1 (and the rest have probability zero). As a result, $\mathbf{Brier}_S(\pi,@) \leq \frac{2}{N}(1-\pi_1)$ with equality if and only if exactly one of the atoms $s_n$ is assigned probability 1. So, there are three relevant cases:

(i) If $\pi$ assigns some false atom probability 1, $\mathbf{Brier}_S(\pi,@) = \frac{2}{N} \cdot (1-0) = \frac{2}{N}$.

(ii) If $\pi$ assigns the true atom probability 1, $\mathbf{Brier}_S(\pi,@) = \frac{2}{N} \cdot (1-1) = 0$.

(iii) If $\pi$ does not assign any atom probability 1, $\mathbf{Brier}_S(\pi,@) < \frac{2}{N} \cdot (1-c_1) \leq \frac{2}{N}$.

So, since **c** fits case (i) and **b** fits case (ii) or (iii) we have the desired result. QED

**Theorem-3**: Let $\mathcal{A}$ be an algebra of propositions generated by atoms $a_1, \ldots, a_{2N}$, where $a_1$ is the truth. Let $P$ and $Q$ be probability functions defined on $\mathcal{A}$. $P$ assigns all its mass to the first $N$ atoms, so that $P(a_1 \vee \ldots \vee a_N) = 1$, and it also assigns some positive probability to $a_1$. $Q$ assigns all its mass to the false atom $a_{2N}$, so that $Q(a_{2N}) = 1$. Then, for any proper score **I** satisfying Truth-directedness, Extensionality and Negation Symmetry we have $\mathbf{I}(v_1, P) < \mathbf{I}(v_1, Q)$ where $v_1$ is the truth-value assignment associated with $a_1$ (i.e., where $v_1(X) = 1$ if and only if $a_1$ entails $X$).

**Proof**: We can divide the algebra $\mathcal{A}$ into four quadrants

$$\begin{aligned}\mathcal{A}^1 &= \{X \in \mathcal{A}: a_1 \vDash X \text{ and } a_{2N} \vDash X\} \\ \mathcal{A}^2 &= \{X \in \mathcal{A}: a_1 \vDash X \text{ and } a_{2N} \nvDash X\} \\ \mathcal{A}^3 &= \{X \in \mathcal{A}: a_1 \nvDash X \text{ and } a_{2N} \vDash X\} \\ \mathcal{A}^4 &= \{X \in \mathcal{A}: a_1 \nvDash X \text{ and } a_{2N} \nvDash X\}\end{aligned}$$

We know the following:

- $Q$ is maximally accurate on $\mathcal{A}^1 \cup \mathcal{A}^4$. Every proposition in $\mathcal{A}^1$ is true, and $Q$ assigns it a probability of 1. Every proposition in $\mathcal{A}^4$ is false, and $Q$ assigns it a probability of 0.
- $Q$ is maximally inaccurate on $\mathcal{A}^2 \cup \mathcal{A}^3$. Every proposition in $\mathcal{A}^2$ is true, and $Q$ assigns it a probability of 0. Every proposition in $\mathcal{A}^3$ is false, and $Q$ assigns it

a probability of 1.

- $P$ is maximally accurate on $\mathcal{A}^3 \cup \mathcal{A}^4$. Every proposition in $\mathcal{A}^3 \cup \mathcal{A}^4$ is false, and $P$ assigns it a probability of 0.
- Each quadrant has $2^{2N-2}$ elements.

**Lemma-3.1**: When $a_1$ is true, the accuracy score of $P$ over the propositions in $\mathcal{A}^1$ is identical to the accuracy score of $P$ over the propositions in $\mathcal{A}^2$.

**Proof**: Note first that the function $F: \mathcal{A}^1 \rightarrow \mathcal{A}^2$ that takes $X$ to $X \wedge \neg a_{2N}$ is a bijection of $\mathcal{A}^1$onto $\mathcal{A}^2$. Since every proposition in $\mathcal{A}^1 \cup \mathcal{A}^2$ is true, we can then write the respective accuracy scores of $\mathcal{A}^1$ and $\mathcal{A}^2$ as

$$\begin{aligned}\mathbf{I}_{\mathcal{A}^1}(a_1, P) &= 2^{2-2N} \cdot \sum_{X \in \mathcal{A}^1} \mathbf{I}\,(1, P(X)) \\ \mathbf{I}_{\mathcal{A}^2}(a_1, P) &= 2^{2-2N} \cdot \sum_{X \in \mathcal{A}^1} \mathbf{I}\,(1, P(X \wedge \neg a_{2N}))\end{aligned}$$

Note: $X$ ranges over $\mathcal{A}^1$ in both summations. But since $P(a_{2N}) = 0$ we have $P(X) = P(X \wedge a_{2N})$ for each $X$in $\mathcal{A}^1$. Since $\mathbf{I}$ is extensional, this means that $\mathbf{I}(1, P(X)) = \mathbf{I}(1, P(X \wedge a_{2N}))$ for each $X$in $\mathcal{A}^1$. And, it follows that $\mathbf{I}_{\mathcal{A}^1}(a_1, P)$ and $\mathbf{I}_{\mathcal{A}^2}(a_1, P)$ are identical. (Note that even if $P(a_{2N}) > 0$, Truth-directedness entails that $\mathbf{I}_{\mathcal{A}^1}(a_1, P) < \mathbf{I}_{\mathcal{A}^2}(a_1, P)$.)

**Lemma-3.2**: When $a_1$ is true, the accuracy score of $Q$ over $\mathcal{A}^2$ is identical to the accuracy score of $Q$ over $\mathcal{A}^3$.

**Proof**: To see this, note first that the function $G: \mathcal{A}^2 \rightarrow \mathcal{A}^3$ that takes $X$ to $G(X) = \neg X$ is a bijection (i.e., the negation of everything in $\mathcal{A}^2$ is in $\mathcal{A}^3$ and vice-versa). This, together with the fact that $\mathcal{A}^2$ contains only truths and $\mathcal{A}^3$ contains only falsehoods, lets us write

$$\begin{aligned}\mathbf{I}_{\mathcal{A}^2}(a_1, Q) &= 2^{2-2N} \cdot \sum_{X \in \mathcal{A}^2} \mathbf{I}\,(1, Q(X)) \\ \mathbf{I}_{\mathcal{A}^3}(a_1, Q) &= 2^{2-2N} \cdot \sum_{X \in \mathcal{A}^2} \mathbf{I}\,(0, Q(\neg X))\end{aligned}$$

But since $\mathbf{I}$ is negation symmetric, $\mathbf{I}(1, Q(X)) = \mathbf{I}(0, Q(\neg X))$ for every $X$, which means that $\mathbf{I}_{\mathcal{A}^2}(a_1, Q) = \mathbf{I}_{\mathcal{A}^3}(a_1, Q)$. (Note that this proof made no assumptions about $Q$ except that it was a probability.)

**Lemma-3.3**: If $P(a_1) > 0$, the accuracy score of $P$ over $\mathcal{A}^2$ is strictly less than the accuracy score of $Q$ over $\mathcal{A}^2$.

**Proof**: Since $Q(X) = 0$ everywhere on $\mathcal{A}^2$ we have

$$\begin{aligned}\mathbf{I}_{\mathcal{A}^2}(a_1, P) &= 2^{2-2N} \cdot \sum_{X \in \mathcal{A}^2} \mathbf{I}\,(1, P(X)) \\ \mathbf{I}_{\mathcal{A}^2}(a_1, Q) &= 2^{2-2N} \cdot \sum_{X \in \mathcal{A}^2} \mathbf{I}\,(1{,}0)\end{aligned}$$

But, by Truth Directedness $\mathbf{I}(1{,}0) > \mathbf{I}(1, P(X))$ since $P(a_1) > 0$ implies that $P(X) > 0$ for all $X \in \mathcal{A}^2$. Thus $\mathbf{I}_{\mathcal{A}^2}(a_1, Q) > \mathbf{I}_{\mathcal{A}^2}(a_1, P)$.

To complete the proof of the theorem we need only note that

$$\begin{aligned}\mathbf{I}_{\mathcal{A}}(a_1, P) &= \frac{\mathbf{I}_{\mathcal{A}^1}(a_1, P)}{4} + \frac{\mathbf{I}_{\mathcal{A}^2}(a_1, P)}{4} && \text{(since } P \text{ is perfect on } \mathcal{A}^3 \cup \mathcal{A}^4\text{)} \\ &= \frac{\mathbf{I}_{\mathcal{A}^2}(a_1, P)}{2} && \text{Lemma} - 3.1 \\ &< \frac{\mathbf{I}_{\mathcal{A}^2}(a_1, Q)}{2} && \text{Lemma} - 3.3 \\ &= \frac{\mathbf{I}_{\mathcal{A}^2}(a_1, Q)}{4} + \frac{\mathbf{I}_{\mathcal{A}^3}(a_1, Q)}{4} && \text{Lemma} - 3.2 \\ &= \mathbf{I}_{\mathcal{A}}(a_1, Q) && \text{(since } Q \text{ is perfect on } \mathcal{A}^1 \cup \mathcal{A}^4\text{)}\end{aligned}$$

# WHAT IS THE EPISTEMIC SIGNIFICANCE OF DISAGREEMENT?

N. Gabriel MARTIN

ABSTRACT: Over the past decade, attention to epistemically significant disagreement has centered on the question of whose disagreement qualifies as significant, but ignored another fundamental question: what is the epistemic significance of disagreement? While epistemologists have assumed that disagreement is only significant when it indicates a determinate likelihood that one's own belief is false, and therefore that only disagreements with epistemic peers are significant at all, they have ignored a more subtle and more basic significance that belongs to all disagreements, regardless of who they are with—that the opposing party is wrong. It is important to recognize the basic significance of disagreement since it is what explains all manners of rational responses to disagreement, including assessing possible epistemic peers and arguing against opponents regardless of their epistemic fitness.

KEYWORDS: social epistemology, disagreement, epistemic peers

In epistemology over the past decade a lively discussion about disagreement has focussed on the conditions under which disagreement becomes epistemically significant. This dispute ignores the more basic question—what significance can or does disagreement have?

Although this basic question has not been explored in the literature, the way that the literature asks its own question presupposes an answer. In this article I will raise the question of whether the significance of disagreement presupposed by the epistemology of disagreement really is the significance disagreement has. I will argue that the significance presupposed throughout the sub-field—that disagreement qualifies the likelihood of the falsity of one's own belief—is not its most general or basic. Disagreement's significance does not concern oneself but rather one's opponent. It is that the person with whom one disagrees is wrong.

I will defend this claim and explain why it matters. First, I will explain how the discussion of the epistemological consequences of disagreement presuppose what I will call a self-reflexive significance. This will allow me, in section two, to show why this significance cannot belong generally to all disagreements, but can only belong to disagreements when they possess certain qualifying characteristics. That will in turn make it possible, in section three, to settle a current debate within the sub-field concerning whether the peerhood of interlocutors is to be presupposed. Disagreement only possesses this significance conditionally, and it is only when it has been determined to meet certain conditions that it can be considered significant. This raises the crucial question—what reason is there to evaluate disagreements on the basis of peerhood? There must be some basic and general significance belonging to disagreement as such that makes evaluating opponents make sense. I will address this question in section four. Finally, in section five, I will explain why this matters—the epistemological significance of disagreement is not simply what it indicates about your own belief, but what it indicates about the beliefs of others. This means that the epistemic significance of disagreement is fundamentally social and intersubjective.

## 1. Disagreement's Self-Reflexive Significance

Let me restate the question: what significance can or does disagreement have? An answer to this question would have to disclose the significance of disagreement itself—whether disagreement on some matter, in and of itself, has any bearing on that matter. Put another way, the question is whether any light can be shed on that which we disagree about (the *disputandum*) by the very fact that we disagree about it. The strictly epistemic question is insensitive to the many additional questions about the context of the disagreement, including the psychology of its participants or their social relations, that could be raised. Doubtless, dispute tells us something about the attitudes of the people involved, and controversy tells us something about the culture in which it exists, but the epistemology of disagreement sets these matters asides. It is concerned narrowly with whether the mere fact that there is a disagreement can indicate something about the disputandum, either directly or indirectly.

The question of the significance of disagreement also excludes questions about the significance of the positions in conflict themselves. Of course a disagreement consists of positions, hopefully supported by reasons and evidence, which bear upon the disputed matter in all sorts of relevant ways. This is not what epistemology of disagreement is concerned with either. The epistemic significance of disagreement itself is not due to the significance of those positions or what supports them: it is due solely to the significance of the fact that the matter is in dispute.

It has been proposed that the epistemic significance of disagreement, defined in such a narrow way, is profound. Disagreement may bring with it sceptical consequences if it can indicate an increased likelihood of error on one's own part.[1] This view is shared by many within the sub-field.[2] If disagreement can indicate, either generally or under certain circumstances, that there is considerable chance that you have formed an incorrect belief, then diminished confidence in your position is warranted. This is the way that the problem of disagreement was originally framed by Sextus Empiricus.[3]

Most contributors to the literature are in agreement that the appropriate response to epistemically significant disagreement is to become less confident in the correctness of one's own position.[4] That is, there is a consensus among most social epistemologists that faced with a disagreement of epistemic significance a person should check the confidence with which they hold their controversial position. That diminished confidence is the appropriate response to any disagreement which possesses epistemic significance is rarely disputed.[5] Instead,

[1] As considered here, the problem of disagreement only arises for those involved—it is not a question of what disagreement means for one who occupies a neutral position, but what disagreement means when you are one of the parties embroiled in it.

[2] For example, see Adam Elga, "Reflection and Disagreement," *Noûs* 41, 3 (2007): 497; Robert Mark Simpson, "Epistemic Peerhood and the Epistemology of Disagreement," *Philosophical Studies* 164, 2 (2013): 561-577; Ernest Sosa, "The Epistemology of Disagreement," in *Social Epistemology*, ed. Adrian Haddock, Alan Millar and Duncan Pritchard (Oxford: Oxford University Press, 2010), 278-297. Others disagree. According to Thomas Kelly, "The Epistemic Significance of Disagreement," in *Oxford Studies in Epistemology, Volume 1*, ed. John Hawthorne and Tamar Gendler (Oxford: Oxford University Press. 2005), 191, diminished confidence in the correctness of one's own position should not be the consequence of any disagreement, but that is because he denies that there are any epistemically significant disagreements.

[3] Sextus Empiricus, *Outlines of Scepticism*, trans. Julia Annas, and Jonathan Barnes (New York: Cambridge University Press, 1994), 41.

[4] See especially Adam Elga, "Reflection and Disagreement," 497. There is, however, no obligation to diminish confidence according to those who deny the claim that there is a unique rational doxastic response to any body of evidence (See Nathan Ballantyne, E.J. Coffman, "Uniqueness, Evidence, and Rationality," *Philosophers' Imprint* 11, 18 [2011]: 1-13; Roger White, "Epistemic Permissiveness," *Philosophical Perspectives* 19, 1 [2005]: 445–459). However if permissiveness should apply to two conflicting theories, it does not seem appropriate to call this a disagreement, since while the two theories conflict they do not invalidate one another.

[5] An exception is Gurpreet Rattan "Disagreement and the First-Person Perspective," *Analytic Philosophy* 55, 1 (2014): 331–353. Rattan argues that disagreement between epistemic peers indicates that there is a misunderstanding or equivocation at work, and that "the epistemic limits of intersubjective understanding" (Rattan, "Disagreement and the First-Person Perspective," 351) justify what he calls "limited permission to persist in confidence" (Rattan, "Disagreement and the First-Person Perspective," 350) until the matter is cleared up.

the debate is over whether there exist any disagreements which in and of themselves call for that response.[6]

I will ask a different question. Disagreement that has sceptical consequences for a reasonable participant must have a certain kind of significance. My question is: what kind of significance is it that would call for diminished confidence?

## 2. Self-Reflexive Significance Is Limited to Only Some Disagreements

For starters, this shows that the significance of disagreement is conceived in a strictly negative way—its significance is in indicating (in some way we have not determined yet) that beliefs about the matter are wrong. There are no other possibilities given that what we are considering is in no case direct or first-order evidence about the disputandum, but second-order evidence. It therefore pertains to the disputandum indirectly, by giving evidence about the truth or falsity of the beliefs about the disputandum itself. As such, all it can do is undercut the confidence with which a belief is held.

The simple possibility of error cannot be what calls for doubt in the face of disagreement. It is possible for your belief to be in error, but it is the fallibility of the belief itself which signifies this, and a disagreement can only be a reminder of the possibility of error if it is already acknowledged. If disagreement can tell you anything more about the possibility that you are wrong in a given case, it is because it is already a characteristic of your belief that it may not be right. The possibility of being wrong is a precondition for a fact to provide evidence that you are wrong—without the possibility of error, facts surrounding your belief could never have anything to do with the possibility that you could be wrong.

Disagreement must tell you something about the possibility that your belief is false if it calls for you to be less confident in your position. Only an indication concerning the likelihood that your belief is wrong can give you any reason to doubt it. The significance of disagreement must indicate something about the

[6] See David Christensen, "Epistemology of Disagreement: The Good News," *Philosophical Review* 116, 2 (2007): 187–217; David Christensen, "Disagreement as Evidence: The Epistemology of Controversy," *Philosophy Compass* 4, 5 (2009): 756–767; Richard Foley, I*ntellectual Trust in Oneself and Others* (New York: Cambridge University Press, 2001); Ernest Sosa, "The Epistemology of Disagreement," in *Social Epistemology*, eds. Alan Millar and Duncan Pritchard Adrian Haddock (Oxford: Oxford University Press, 2010), 278–97.. In "Disagreement as Evidence," David Christensen characterises this as a debate between what he calls 'conciliatory' and 'steadfast' views. That is, between interpretations of the significance of disagreement which hold that the rational response is to move closer, in some way or another, to the views of your interlocutor, and interpretations that hold that, in the face of disagreement, you are obligated to retain the confidence in your beliefs that you had going in.

*likelihood* of error if it is to warrant diminished confidence, because it is only if it qualifies the already certain possibility of error that it tells you something new.

This response must be justified by additional information about the likelihood that your belief is in error. That is to say, more must be determined about the quality or character of the disagreement. Only disagreement of a particular character can have the kind of significance which would call for you to lose *even some* confidence in your belief. But can disagreement of a certain type indicate the likelihood that one is in error?[7]

The insignificance of *unqualified* disagreement is the reason that the epistemology of disagreement is only concerned with *qualified* disagreement. It is only the disagreement of peers or superiors, which is to say those who are at least as likely as oneself to have knowledge of the matter in dispute, that is meaningful.

Peerhood, or the relative epistemic fitness of those with whom one finds oneself in dispute, is a handy way to indicate what kind of qualification of a disagreement would have the epistemic significance that we are talking about.[8] If I know that my opponent is as likely as I am to be right about the disputandum, then I also know that the likelihood of error on my own part is high; at least 50 per cent.[9]

The qualification of a given disagreement as a peer disagreement distinguishes it considerably from disagreement in general. The possibility that any one of my beliefs could be false is not determinate in any way (I cannot ascribe any statistical or comparative character to it), but a disagreement between peers is qualitatively determinate. *Peer* disagreement carries a significance that pertains to the likelihood of error in my position; it indicates that I am no more likely to be right than the person who challenges me. This does not mean that in a peer disagreement I am more likely to be wrong than in any other, unqualified disagreement—it means that in a peer disagreement the likelihood that I am wrong is certain, whereas in the other it is totally indeterminate.

[7] That it is by way of indicating an increased likelihood of error that a disagreement could indicate the need to revise one's confidence is assumed generally in the sub field. Christensen gets closest to explicitly stating it when he remarks "arbitrating the dispute from one's own perspective need not entail disregarding evidence that one might be wrong" (Christensen, "Epistemology of Disagreement," 762). It is assumed here by Christensen that 'evidence that one might be wrong' is precisely what is in question.

[8] See Bryan Frances, *Disagreement* (Cambridge: Polity, 2014), 45*ff* for what he calls 'likelihood definitions' of peerhood.

[9] Of course, this conclusion is disputed. However, there is broad consensus that a conclusion along these lines follows from disagreement among true peers. More controversy centres around the likelihood or possibility of finding such peers.

Since a peer disagreement is defined by the fact that the likelihood that I am wrong and the likelihood that my opponent is wrong are on par with one another, the response that social epistemologists generally believe peer disagreement calls for—diminished confidence—is justified. When I already know that I could be wrong, the fact that I disagree with one of my peers tells me that there is a good chance that I actually am wrong. Once again, we can compare this to disagreement of an unqualified kind, which tells me nothing about the relationship of this instance to that invariant possibility.

Is this the significance of disagreement—that it can inform me about the likelihood that I am in error? That is the consensus in the sub-field, and it is why most of the discussion concerns the conditions under which an opponent can and must be considered one's peer. In the section that follows, I will briefly explain some of the key positions in this debate. As I explain, a key determinant as to whether the disagreements we find ourselves in can be expected to show themselves to be peer disagreements comes down to whether or not peerhood can be thought to be assumed by default, or whether it must be demonstrated. I argue that it must be demonstrated, and that this points to the existence of a more basic and general significance of disagreement.

## 3. Peerhood Cannot Be Assumed

Is it the case that disagreement itself is enough to disqualify an adversary from peerhood with respect to the matter, as Thomas Kelly and Ernest Sosa argue?[10] Can you consider your opponent less likely than you are to be right, simply on the basis of their being your opponent, or having beliefs which you believe to be wrong? It would seem to follow from your having a belief that anyone who rejected that belief was in your lights far less likely to have knowledge of the matter.

Or, is it necessary to assess your interlocutors on grounds that are independent of the reasoning supporting the position that brings you into conflict, as David Christensen claims?[11] He argues that, in order to avoid begging the question against your interlocutors, assessment must be on grounds other than that which is at issue in the disagreement in question.

Even if independent grounds for dismissing an adversary as sub-par are required, in real-world controversies are peers likely to be thin on the ground, as

[10] See Kelly, "The Epistemic Significance of Disagreement," 2005 and Sosa, "The Epistemology of Disagreement," 2010.

[11] Christensen, "Epistemology of Disagreeement," 2007.

Adam Elga argues,[12] because the basis on which to assess the likelihood of being right of those who do not share your views is lacking?

Elga makes the latter claim because our disagreements about controversial questions are generally not anomalies. Rather, disagreements arise within polarised and polarising "clusters of controversy"[13] and when those with whom you disagree on one thing tend to disagree with you on all or most related questions as well, you are not going to find the means to assess their level of expertise in that general area.

All of the conceptions of interlocutor assessment discussed here in brief concern how peerhood or non-peerhood is to be determined. What none of these considerations take into account is whether or not interlocutors must be considered peers by default or whether peerhood must be earned. However, this question is, if not decisive, at least of great significance with respect to the determination. This is because, as Elga points out, assessment will often be impossible to accomplish and therefore the question of whether or not a particular interlocutor is one's peer will come down to whether or not they must be considered one prior to assessment. This question has become the focus of some more recent attention in the sub-field.

Peerhood cannot be a default. That is because it means that you are no more likely to be right than you are to be wrong. Not just any disagreement can have that significance, nor can just any adversary be considered peer. This is not because a peer disagreement necessarily means that your likelihood of being wrong is higher than it is in a disagreement with an unqualified interlocutor. It isn't. The relative likelihood that you are wrong in an unqualified disagreement is uncertain, so it may be higher or lower or identical to that in a peer disagreement. Peerhood cannot be a default precisely because it denotes demonstration of the relative likelihood of the possibility that yours is the position that is wrong in the dispute. It is only once your opponent has been found to be at least as likely as yourself to be right about the matter that their disagreement has the kind of significance concerned: the ability to "say" something about the likelihood of error in your own position. Without a demonstration of the relative likelihood of your opponent's being right, the fact that a true disagreement means that one or the other of you must be wrong does not in any way qualify the general, indeterminate possibility that you could be wrong that you must acknowledge from the start.

Peerhood only means something if it denotes specific characteristics of the interlocutor in question. Therefore it cannot be assumed. This means that the

[12] Elga, "Reflection and Disagreement," 2007.

[13] Elga, "Reflection and Disagreement," 2007, 493.

assessment of interlocutors and discrimination between those that are peers and those that are not peers is crucial to appreciating the significance of disagreement, and if this assessment is not or cannot be performed then the disagreement cannot be considered 'peer'.

There are a few epistemologists who argue otherwise.[14] Contesting Elga's conclusion about controversy clusters, Robert Mark Simpson states that "remaining steadfast in the face of a disagreement is only justified when one has some basis for thinking that one's opponents are epistemically less well-credentialed than oneself with respect to the subject of the disagreement."[15] In other words, one must have a reason for considering an opponent *sub*-par. Peerhood, Simpson assumes, is the default.

In arguing against Elga from the assumption of default peerhood, Simpson also reveals that Elga's argument is based on the contrary assumption—that peerhood must be demonstrated. 'Clusters of controversy' refers to the tendency of those with whom one disagrees about serious, real-world issues to also have conflicting views, by and large, about other, related matters. The consequence of these controversy clusters is that we can expect finding real disagreements with someone who can be deemed a peer to be rare. This is because determining that a particular interlocutor is an epistemic peer with respect to a particular matter is only possible after assessing their ability to get such things right. However, because controversies tend not to exist in isolation but in clusters, it is likely that when you disagree with someone on one question you will also disagree with them on related questions and so you will lack the kind of evidence of their being right (by your own lights) on the relevant sorts of questions necessary to establish their peerhood. The way that disagreements form in society makes it improbable for peerhood to be attributed to an opponent in a disagreement.

The example offered by Elga is a disagreement about the ethics of abortion.[16] Elga supposes that we should expect two people who disagree about abortion also to disagree about related controversies, such as political affiliation, religion, and the definitions of life and personhood. The broad disagreement between the two means they lack any basis on which to establish the other's peer *bona fides*. This

[14] See Catherine Z. Elgin, *Considered Judgment* (Princeton: Princeton University Press, 1996); Richard Feldman, "Epistemological Puzzles about Disagreement," *Epistemology Futures*, ed. Stephen Hetherington (New York: Oxford University Press, 2006): 216–236; Robert Mark Simpson, "Epistemic Peerhood and the Epistemology of Disagreement," *Philosophical Studies* 164, 2 (2013): 561–577; Ben Sherman, "Unconfirmed Peers and Spinelessness," *Canadian Journal of Philosophy* 45, 4 (2015): 425–444.

[15] Simpson, "Epistemic Peerhood and the Epistemology of Disagreement," 576.

[16] Elga, "Reflection and Disagreement," 493.

means that each party will lack any basis on which to perform the required assessment of the epistemic capacity of their interlocutor. Essentially, Elga argues that if an opponent in a disagreement is going to qualify as one's peer, their capacity to get the matter right must be demonstrated. And what is needed in order to demonstrate it is the possibility of pointing to the correctness of their views on related matters.

Rather than taking on Elga's factual claim that real-world disagreements exist within clusters of controversy, as Sarah McGrath and others do,[17] Simpson disputes Elga's views on what we are to make of the disagreements of those we cannot assess. Elga, we saw, takes the lack of any basis on which to establish an opponent's credentials as sufficient grounds for their dismissal as sub-par. This makes sense because of his assumption that others are not to be considered one's epistemic peers by default; that instead they must earn the right to be considered peers.

Simpson, on the other hand, claims that one must consider a disputant one's epistemic peer until one has evidence that they are *not*. Thus, the phenomenon of controversy clusters Elga will have a consequence opposite to that attributed to it by Elga.

This difference in the epistemologies of disagreement of Elga and Simpson does not belong to the more frequently raised debate about that basis on which peerhood can be assessed, it is about whether assessment is necessary in the first place, and what happens if it cannot be performed. Although this only appears as an explicit theme in epistemology of disagreement in Ben Sherman (even Elga and Simpson do not make the importance of the question explicit),[18] it determines for the most part whether or not peer disagreements, with the serious consequences

[17] In "Moral Disagreement and Moral Expertise," *Oxford Studies In Metaethics*, Vol. 3, ed. Russ Shafer-Landau (Oxford: Oxford University Press, 2008): 87–108, Sarah McGrath argues that while controversies do cluster, these clusters form upon common ground sufficient to allow for the assessment of opponents.

[18] In "Unconfirmed Peers and Spinelessness" Sherman brings attention to the role played by the decisive difference between what he calls "presumption of peerhood" and "presumption in favour of self-trust" (Sherman, "Unconfirmed Peers and Spinelessness," 430). Shortly, I will explain how the problematic of presumption of peerhood differs from the problem of default peerhood. Whether or not a given theory of disagreement has it that others are to be considered peer by default plays an enormous role in determining whether peers, as that theory defines them, will be such as can be expected to be found. His answer to the question (See Sherman, "Unconfirmed Peers and Spinelessness," 433–434) supports the complaints of Christensen and Simpson that to discount an adversary without adequate evidence of their epistemic inferiority is "question-begging" (See Christensen, "Epistemology of Disagreement," 198 and Simpson, "Epistemic Peerhood and the Epistemology of Disagreement," 575–576).

for doxastic confidence they bring with them, are likely to characterise the important controversies that dominate philosophy and the broader culture.

The conflicting assumptions, implicit for the most part, about whether or not peerhood can be presupposed, are applicable in the kind of case considered by Elga, McGrath, and Simpson, in which what is at issue is what it is rational to believe about your interlocutor's epistemic abilities in the absence of conclusive evidence one way or the other. Simpson's opposition to what Sherman calls the "presumption in favour of self-trust"[19] is motivated by the concern that to discount an adversary without adequate evidence of their epistemic inferiority is "question-begging" or "bootstrapping."[20] Foley defends the presumption on the basis that a general *prima facie* trust in oneself is necessary for any knowledge at all.[21]

Along the same lines, Sherman argues that there should be a "presumption of peerhood."[22] Discussing whether disagreement should come with a "presumption in favour of self-trust" as Foley argues,[23] or a "presumption of peerhood", Sherman argues that the right to legitimately dismiss an opponent as sub-par must be 'earned'. Sherman calls this "earning a spine" in reference to Elga's concern about "spinelessness" as a consequence of peer disagreement.[24] Spinelessness is Elga's pejorative expression for accepting diminished confidence in the face of disagreement. Sherman's point is that spinelessness can be overcome, but not without work. In other words, peerhood must be presupposed.

Foley defends the presumption in favour of self-trust on the basis that the presupposition of a general *prima facie* trust in oneself is necessary for any

[19] Sherman, "Unconfirmed Peers and Spinelessness," 430.

[20] See Christensen, "Epistemology of Disagreement," 198; Simpson, "Epistemic Peerhood and the Epistemology of Disagreement," 575–576; Sherman, "Unconfirmed Peers and Spinelessness," 433–434.

[21] Richard Foley, *Intellectual Trust in Oneself and Others* (New York: Cambridge University Press, 2001), 108.

[22] Sherman, "Unconfirmed Peers and Spinelessness," 430.

[23] Foley defends the presumption in favour of self-trust on the basis that the presupposition of a general *prima facie* trust in oneself is necessary for any knowledge at all (See Foley, *Intellectual Trust in Oneself and Others*, 108). Self-trust is epistemically essential: even the decision to trust another is only possible because of a more fundamental trust in one's own judgment that that trust is warranted. But just because there is always self-trust involved in any judgment, that does not mean that one is compelled to prefer one's own judgments, reconsidered, over another's. When one reconsiders one's position in the way that is necessary in order to consider a disagreement, it is necessary to trust in the judgment being performed at the moment, but there is no epistemic necessity to trust the previous judgement just because it had been made by the same person, let alone prefer it to the judgment of an interlocutor.

[24] Sherman, "Unconfirmed Peers and Spinelessness," 431.

knowledge at all.[25] Self-trust is epistemically essential: even trust in another is only possible because of a more fundamental trust in one's own judgment that they are to be trusted.

This does not mean, however, that one is compelled to prefer one's own judgments over another's. When reconsidering one's position in order to assess the significance that may or may not be presented by a disagreement it is necessary to trust in the judgment one is utilising at the moment. This does not mean that one is obligated to trust one's previous judgement just because it happened to have been made by oneself, since it is not the judgment being relied upon at the moment. The fact that you are the same person who made the judgment that you must decide whether or not to prefer to that of your interlocutor does not preclude you from changing your mind and siding with them, because although the judgment was made by you it is a distinct act of judgment from the one which you are employing at present. Therefore, there is no obligation to trust it.

Simpson and Sherman raise the important question of whether or not one's adversaries must be assumed to be peers, or whether they can only be thought to be peers when positive evidence of their peerhood has been brought to bear. However, they approach the question from the wrong direction. What matters with respect to whether or not adversaries are to be considered peers by default is not what is rational to believe about the epistemic ability of those with whom we find ourselves in disagreement, but what the conditions are for finding disagreements in which we are involved to be epistemically significant. In this regard there can be no doubt, self-reflexive epistemic significance belongs to disagreements only on the condition that those disagreements are qualified by the peer condition or some similar factor which cannot be granted universally by default.

It is not because of an epistemic principle, either that commanding self-trust or that proscribing question-begging, that peerhood cannot be presupposed. Peerhood cannot be presupposed because it stipulates a quality that distinguishes it from just any disagreement. It is only when peerhood names a quality that makes it possible to determine that the likelihood that one's opponent in a disagreement is wrong is no higher than the likelihood that one is wrong oneself that peerhood can fulfil its essential function of distinguishing the class of disagreement that belongs to it from disagreement in general. If peerhood is presupposed, then it does not contain the crucial qualifying character that it is supposed to. If peerhood is default, then a peer disagreement no longer has the significance it is meant to

[25] Foley, *Intellectual Trust in Oneself and Others*, 108.

have—indicating the likelihood of error on one's own part. Default peerhood, therefore, is a meaningless notion.

As long as disagreement is going to have any consequence involving diminished confidence in one's own position, it will be necessary for some characteristic of the disagreement to contribute information about the likelihood of an error on one's own part. If the characteristic in question is not the peerhood of one's opponent, it must be some other qualifying characteristic.

The quality of peerhood only makes sense if the difference between peer and non-peer is enforced. If everyone is your peer by default, then peerhood is not dependent upon evaluation. This equality of all speakers could not be called peerhood, which denotes a class, and it could not signify equality with respect to the likelihood of being right of you and your interlocutor. Only peerhood ascertained through assessment can denote equal likelihood of being right, while general equality of all speakers, which is a presupposition or a principle that is not given in evidence, can only denote that the relative likelihood of being right is indeterminate. If disagreement means only that you cannot know whether or not you are more or less or just as likely as your interlocutor to be right about the disputandum (which seems to me to be the correct interpretation of the meaning of disagreement according to Sextus) it does not bring with it the self-reflexive epistemic significance that calls for diminished confidence.

## 4. The Intersubjective Significance of Disagreement

In section one, I established that if the reasonable response to disagreement is reduction in the confidence one places in one's own position, then the significance of disagreement must pertain to the likelihood that that position is in error. I also argued, in section two, that disagreement in general has no bearing on the probability of error—that in order for disagreement to indicate the likelihood of error, particular qualities belonging to a given disagreement or class of disagreements must indicate its likelihood. Consequently, peerhood cannot be the default for disagreement, as I argued in section three. As long as disagreement is going to have any consequence involving diminished confidence in one's own position, it will be necessary for some characteristic of the disagreement to contribute information about the likelihood of an error on one's own part. If the characteristic in question is not the peerhood of one's opponent, it must be some other qualifying characteristic.

That disagreement cannot call for diminished confidence until investigation has shown that one's opponent is the kind whose disagreement can indicate even odds of error in either position (or until something else has contributed comparatively relevant significance) gives rise to a surprising corollary. The necessity of assessment also means that the significance of disagreement which calls for diminished confidence cannot be the most basic significance of disagreement. If the significance of disagreement is such as would call for diminished confidence, then that significance does not belong to all disagreements by default. This raises the question: what significance does disagreement have already? What calls for the investigation to determine when disagreement calls for diminished confidence, and when it does not.

The assessment necessary for any diminishment of confidence must itself be called for. Why would we ever assess the epistemic ability of our opponent in a disagreement? If disagreement is without significance until assessment has been carried out, then why would anyone ever take the trouble to carry out an assessment? It is only when something grabs our attention that we pay it any mind, let alone investigate it. Since it is taken for granted by everyone that assessing those with whom one is in disagreement makes sense, it must be the case that something about being in disagreement with another grabs our attention. In other words—disagreement is significant.

Evaluation is only rational in case there is a more fundamental significance of disagreement that is not dependent upon peerhood. There must be a more basic significance, belonging to any disagreement whatsoever, that motivates the evaluation of opponents in the first place. A basic significance of disagreement must provide the impetus for the curiosity that drives investigation into the matter.

If, as social epistemologists claim, disagreement with an opponent of a certain character (an expert or a peer) can possess altogether different significance, that significance can only be attached to the disagreement after an assessment of the opponent's relative epistemic fitness has been undertaken; an activity which no one would have any reason to engage in if there was not already some significance which called for it. Paradoxical though it may seem, the significance of peer disagreement—that one should be concerned about being in error oneself—is only possible because of (and as a correction to) the prior significance that one's *opposition* is in error.

Does disagreement possess this more fundamental kind of significance? Does it possess a significance which is relevant to knowledge of the object of the dispute in question, yet which does not rely on evaluation of one's opponents or, indeed, which does not hinge on *any* further particular qualities of that dispute? Does disagreement possess a kind of significance just in virtue of being disagreement? If so, what would that significance be?

Epistemology of disagreement has treated disagreement which is not qualified according to the peer condition as insignificant because it has not recognised any quality that belongs to disagreement in general which differentiates it from the merely possible disagreement which could attach to any belief. As long as actual unqualified disagreement is not distinguished from merely possible disagreement, or the mere possibility of there being a view that conflicts with your own, there is nothing about it to attract your attention. Only if the fact that someone happens to champion one of the indefinite number of possible positions that contradict one of your own can be differentiated from the mere possibility that such a position might be held, can an epistemic significance which belongs to disagreement as such be discerned. Only if it is shown that the significance of the fact that there is a person confronting you with an opposing point of view cannot be reduced to the significance of the reasons that might be produced to support that view can a general significance of disagreement be said to exist.

There is such a significance. Actual disagreement has a significance that is not reducible either to the significance of a possible disagreement and the reasons in favour of it, nor to the merely heuristic function that a representative of a possible view might have, such as providing formulations and arguments that make the view easier to consider (and therefore to refute). The difference is this: when I consider a merely possible objection to my belief, I am not compelled to adopt the further belief that another person is wrong, but when another person disagrees with my belief, I am compelled to accept that there is a person who is wrong as a corollary to the fact that our beliefs are in conflict with one another

This is what disagreement introduces: my knowledge that a person disagrees with my belief signifies *their wrongness.* The being in error of another person is signified only by actual disagreement, not by merely possible disagreement, nor by the mere existence of opposing views.

## 5. Why Intersubjective Significance Matters

It is fair to call this significance obvious, but it is not trivial. A circumstance in which you hold a belief that you know someone might possibly take issue with and a circumstance in which that belief brings you into actual conflict with another person are not equivalent. In the case of actual disagreement, your position commits you not only to the belief in that position itself but also to the belief that your opponent is wrong, while in cases of merely possible disagreement your views do not commit you to any similar judgments about another person's epistemic state.

As long as your position was justified, the same reasons that support it should have no trouble supporting the weight of its corollaries, so the indication that another person is wrong should not be an occasion for self-doubt. This means that the basic significance I am drawing attention to is not a roundabout way of getting at an indication of likely error, and if the wrongness of another person is indicated, that is not meaningful solely because it indicates that you might be wrong instead.

This raises the question of what difference this significance makes. If disagreement signifies to you that your opponent is wrong, what does that matter? To be specific, what is it about your opponent being wrong that would call for assessment of the epistemic fitness of your opponent, or any other further discussion or investigation? This is the question that brought us here: what calls for assessment of opponents? However, it is not immediately obvious why assessment would be called for by the indication that your opponent was wrong. On the contrary, it seems that this significance would preempt any call for assessment; does it not indicate, on the face of it, that your interlocutor is not your peer, since it indicates that they are wrong?

The basic, intersubjective significance of disagreement is not merely an indirect route to the self-reflexive significance that you should be concerned about the chance that you could be wrong. The wrongness of the other person signifies something about the other person, and that is where its meaning lies. This is to say that the basic significance of disagreement is irreducibly intersubjective. It is the epistemic failure of the other, and not something about my own epistemic state, that disagreement signifies. Further epistemic activity is called for directly by the indication that the other is wrong, and it is called for even if nothing in the intersubjective significance of disagreement ever leads to self-reflexive significance. I will explain why.

Disagreement *indicates* that the other person is wrong, it does not demonstrate it conclusively. It suggests something in the way that evidence that is not conclusive evidence suggests. It may be univocal, but without being conclusive. As a suggestion, it does not settle the question. Rather, it raises it. It does not demonstrate conclusively that the other is absolutely wrong in their belief, thereby rendering any further inquiry superfluous. Instead, it raises the question of the other's wrongness precisely by indicating that that is the case. It makes perfect sense, then, that this would call for further investigation of the other person, including assessment of their epistemic ability, as long as uncertainty remains about whether what is indicated by the disagreement is true.

## Conclusion

Despite the interest in the epistemic significance of disagreement in social epistemology, there has not been any attention paid to what the significance of disagreement is. Neglect of this question has led to two problems. For one, it has led to an impossible notion of default peerhood which ignores the necessity (belonging to the very concept of peerhood) that peerhood must be conferred upon assessment. More seriously, it has preserved an excessive narrowness in the way that epistemology of disagreement considers the epistemological meaning of disagreement. This narrowness is by design, and it has beneficially excluded contextual sociological and psychological questions from the narrow scope of the question of what, if anything, is learned about the disputandum from the fact that there is controversy surrounding it.

However, the narrowing of the scope of the question goes too far in considering the epistemic significance of disagreement solely in terms of how and under what circumstances it impinges on your own ability to be reasonably certain. I have argued that, considered strictly epistemologically, the beliefs of others and the correctness of those beliefs are relevant to the epistemic circumstance of the disputandum itself, which is to say to the narrowly determined question of disagreement at issue in social epistemology. It is because of what disagreement tells us about the beliefs of others that it can tell us about the disputandum, in which our primary interest lies. This means that, in order to understand the epistemic significance of disagreement, we cannot exclude what it tells us about the intersubjective context in which our own beliefs exist.[26]

[26] Acknowledgements: This paper was presented at the Philosophy Colloquium at the American University of Beirut. Thanks to Tim Carter for comments and copy editing, and to an anonymous referee for comments.

# BRIDGING THE INTELLECTUALIST DIVIDE: A READING OF STANLEY'S RYLE

Jesús NAVARRO

ABSTRACT: Gilbert Ryle famously denied that knowledge-how is a species of knowledge-that, a thesis that has been contested by so-called "intellectualists." I begin by proposing a rearrangement of some of the concepts of this debate, and then I focus on Jason Stanley's reading of Ryle's position. I show that Ryle has been seriously misconstrued in this discussion, and then revise Ryle's original arguments in order to show that the confrontation between intellectualists and anti-intellectualists may not be as insurmountable as it seems, at least in the case of Stanley, given that both contenders are motivated by their discontent with a conception of intelligent performances as the effect of intellectual hidden powers detached from practice.

KEYWORDS: knowing-how, intelligence, intellectualism, dispositions, behaviourism

## 1. The Debate About Know-How: What Is at Stake?

We have assisted in recent years to a live debate about the nature of knowledge-how, but it seems to be difficult to identify what is exactly under dispute in it. The origin of the debate are two famous texts by Gilbert Ryle,[1] where he famously defended that knowing-how is not knowing-that, complaining about the intellectualist slant manifested by those who tried to reduce the former to the latter.

> Philosophers have not done justice to the distinction which is quite familiar to all of us between knowing that something is the case and knowing how to do things. In their theories of knowledge they concentrate on the discovery of truths or facts, and they either ignore the discovery of ways and methods of doing things or else they try to reduce it to the discovery of facts. They assume that intelligence equates with the contemplation of propositions and is exhausted in its contemplations.[2]

Attempting to manifest the shortcomings of intellectualism, Ryle would have shown that agents do not know how to $\varphi$ when they have grasped some

[1] "Knowing How and Knowing That," *Proceedings of the Aristotelian Society* 46 (1946): 1-16; *The Concept of Mind* (London: Routledge, 2009, first published in 1949).

[2] Ryle, "Knowing How and Knowing That," 4.

truths about the practice of $\varphi$-ing, but when they have the power to $\varphi$, the skill to $\varphi$ well, the ability to achieve success in $\varphi$-ing in the relevant circumstances, etc—all of which are issues related to what the agent is able to do, and not to what propositional attitudes she endorses.

Ryle's views became a kind of general consensus, which was underwritten by Jason Stanley and Timothy Williamson's defence of intellectualism in a paper that proved to be as unexpected as influential.[3] According to Stanley and Williamson, there is no fundamental distinction between knowledge-how and knowledge-that, given that the former is, in their opinion, a species of the latter—a view that they defended with the help of much apparently solid linguistic evidence.[4]

In the last years different positions have been proposed on one side or the other of the intellectualist divide, mostly arising from development or criticism of Stanley and Williamson's original proposal.[5] However, in my view, the terms of

[3] Jason Stanley and Timothy Williamson, "Knowing How," *The Journal of Philosophy* 98, 8 (2001): 411-444.

[4] According to Stanley and Williamson, knowing-how would have exactly the same syntactic structure as knowing-what, knowing-when, or knowing-why, all of which are just a matter of knowing facts, and thus are cases of propositional knowledge. The case of knowing-how would be quite a *sui generis* variety of knowing-that: one where the agent knows *de se* that there is a *way* for her to perform the action in question, a way that she must grasp under a practical mode of presentation. At the very same time that Stanley and Williamson's seminal paper came out, Jesús Vega was problematizing the Rylean idea of "practical understanding" and showing that it needed a better articulation with propositional knowledge, mediated by experience and practice. See his "Reglas, medios, habilidades. Debates en torno al análisis de «S sabe cómo hacer X»," *Crítica* 33, 98 (2001): 3-40.

[5] Several authors have followed their way, defending different varieties of intellectualism. For instance: Paul Snowdon, "Knowing How and Knowing That: A Distinction Reconsidered," *Proceedings of The Aristotelian Society* 105, 1 (2004): 1-29; John Bengson and Marc A. Moffett, "Know-How and Concept Possession," *Philosophical Studies* 136, 1 (2007): 31-57 and "Two Conceptions of Mind and Action. Knowing How and the Philosophical Theory of Intelligence," in *Knowing How: Essays on Knowledge, Mind, and Action*, edited by John Bengson and Marc A. Moffett (New York: Oxford University Press, 2011), 3-55; Berit Brogaard, "Knowledge-How: A Unified Account," in *Knowing How*, edited by Bengson and Moffett, 136-160; Yuri Cath, "Revisionary intellectualism and Gettier," *Philosophical Studies* 172, 1 (2015): 7-2; Carlotta Pavese, "Know-How and Gradability," *Philosophical Review* 126, 3 (2017): 345-383. Stanley and Williamson themselves come back to the issue in "Skill," *Noûs* 51, 4 (2017): 713-726. On the opposite side of the divide, the moves have been, with some exceptions, more aggressive or defensive than constructive, by which I mean that different attacks have been levelled at the new intellectualist arguments, but little has been done yet in order to build up a full-blown positive epistemological alternative. Amongst the most recent ones are Ellen Fridland, "Problems

the debate are far from being unanimously established, as well as its proper object. Different positions in the philosophy of mind, philosophy of language and metaphysics interfere with the strictly epistemological problem, making it hard to figure out what the genuine bone of contention is. For that reason, I would like to stipulate for the sake of this paper some basic terminology, differentiating three levels under dispute: epistemological concerns, pre-conceptual assumptions and metaphysical theories.

First, I will use the term "intellectualism" to label a very specific epistemic position about the nature of knowledge-how:

> INTELLECTUALISM: Knowing how to *φ* is knowing that *p* is the case.[6]

This epistemic thesis (i.e. a claim about what that particular kind of knowledge is) was the focus of Gilbert Ryle's criticism in the aforementioned papers, both entitled "Knowing How and Knowing That." Given that Ryle was opposing INTELLECTUALISM, it is not weird that his own opinion was latter labelled as "anti-intellectualism," but I find this utterly misleading because, strictly speaking, "anti-intellectualism" is no positive thesis, but just the denial of INTELLECTUALISM. At least, that is the sense that I will give to the term here:

> ANTI-INTELLECTUALISM: Knowing how to *φ* is not knowing that *p* is the case.

Notice that ANTI-INTELLECTUALISM, unlike INTELLECTUALISM, just says what knowledge-how is *not*, not making any positive claim about what it

with intellectualism," *Philosophical Studies* 165, 3 (2013): 879-891; "Knowing-How: Problems and Considerations," *European Journal of Philosophy* 23, 3 (2015): 703–727; J. Adam Carter and Duncan Pritchard, "Knowledge-How and Cognitive Achievement," *Philosophy and Phenomenological Research* 91, 1 (2015): 181–199; J. Adam Carter and Bolesaw Czarnecki, "(ANTI)-Anti-Intellectualism and the Sufficiency Thesis," *Pacific Philosophical Quarterly* 98, 1 (2017): 374-97; William Hasselberger, "Propositional Attitudes and Embodied Skills in the Philosophy of Action," *European Journal of Philosophy* 26, 1 (2018): 449-76; Carter, J. Adam, y Jesús Navarro, "The Defeasibility of knowledge-how," *Philosophy and Phenomenological Research* 95, 3 (2017): 662–685. See J. Adam Carter and Ted Poston, *A Critical Introduction to Knowledge How* (London: Bloomsbury Publishing, 2018) for a critical overview. Finally, a more constructive attitude on the anti-intellectualist side may be found in David Löwenstein, *Know-how as Competence: a Rylean Responsibilist Account* (Frankfurt am Main: Vittorio Klostermann: 2017).

[6] Nowadays, intellectualism is sometimes defined as the much weaker claim that knowing how is *at least partially grounded* in some propositional attitude—see for instance Bengson and Moffett "Two Conceptions," 7. Nevertheless, for reasons that will be explained, I find it disputable that such a weak thesis was the target of Ryle's original attacks, as Bengson and Moffett themselves seem to assume (*Ibid*, 9 note 11).

actually is.[7] Denying a concrete response to one question does not imply the acceptance of any other particular positive answer to that same question. Ryle might have had a positive thesis on the nature of knowledge-how—what I will later call "Ryleanism"—but I find it utterly misleading to label his alleged positive view as "anti-intellectualism," since there may be other positive views on the nature of knowledge-how, besides Ryleanism, that would share the negative point that it is not a species of knowledge-that.

In any case, Ryle's epistemological focus in these papers was framed in a wider philosophical project, whose aim was beyond epistemology (or beneath it). He aimed to impeach a general pre-conceptual understanding of the mind that, according to Ryle, was dominant at his time, which he labels in different ways: "the prevailing doctrine," "the official theory," "the intellectualist legend"… a way of thinking that he finds somehow related to INTELLECTUALISM in epistemology. Many of his readers have found it annoying that Ryle does not take any particular author as his enemy, constructing a mysterious "legend" as a kind of straw man that nobody actually ever defended. This accusation is unfair, given that contenders of Ryle were flesh and blood authors,[8] but I still believe that there is an explanation for the uneasiness that Ryle produces in his many of readers by being so reluctant to discuss particular theories. The reason for this is that Ryle was not attacking any explicit theoretical view, either in the field of metaphysics or the philosophy of mind, but a kind of unarticulated and pre-conceptual assumption beneaththeoretical activity in those fields. A kind of implicit presupposition that had become a piece of common sense—at least common in the limited academic community. I will articulate that intuition in the following terms:

> HIDDEN: intelligence is not something that may be directly observable in the agent's behaviour, but only predicated of it in so far as it is a manifestation of some hidden state or process, which is not itself observable.

[7] Notice that INTELLECTUALISM, thus defined, is even compatible with the views of some authors that consider themselves nowadays as "intellectualists," such as Bengson and Moffett's (see note 6), in so far as they assume a weaker thesis than the one that strictly *identifies* knowledge-how with knowledge-that.

[8] Michael Kremer has shown there was a real intellectual debate around intellectualism before the Second World War, with authors who actually held positions very similar to the ones contested by Ryle. See his "Ryle's 'Intellectualist Legend' in Historical Context," *Journal for the History of Analytical Philosophy* 5, 5 (2017): 16-39. However, as Kremer convincingly shows, what Ryle was attempting to undermine was the common assumption behind that debate, which shows why his own view ought not be simply understood as the denial of intellectualism. See also Will Small, "Ryle on the Explanatory Role of Knowledge How," *Journal for the History of Analytical Philosophy* 5, 5 (2017): 56-76.

Realising that HIDDEN is not some author's thesis, hypothesis or theory is of utmost importance.[9] It does not work as a positive statement that could be made explicit and defended by solid arguments, but as a kind of pre-theoretical assumption that motivates a certain direction in the inquiry about the mental, shaping what any valuable answer may look like.[10] HIDDEN is what commits any explicit philosophical conception of the mind to explain why mental epithets and, in particular, those related to 'intelligence,' may be predicated of people's actions, given that it is not the sort of thing that we could ever *see* in their behaviour. Those assuming HIDDEN are committed to the task of explaining what it is that makes behavioural patterns intelligent—*viz.* in virtue of which kind of inner and hidden processes, unobservable by others, occurring in each agent's private 'grotto,' is their behaviour intelligent.

HIDDEN, in and by itself, is no metaphysical claim—although it could certainly favour some metaphysical views over others. Assuming HIDDEN as an implicit starting point,authors might defenddualist, materialist, functionalistor emergentist views about the nature of the mind, just to mention a few possibilities, because HIDDEN says nothing about the nature of the alleged hidden processes where intrinsic intelligence is supposed to be located, or about the sort of connections that such process have with those occurrences that we actually see. HIDDEN just invites us to look for the mental somewhere else—as opposed toin—what we actually see in behaviour.

One may think that, in contrast to HIDDEN, INTELLECTUALISM is a positive metaphysical statement. But strictly speaking it is not, since it says nothing about the nature of the mind or its processes either, or about the way it deals with propositional contents, or about the kind of relation (causal, functional, explanatory…) that the mind has with those performances that we observe. Unlike HIDDEN, INTELLECTUALISM is a theoretical thesis, but not one that belongs to metaphysics, or to the philosophy of mind, but to the theory of knowledge.

[9] Will Small holds that "The central target (...) of Ryle's discussion in the second, third, and fourth chapters of (*The Concept of Mind*) taken together is the view that to credit some piece of behaviour with displaying qualities of mind we must appeal to inner mental causes of it. I will call this general view causalism" ("Ryle on the Explanatory Role," 59). I would not disagree with Small's exegetical point in those specific texts, but I believe it is important to realise that causalism too is just a case that exemplifies the general pattern that is Ryle's target.

[10] This is the reason why I prefer the label HIDDEN to the one Hasselberger uses for a very similar view, namely "Neo-Carthesian presupposition" ("Propositional Attitudes," 15).

## 2. How Not To Introduce the Debate

My intention while introducing those terminological distinctions is not so much exegetical as instrumental. I do not hold that these are the exact definitions that Ryle had in mind, but that distinguishing the terms in this way will prove beneficial while we approach the theoretical arena.

The predominant strategy that anti-intellectualists have adopted until quite recently has been to defend Ryle against the attacks of the new trend of intellectualism that stems from Stanley and Williamson, either by modifying Ryle's position or by showing that the arguments levelled against it are not solid. My proposal here is to adopt a different strategy, in the wake of what may be considered as a new wave in the anti-intellectualist party: namely, to show that Ryle's views have been seriously misconstrued in this debate.[11] In this sense, I would like to defend Ryle but, most importantly, *not* Stanley and Williamson's Ryle, which in my view is a misconstruction that results, as I will show, from a slanted reading of his work. I will focus in particular on Jason Stanley's later developments of intellectualism[12] with a double intention: first, to put forward a better understanding of Ryle's views resulting from a more charitable reading of his work; and second, to show that, surprisingly enough (at least for me!), this different reading paves the way for a possible understanding between Stanley and what I take to be the original Ryle. Preparing the ground for such understanding is the final goal of this paper, and what explains its title.

Before reaching Stanley, I will stop for a moment to consider the way John Bengson and Mark A. Moffett introduce the debate on knowledge how in their conscientious introduction to the volume they co-edited on the topic, which I find paradigmatic of an unfortunate approach that confuses the different levels that I tried to separate in the previous section:

> Intelligence-epithets often modify overt behaviours, such as pruning trees. But Ryle is keenly aware that Intelligent actions, such as pruning trees skilfully, are not distinguishable from non-Intelligent actions in virtue of any *overt* features of the performance; rather, we must "look beyond the performance itself."[13]

In my view, Bengson's and Moffett introduction to the debate dooms it to degenerate into a sort of dispute about which is the better way to respond to

[11] Dissatisfaction with respect to Stanley and Williamson's reading of Ryle has been a part of the debate since the beginning, but a milestone in this respect is the monographic issue edited by Julia Tanney in the *Journal for the History of Analytical Philosophy* (2017).

[12] In particular, *Know How* (Oxford: Oxford University Press, 2011) and "Knowing (How)," *Noûs* 45, 2 (2011): 207-238.

[13] "Two Conceptions," 6.

HIDDEN, taking for granted that such intuition must somehow or another be satisfied by any theory we may seriously consider. Bengson and Moffett are quoting Ryle here, but they make sense of his "beyond" in a way that forces him to search for intelligence *elsewhere* when, perhaps, it could be there, at sight, in behaviour itself, in the light of the possibilities and eventualities that it makes manifest. It is right that this "elsewhere" does not have to be inherently mysterious, or essentially inaccessible, but still in Bengson and Moffett's reading it could not simply *be there*, at sight. The possibility of holding that intelligence is *in the act itself* seems to be a non-starter from Bengson and Moffett's perspective. That is why the different positions in the debate show up in their description of the scene as alternative ways to account for one structurally similar intuition:

> The core contention of the intellectualist side of this line is that states of Intelligence and exercises thereof are at least partially grounded in propositional attitudes. The core contention of the anti-intellectualist side, by contrast is that states of Intelligence and exercises thereof are grounded in powers (abilities or dispositions to behavior), not in propositional attitudes.[14]

Unfortunately, HIDDEN appears here as the common ground where all the contenders must find their own place, Ryle included, who is presented as the one who defends the view that the invisible place where we have to look for intelligence is in the agent's abilities or dispositions:

> Whereas anti-intellectualism allows that we detect *abilities* or *dispositions* in virtue of witnessing actual performances (in diverse circumstances, on multiple occasions, etc.), intellectualism allows that we detect *attitudes* in virtue of witnessing such performances. Either way, we manage to "look beyond the performance itself" to a power (ability, disposition) or intellectual state (attitude) of the individual that is distinct from any particular overt behaviour.[15]

Bengsonand Moffett's apparently balanced formulation is highly problematic on closer inspection because INTELLECTUALISM and ANTI-INTELLECTUALISM are not structurally similar hypotheses—a point that will take some unpacking.

Given that knowledge-that involves a psychological attitude towards some propositional content, defendants of INTELLECTUALISM have to claim that knowledge-how is also constituted by such propositional attitude. That is why INTELLECTUALISM is in natural accordance with HIDDEN: it would give an answer to the question for the 'elsewhere' intelligence stems from: the agent's propositional attitudes. But Ryle's position ought not be introduced by the same sort of argument just by substituting "psychological or propositional attitude" for

[14] *Ibid*, 18.

[15] *Ibid*, 30.

"ability," "disposition" or "power." Otherwise, we could not construe him but as another positive attempt to satisfy HIDDEN.

Quoting in length the passage cited by Bengson and Moffett will help realise the infelicity of their presentation of the different views under dispute:

> In judging that someone's performance is or is not intelligent, we have, as has been said, in a certain manner to look beyond the performance itself. For there is no particular overt or inner performance which could not have been accidentally or 'mechanically' executed by an idiot, a sleepwalker, a man in panic, absence of mind or delirium or even, sometimes, by a parrot. But in looking beyond the performance itself, we are not trying to pry into some hidden counterpart performance enacted on the supposed secret stage of the agent's inner life. We are considering his abilities and propensities of which this performance was an actualisation.[16]

We cannot express Ryle's views as the claim that the mysterious*something*we must be looking for is the ability, the capacity, the power or the disposition,which may not be directly observable, and must be somewhere hidden in the agent, making it the case that her behaviour manifests intelligence. However, from Bengson and Moffett's point of view, all contenders would agree on the idea that what makes a performance intelligent is some additional feature that can only be conjectured, hypothesized, or just indirectly inferred, which is precisely the very idea that Ryle intended to criticise. Bengsonand Moffett's introduction to the debate is thus committing all contenders to respond to the intuition of HIDDEN, searching for the place where intelligence *really* happens, given that in principle it cannot be out there, at sight.

## 3. Stanley's Ryle

I will now focus on Stanley's 2011 pieces (*Know How* and "Knowing (How)"), which are developments of the view he put forward with Timothy Williamson in their 2001 paper. In section one we have seen that Ryle's original criticism of INTELLECTUALISM was motivated by the fact that that epistemological thesis is somehow in accordance with the kind of unfortunate pre-theoretical intuition that I have labelled HIDDEN. At this point, it seems pertinent to ask whether Stanley's defence of INTELLECTUALISM may also be considered as being in accordance with HIDDEN. The answer to this question will be crucial to take a stand on Stanley's understanding of the Rylean project. The question then is: is Stanley's aim to defend a notion of intelligence that confines it to the privacy of the mind

[16] Ryle, *The Concept of Mind*, 33.

(HIDDEN) when he claims that knowledge-how is a species of knowledge-that (INTELLECTUALISM)?

I believe not: a careful reading of his proposal shows that Stanley's INTELLECTUALISM is not a defence of HIDDEN, but a different attempt to escape from it—whether a successful one or not is something that remains to be elucidated. Unfortunately, Stanley does not elaborate on this point, and his position regarding the kind of intuitions I have phrased as HIDDEN remains obscure. Instead of positioning himself explicitly for or against them, he focuses on the epistemological claim of INTELLECTUALISM, raising a direct confrontation with Ryle that, as I will show, loses track of what was originally at stake in his proposal. Had Stanley directly discussed Ryle's main goal, he would probably have found that the kind of position he himself is championing has much in common with Ryle's original project. However, instead of pursuing this line of thought, he reads Ryle in a way that, from the outset, seems to be far from charitable, discrediting him for holding old fashioned views that "No one thinks anymore," and are "now universally rejected."[17] The result is a reading that some authors have found highly disputable.[18]

I will summarise Stanley's reading of Ryle in six points, all of which I find mistaken. According to Stanley, Gilbert Ryle is:

(1) Unclear about his own positive position.

(2) A verificationist on meaning.

(3) A behaviourist on the nature of the mind.

(4) A fictionalist on mental states and processes.

(5) A 'preachivist' on knowledge-that.

(6) A 'distinctivist' on the relationship between action and theory.

The nature of claims (1) to (4) is crucially different from the one of (5) and (6). The former group, which I will analyse in sections four and five, are explicit attacks that Stanley directs towards Ryle, in the sense that Stanley is aware that

[17] Stanley, *Know How*, 7.

[18] For instance, Stephen Hetherington, "Knowledge and Knowing: Ability and Manifestation," in *Conceptions of knowledge*, edited by Stefan Tolksdorf (Berlin: de Gruyter, 2012), 73-100; Jennifer Hornsby, "Ryle's Knowing-How, and Knowing How to Act," in *Knowing How*, edited by Bengson and Moffett, 80-98. In the same volume, Paul Snowdon's contribution ("Rylean Arguments: Ancient and Modern," 59-79) is also critical, although less strongly. For more recent criticism, see Julia Tanney, "Gilbert Ryle on Propositions, Propositional Attitudes, and Theoretical Knowledge," *Journal for the History of Analytical Philosophy* 5, 5 (2017), and both Kremer's and Small's contribution to that volume.

Ryle rejects those accusations. In contrast, (5) and (6) are not accusations, but attempts to objectively paraphrase Ryle's views in ways Ryle would allegedly consider valid, according to Stanley. I will hold that the fact that Stanley sees those two last theses as faithful summaries of Ryle's views is still more pernicious than the fact that he makes the precedent unfair accusations, because it shows that he is missing the core of his opponent's position. That is the reason why, in section six, I will analyse in depth those two later points, contrasting them with a reconstruction of Ryle's original arguments.

## 4. Lack of Clarity

Even if, at first glance, Ryle's style might be the most crystalline one a philosopher might have ever achieved, the complaint that his positive position on the nature of knowledge-how—what I have called "Ryleanism"—is unclear is quite widespread, even among those who are willing to follow his lead. He did say, quite indisputably, that, when attributing knowing how, we are normally talking about people's abilities and capacities—*viz.* what they are able to do, their powers—, and not about the intellectual truths that they have grasped. And he did say that knowing how "is a disposition, but not a single-track disposition like a reflex or habit."[19] But would Ryle defend a strict reduction of knowing how to those abilities, powers and dispositions? The answer is anything but clear. Some authors (mostly intellectualists) identify his view with a sometimes called "ability thesis," whereas others (mostly anti-intellectualists) deny that such a simple view was ever held by Ryle, or at least find the idea disputable.[20]

To make things worse, not only Ryle's positive views on the nature of knowledge-how is enigmatic, but also his positive views about the relationship between knowledge-how and knowledge-that—i.e., his response to what Kremer calls the challenge of "accounting for the unity of knowledge."[21] In this sense, Ryle may be interpreted in at least three possible ways: practicalism, unitarianism and pluralism. First, he may be read as not just attacking INTELLECTUALISM, but as

[19] *The Concept of Mind*, 34.

[20] See for instance Jeremy Fantl, "Knowing-How and Knowing-That," *Philosophy Compass* 3 3 (2008): 455 and n10; Jennifer Hornsby, "Ryle's Knowing How," 82; Benjamin Elzinga, "Self-Regulation and Knowledge-How," *Episteme* 15, 1 (2018): 119-140; David Löwenstein, *Know How as Competence*, 6. There is some insightful criticism from an intellectualist position in Natalia Waights Hickman, "Knowing in the 'Executive Way': Knowing How, Rules, Methods, Principles and Criteria," *Philosophy and Phenomenological Research* 10.1111/phpr.12488 (2018), 5.

[21] "A Capacity to Get Things Right: Gilbert Ryle on Knowledge," *European Journal of Philosophy* 25, 1 (2016): 25.

defending the opposite thesis, attempting to reduce knowledge-that to a species of knowledge-how—a view that is sometimes called "strong anti-intellectualism" or "practicalism."[22] It is hard to deny that, at some points, Ryle seems to be quite akin to this idea, for instance, when he explicitly claims that knowledge-how is logically prior to knowledge-that,[23] or when he holds that knowledge-that presupposes knowledge-how as its precedent (because one may only know a truth if one is able to previously perform actions that amount to knowledge-how, or because one may only count as knowing that such and such is the case if one also knows how to give good reasons to hold it).[24] And, on top of that, Ryle holds that understanding is a part of knowing how,[25] which, if right, and given that knowing that *p* appears to require understanding *p*, seems to imply that know-how must be at least a constitutive element of knowledge-that.

A second possible reading of Ryle, unitarianism,would construct his view not as the one that knowing-that may be reduced to knowing-how, but as claiming that both concepts have a common root. Such an interpretation has been put forward by Michael Kremer, who holds that there is a core meaning involved in the different uses of the verb 'knows,' one that covers both knowing-how and knowing-that.[26] According to his interpretation of Ryle, which he bases on views by John Hyman,[27] to know is to have a 'capacity to get things right'. Even if Kremer's reading is compellingly defended, I find it difficult to prevent it from collapsing into some form of intellectualism—as it overtly occurs with an account of knowing how like Hickman's, that shares with Kremer the influence of Hyman. My worry in this respect is that the idea of correctness involved in "getting it right" seems to strongly suggest that truth conditions are somehow grasped by the agent, and thus that all knowledge is some way or another based on representational states.[28]

[22] Fantl labels the view as 'strong anti-intellectualism' ("Knowing How and Knowing That," 452) and by Hetherington as 'practicalism' ("Knowledge and Knowing," 73), but both are careful enough not to attribute it to Ryle.

[23] "Knowing How and Knowing That," 4.

[24] *Ibid*, 9.

[25] *The Concept of Mind*, 41.

[26] "A Capacity to Get Things Right," 28.

[27] "How Knowledge Works," *The Philosophical Quarterly* 49 (197): 433-451.

[28] That is the effect of expressions like "the content of knowledge-how" (Hickman, "Knowing in the 'Executive Way'," 17), which I find shocking, even if conceived as non-conceptual. Instead of with idea of "getting it right," the unitarianist view may perhaps be better defended in terms of achievements or failures, not assuming that the aim is in any sense a correct *representation*. That is: we would need an account of performance assessments that does not rely on how the agent

Finally, a pluralist reading of Ryle would hold that knowing-how and knowing-that are not reducible to each other (as both intellectualists and practicalists hold, in different directions), nor to a third more basic genus (as unitarists hold), but simply different concepts with strong and interesting connections but no common core. David Wiggins, for instance, holds that "Ryle is in a position not merely to allow but also to assert that, in their full distinctness, knowing how to and knowing that need one another."[29] According to such a reading, theoretical knowledge relies on the practical, and practical knowledge rests on the propositional. The problem with this interpretation is that it would still have to show what response Ryle would give to the challenge of accounting for the unity of knowledge. The disparity and irreducibility of those two concepts could be understood as a denial that there is *one* think called knowledge besides that terminological coincidence—a position that may in the end favour the standard tacit assumption that epistemologists ought only to be concerned with 'genuine' knowledge, i.e. of the propositional kind, an unfortunate idea that may be found in virtually all introductions to the field.[30]

By my side, I am reluctant to accept any of these three possibilities because they seem to be involved in a misleading quest for Ryle's original theoretical views, Ryleanism, as a positive epistemological theory of the nature of knowing-how, whereas I would say that this common assumption is what may be challenged. My point in this respect is that, in general, Ryle was not trying to offer any clear-cut explanatory hypotheses of concepts. We may not find in his work, in particular, any reductive analysis of epistemic concepts, in terms of necessary and sufficient conditions, and his approach to knowing-how is no exception.[31] A possible explanation of this is that his philosophical method was not really driven towards

represents the desired outcome, but on the way or manner in which she is able to conduct performances (successfully, with mastery, skilfully…). For hints in this direction, and doubts about the very idea of non-conceptual content, see Daniel D. Hutto, "Unprincipled Engagements. Emotional Experience, Expression and Response," in *Radical Enactivism* (Amsterdam: J. Benjamins Pub. Co., 2006): 13-38.

[29] David Wiggins, "Knowing How to and Knowing That," in *Wittgenstein and Analytic Philosophy: Essays for P. M. S. Hacker*, edited by Hans-Johann Glock and John Hyman (Oxford: Oxford University Press, 2009): 264–5.

[30] If the concepts of knowing-how and knowing-that were finally so irreducible to each other, nor to any common term, that would strongly suggest that the former is in the end of no genuinely *epistemic* concern, being more related to the philosophical study of powers. See Vega, "Reglas, Medios, Habilidades," 7.

[31] Hornsby, "Ryle's Knowing How," 81-2.

theory. Stanley himself recognises this anti-theoretical tendency,[32] but still, as I will show, he recurrently reads Ryle as an author that does puts forward and defend positive views. In contrast, different interpreters hold today—and I would forcefully agree with them—that Ryle's philosophical project was quite a different one, with important resemblances to Wittgensteinian therapy.[33] Such an intellectual project might not seem as trendy today as it once was, at least for those who expect that their philosophical work will have some clear impact on the mainstream development of cognitive sciences—and I cannot think of many more evident examples than Stanley's case.[34] Nevertheless, even if one does not sympathise with the kind of anti-theoretical slant that Ryle manifests, approaching his work with the fundamental aim of reconstructing and objectively evaluating his positive theoretical views may not be the most charitable way to read him.

I do not want to deny that Ryle has a positive view on the topic under discussion, nor do I want to hold that his 'logical geography' is fully deprived of positive theses. Still, even if there were such theses, and even if it were evident today that such theses are wrong, that does not invalidate his achievements with respect to his primary negative and therapeutic goal. And, as Ryle himself avows:

> My argument has been intended to have the predominantly negative point of exhibiting both why it is wrong, and why it is tempting, to postulate mysterious actions and reactions to correspond with certain familiar biographical episodic words.[35]

If we take Ryle's reflections at face value, any reading of his works that were primarily focussed on constructing his alleged own positive position could run the risk of missing his "predominantly negative point"—and, in this respect, it does not matter much if the interpreter is in favour of Ryle or against him.

[32] "Ryle was a committed ordinary language philosopher, unreflectively and immediately hostile to analysis and reduction of any kind." Stanley, "Knowing (How)," 10.

[33] See Julia Tanney, "Rethinking Ryle: A Critical Discussion of *The Concept of Mind*," in Gilbert Ryle: *The Concept of Mind,* 60th Anniversary Edition (Abingdon: Routledge, 2009): xi. Stephen Hetherington discusses the resemblance between Ryle's reflections on know-how and Wittgenstein's ones on rule following in "Knowledge and Knowing," 31. See also: David Löwenstein, "Knowledge-How, Linguistic Intellectualism, and Ryle's Return," in *Conceptions of knowledge*, edited by Stefan Tolksdorf (Berlin: De Gruyter, 2012): 301; Hasselberger, "Propositional Attitudes."

[34] For a critical view on this trend see Max R. Bennett and Peter M. S. Hacker, *Philosophical Foundations of Neuroscience* (Malden, MA: Blackwell, 2003).

[35] *The Concept of Mind*, 135.

## 5. Verificationism and Behaviourism

Accusations of verificationism (2) and behaviourism (3) go hand in hand. The former is a position in the philosophy of language according to which a sentence may only be meaningful in so far as it is verifiable, at least in principle, and the latter is a statement in the philosophy of mind, that would force us to account for all psychological states and processes in terms of behavioural patterns. Those are supposed to be objectively verifiable, in contrast to psychological states and processes themselves, which (unless they are one's own) allegedly depend on rational reconstruction and speculative hypothesis about the unseen. For that reason, behaviourism shows up as a position in psychology and the philosophy of mind that is in accordance with verificationism in semantics and the philosophy of science. These views are usually assumed as handicaps of Ryleanism, given that they are views that did not survive the arrival of functionalist and cognitivist approaches to the mental. Now, Stanley's reading of Ryle does not just identify his position with these views but, furthermore, reads him as systematically producing positive defences of them, for instance, when he claims that:

> *The Concept of Mind* is devoted to advancing Ryle's behaviourist views. It is not immediately evident how the topic of knowing how fits into this now unpopular agenda,[36]

or that:

> Ryle assumes a theory of meaning that connects linguistic meaning to verifiability: a term is meaningful only if it is possible in principle to verify whether or not it applies to something.[37]

Stanley's interpretation of Ryle then assumes his texts as pursuing the basic goal of advancing positive theses, behaviourism on the one hand and verificationism on the other, two positions that would both be motivated by one same epistemic fear of the unknowable.

However, at the same time, Stanley is perfectly aware that both positions are explicitly rejected by Ryle, or at least set aside as unclear and problematic—the former in his papers "Unverifiability-By-Me"[38] and "The Verification Principle,"[39] and the later in different papers complied in *On Thinking*.[40] That is the reason why I call these "accusations," and not simply "restatements" of Ryle's views. The fact

[36] Stanley, "Knowing (How)," 1.

[37] *Idem*, 7.

[38] In Gilbert Ryle, *Collected Papers*, vol. 2 (London: Hutchinson, 1971): 126-236.

[39] *Idem*, 300-306.

[40] Gilbert Ryle, *On Thinking*, edited by Konstantin Kolenda (Oxford: Basil Blackwell, 1979).

that Ryle explicitly rejects those views, or at least holds that they require much qualification, is certainly not determinant, since he could be a verificationist and a behaviourist *malgré lui*. And it is understandable indeed that both views could seem appealing to somebody pursuing Ryle's project in so far as, if those two theses were correct, we would have excellent reasons to definitely reject HIDDEN. Verificationism and behaviourism bring to the foreground *everything* that is allegedly beyond the performance itself, and *always*—but at too high a price. Ryle is not forced to endorse such radical views in order to hit his target with respect to HIDDEN. If those principles were correct, they would prove that *all* intelligence is out there, and that *all* the mental is at sight for external observers—but Ryle's target requires much less than that. It would be enough for his purposes to show that *some* acts of intelligence may be there, at sight, and that *some* mental attributions are not hypotheses on what happens in the agent's secret grotto, but something that we may actually see in what she is doing. In other words: Ryle does not need the sledge-hammers of verificationism or behaviourism in order to crack the nut of HIDDEN.

Much more could be said about this, but it will suffice to have shown, first, that Ryle puts both behaviourism and verificationism under critical assessment, not being committed to any of them; and, second, that those views seem to be much stronger than the ones he would require to achieve his goal of undermining HIDDEN. It is not clear why Stanley insists so much on Ryle's arguments having these today unfashionable burdens.

## 6. Fictionalism

This point brings us to accusation (4), according to which Ryle can only have a fictionalist account of mental processes.[41] Fictionalism is a view according to which our talk about mental states and processes is nothing but a *façon de parler*, which does not aim at literal truth. Our attributions of beliefs, desires or intentions would not token any real events occurring in the world, and would thus be mere fictions. Once again, such interpretation is an accusation in clear contradiction with what Ryle claims about his own position. It is hard to deny that he *does* explicitly affirm the existence of mental processes occurring in the privacy of the agent's mind, something that we, external observers, cannot see from the outside: as Stanley recognises, silent soliloquies, mental imagery and acts of remembering are present all over his texts as real occurring events.[42] Ryle never denies the existence of

[41] Stanley, "Knowing (How)," 9.

[42] In this respect, see Brian Weatherson, "Doing Philosophy With Words," *Philosophical Studies* 135, 3 (2007): 429-437; Eric Schwitzgebel, "Gilbert Ryle's Secret Grotto," in *The Splintered Mind*

private mental processes of that kind. What he denies is their *essentially* private nature, the idea that those processes are something that, in principle, could *never* happen on the outside, at sight of others. Those processes, he claims, may happen in the agent's privacy, but they could have occurred in the public scene just the same. And, most importantly, when they do happen in the public scene, they do not denote intelligence because there is something simultaneously occurring behind the scene, something that makes them be truthful hallmarks of intelligence: the occurrence of intelligent behaviour at sight is not a secondary manifestation of what is primarily happening in the privacy of the agent's mind.

The accusation of fictionalism is related in Stanley's interpretation of Ryle to the attribution of another opinion that seems untenable:

> On Ryle's picture of action, intentional actions are not the effects of inner categorical causes. Thus, his picture of knowing how coheres with his conception of intentional action. Ryle's metaphysical picture is widely regarded as implausible, since it involves ungrounded dispositions—that is, the possession of dispositions without any categorical basis.[43]

Stanley is probably identifying Ryle with a variety of the anti-causalist account of rational action, i.e. the idea that *reasons are not causes*, championed by authors like Wittgenstein and Anscombe—a view that Small has recently linked to Ryle's work[44]—, but this view ought not be confused with the blunt idea that intentional actions have simply *no* causes. It may be perfectly defended that rational explanations are not causal explanations without being committed to the much more contentious view that intentional actions have no causal explanation. Anti-causalists may accept that there *is* some causal explanation for every intentional action, but still hold that elucidating such cause is not what rational explanations aim at because such actions are somehow intrinsically normative.[45]

As Small has shown, Ryle holds that intelligence attributions are dispositional but, at the very same time, he is very careful not to identify them with the sort of dispositions that may be reduced mechanical or merely causal explanations, either internal or external, since he does not purport at all to explain

(blog), June 15, 2007, http://schwitzsplinters.blogspot.com/2007/06/gilbert-ryles-secret-grotto.html.

[43] Stanley, *Know How*, 17.

[44] Small, "Ryle on the Explanatory Role," 5.

[45] "Our inquiry is not into causes (and *a fortiori* not into occult causes), but into capacities, skills, habits, liabilities and bents" (Ryle, *The Concept of Mind*, 33). For a defence of the intrinsic normativity of these concepts see Löwenstein, *Know How as Competence*, 13.

knowing-how in terms of pure habits or automatic manifestations.[46] Ryle's dispositionalism would be an attempt to escape such reduction of prudence and intelligence to causal explanations that are blind and mechanical in kind. In order to defend this, he needs to show that at least *some* of our mentalist vocabulary (knowing how attributions, epithets of intelligence or prudence and, in general, all the rich vocabulary that we employ to describe human performances) is not based on causal hypotheses, but that its meaning stems from the way we use this jargon at the personal level. Such vocabulary opens up a logical space where certain kinds of rational assessment and criticism becomes appropriate,[47] expectations of success are backed by some expectations of warrant[48] or control,[49] and new concerns relating responsibility and resilience arise.[50] The impossibility to reduce such explanations to mechanical causes is not a deficit in the explanation itself, but the defining feature of the kind of "imponderable evidence" that constitutes our knowledge of human beings—what since Wittgenstein is known as *Menshenkenntnis.*[51]

## 7. A Reconstruction of Ryle's Argument

In contrast to the former ones, the remaining two theses, (5) and (6), are presented by Stanley as objective restatements of Ryle's positions. That is, according to Stanley, those are views Ryle would be glad to endorse:

> PREACHIVISM: acting on some piece of knowledge-that requires an act of contemplating the proposition in question: an occurring mental process of 'preaching' by which the proposition is considered as a reason for action.
>
> DISTINCTIVISM: what guides us in action is a distinct cognitive capacity from what guides us in reflection.

In my view, Ryle does not endorse any of these views, which means that Stanley would not just have levelled some unfair accusations—(1) to (4)—, but furthermore he would have misidentified Ryle's own position. In order to show the reason of the misunderstanding I will have to reconstruct Ryle's argumentative strategy with some detail.

[46] Small, "Ryle on the Explanatory Role," 74.

[47] Stina Bäckström and Martin Gustafsson, "Skill, Drill, and Intelligent Performance: Ryle and Intellectualism," *Journal for the History of Analytical Philosophy* 5, 5 (2017): 41-55.

[48] Katherine Hawley, "Knowing How and Epistemic Injustice," in *Knowing How: Essays on Knowledge, Mind, and Action*, edited by Bengson and Moffett, 28.

[49] Löwenstein, *Know How as Competence*, 107.

[50] Benjamin Elzinga, "Self-Regulation and Knowledge How," 121.

[51] *Philosophical Investigations* (Oxford: Wiley-Blackwell, 2009) §§358-360.

To be fair, behind the appearances, Ryle's argument is anything but simple. In order to show that INTELLECTUALISM is wrong he puts himself in the shoes of a putative defendant of it, and presents her with a dilemma that has two untoward consequences. The argument is thus a dilemma within a reduction, and the crucial idea that we should keep in mind while recreating such an argument is that, just like any *reductio,* it is *not* based on premises that the author himself would be happy to endorse in any of its branches, but precisely on those that he wants to dismiss—or at least some of them. Failure to notice this is what makes Stanley's reading of Ryle so misguided. He seems to believe that Ryle himself endorses, assumes or at least presupposes the premises of the argument he puts forward.[52]

Let's begin by considering the way Ryle introduces the argument in his Presidential Address:

> The prevailing doctrine (deriving perhaps from Plato's account of the tripartite soul) holds: (1) that Intelligence is a special faculty, the exercises of which are those specific internal acts which are called acts of thinking, namely, the operations of considering propositions; (2) that practical activities merit their titles 'intelligent', 'clever', and the rest only because they are accompanied by some such internal acts of considering propositions (and particularly 'regulative' propositions). That is to say, doing things is never itself an exercise of intelligence, but is, at best, a process introduced and somehow steered by some ulterior act of theorising. (It is also assumed that theorising is not a sort of doing, as if 'internal doing' contained some contradiction).[53]

I have quoted Ryle in length because the problem with this introduction is in the final brackets—and in the very fact that it is said in brackets. If we took what is said in them at face value, Ryle would be claiming that the position he targets is simply inconsistent—at least if we identified 'thinking' with 'theorising'—in the sense that the prevailing doctrine would be an attempt to preserve two claims that contradict each other. There would not be much point in writing a paper against a position that is introduced as overtly inconsistent. However, the rest of Ryle's paper is not futile because HIDDEN, as I said at the

[52] A similarly unfair criticism may be found in Stalnaker, when he says: "I think the more general intellectualist view that (Ryle) was criticizing is a picture that Stanley should also want to reject. (That is, I think Ryle was right to criticize the intellectualist view of knowing-how. His mistake was to accept, or at least presuppose, an intellectualist account of knowing-that)." (Robert Stalnaker, "Intellectualism and the Objects of Knowledge," *Philosophy and Phenomenological Research* 85, 3 (2012): 755). By my side, I see no mistake in assuming a wrong view in order to reject it by *reductio*.

[53] "Knowing How and Knowing That," 1.

beginning, is no allegedly consistent theory in and by itself, but just an unstructured assumption, some kind of blur desideratum: not a set of well-formed positive theses, but an implicit intuition that guides the authors in their search for the mental. Its lack of consistency is the reason why anybody attempting to respond to it in a positive way will have to confront a dilemma: either she assumes that "thinking" *is* an activity (something that we do), or she denies that it is so, understanding it, or at least its purest manifestations, as static contemplation. Let me label each of those alternatives as:

> ACTIONALISM: thought is an activity (a sort of doing).
>
> CONTEMPLATIONALISM: thought is not an activity ("internal doing" is a contradiction).

In order to respond to their own contradictory desiderata, those willing to propose a theory in accordance with HIDDEN have to go either for ACTIONALISM or for CONTEMPLATIONALISM. The former horn of this dilemma leads to the first one of Ryle's arguments: if, for some action to be intelligent, it must be accompanied by some occult act of thought (we thus enter the *reductio* by assuming HIDDEN), and thinking itself *is* a sort of action (and we opt for the first horn of the dilemma: ACTIONALISM), then that further act of thought is something that the agent *does*. But then it must be something she could do intelligently or stupidly. We certainly want her to do it intelligently, but then HIDDEN forces us to assume that it must be accompanied by some further act, which is what makes it intelligent, and an infinite regress is thus initiated.

The consequences of going for the second horn are not more pleasant: if some action's being intelligent means that it must be accompanied by some hidden act of thought (we enter the *reductio* by assuming HIDDEN too), but thinking is *not* itself an action (we opt for the second horn of the dilemma in this case: CONTEMPLATIONALISM), we then have to account for the way thought, as inert static contemplation, may ever have effects in action, which is dynamic, but lacks itself from intelligence. This second horn forces those bewitched by HIDDEN to envisage a sort of impossible mediator, a 'schizophrenic broker,' who should have a bit of theory and a bit of practice, but be none of them. Nothing, according to Ryle, could ever meet such incompatible demands, at least in the framework of HIDDEN. In other words: there is no escape for those assuming HIDDEN: there is an infinite regress waiting for them at the end of the corridor of ACTIONALISM, and a schizophrenic broker at the end of the corridor of CONTEMPLATIONALISM. They'd better leave HIDDEN behind.

Now, the way I see it, the problem with Stanley's reading is that he fails to grasp Ryle's general strategy, the disjunctive structure of this dilemma, attributing to him each premise of his *reductio* at different moments of his reconstruction. Stanley's Ryle would have somehow assumed both ACTIONALISM *and* CONTEMPLATIONALISM, in order to defeat a contradictory straw man, which would have gone for *both* horns of his dilemma at the same time. Ryle appears in Stanley's eyes as someone who holds both the view that knowledge-that requires the 'contemplation of propositions,' a sort of inner 'preaching,' and the idea that we have to introduce an impossible broker between thought and action. But none of those are theses that Ryle himself endorses! They are only considered by him for the sake of the argument in different horns of the dilemma he confronts his opponent with. If anything, they are *his opponent's* theses, those he wants to reject in the end, by means of a *reductio*, and not the premises that Ryle himself would endorse as his own positive views.

The failure to see this is what makes Stanley summarise Ryleanism as a form of DISTINCTIVISM, something that he does since the perplexing first lines of his first chapter:

> Humans are thinkers and humans are agents. There is a natural temptation to view these as distinct capacities, governed by distinct cognitive states. When we engage in reflection, we are guided by our knowledge of propositions. By contrast, when we engage in intelligent action, we are guided by our knowledge of how to perform various actions. If these are distinct cognitive capacities, then knowing how to perform an action is not a species of propositional knowledge. (…) That there is an important distinction between the kinds of states that guide us in action and the kind of states that guide us in reflection is orthodoxy in much of the most influential work in twentieth-century philosophy. (…)But the most systematic attempt to prove what philosophers and laypersons typically assume, that what guides us in action is a distinct cognitive capacity from what guides us in reflection, is due to Gilbert Ryle, in his major work, *The Concept of Mind*.[54]

I have to admit that reading these very first lines of Stanley's book caused me to jump in my chair—a jump that was somehow my first step into writing this paper. In effect, had Ryle ever claimed this, he would have pictured us all as 'schizophrenic brokers,' divided into the irreconcilable sides of theory and practice, thought and action, contemplation and performance. According to Stanley's Ryle, human beings would be essentially fragmented, having two sorts of 'capacities' or 'cognitive states,' some of them directed at doing and some others at thinking; some being the basis of our know-how, and others grounding our knowledge-that; some would have to do with behaviour, and the others with thought. Stanley is right indeed in denouncing this as a dead end—but it is not Ryle's position. At all.

[54] Stanley, "Knowing (How)," 1.

In fact, for those of us that attempt to make a more sympathetic reading of Ryle, it is hard to conceive a less Rylean picture of the human mind.

When placed in Ryle's general strategy, DISTINCTIVISM appears not as Ryle's general view about the relationship between thought and action, but as the undesired conclusion behind the second horn of the dilemma: CONTEMPLATIONALISM. If Ryle's imagined opponent assumed that genuine thought is not itself a kind of action, but static contemplation, then he would have to introduce something between action and thinking, which is precisely what Ryle wants to show is *not* necessary. Considering his general goal, if we had to restate Ryle's positive views, it is much more sensible to construct Ryle as holding that there is no such thing as a 'distinctive cognitive capacity for reflection' that could be told apart from the sort of capacities that guide us in action: his aim is not to defend that there is a gap between behaviourally inert contemplation and unintelligent mechanical movement, one that would require the introduction of some brokering mechanism, but, on the contrary, that there is *no* such gap to overcome.

To summarise, I find two main troubles with Stanley's reading or Ryle with respect to (5) and (6): first, he considers Ryle's arguments in a summative way, as theses he subsequently endorses, while they should be read as disjunctive alternatives, belonging to different horns of one dilemma. And second, and most importantly, Stanley takes the premises of those arguments as opinions that Ryle himself endorses, or even as the essence of his views on the nature of the human mind, whereas they are only theses he assumes for the sake of the argument, attributing them to his opponent in order to turn an unarticulated preconception (HIDDEN) into a viable theory and, then show that such a theory does not stand up to scrutiny. They are thus not positive theses Ryle would be happy to endorse himself at all.

## 8. Bridging the Divide

This should suffice to show where does the misunderstanding begin and how far it gets. Now, although Stanley fails to identify Ryle's views in some crucial concerns, some genuine disagreement remains. As I said at the beginning, the basis of that disagreement is their opposed assessment of INTELLECTUALISM as an epistemological thesis, which Stanley affirms while Ryle denies. In the remaining part of the paper I would like to discuss Stanley's positive views on the nature of knowledge-how in order to show that, once Ryle's position is correctly understood, they are not so deadly rivals as it may seem. On the contrary, despite their divergence on the specific thesis of INTELLECTUALISM, they both seem to share some important attitudes that are utterly against HIDDEN.

Stanley replies to the first Rylean argument (the infinite regress) by claiming that one may act on some piece of knowledge-that with no need to perform any additional act of considering a proposition, in the sense of 'preaching'. He follows Ginet on this, who rightly defended the point that one may act on a piece of knowledge–that directly, just like one may exercise one's know-how directly, with no need to recall regulative propositions.[55] Even if Ginet holds this view as a criticism of Ryle's opinions, it is hard to imagine Ryle disagreeing on this. If the interpretation I have been proposing is correct, Ryle never makes the positive claim that 'preaching' is a necessary requirement for propositional knowledge-that to have practical effects. This is what, in his opinion, advocates of HIDDEN would be forced to assume if they went for the first horn of the dilemma, which he himself never does.[56]

Now, in order to reply to Ryle's second argument, Stanley holds that the function assigned to the schizophrenic broker could be fulfilled by some kind of automatic process, or by a sort of by-product of mental mechanisms, and does not have to be a further action of the agent.[57] He thus defends the possibility of sub-personal mechanisms that are not themselves agential, but implement the machinery of agency. They would be contentful, but nobody would be aware of their contents.[58] This is probably the point where the divergence between Ryle and Stanley would be stronger, and harder to overcome, given that Stanley's functionalist and modular image of cognition seems to be radically alien to a Rylean conception of the mind—an account of intentionality and rationality that is all deployed at the personal level.

However, in my opinion, a better option for Stanley would be to impeach the very need for a schizophrenic broker instead of holding that the broker is conceivable, realising that such a need only arises when one assumes that theorising is not doing. Why should Stanley buy that premise at all? Why should he hold that acts of thinking are not acts, or that 'internal doing' implies a sort of

[55] Carl Ginet, *Knowledge, Perception and Memory* (Dordrecht: Springer Netherlands, 1975): 7.

[56] See Hetherington, "Knowledge and Knowing," 29-31.

[57] Stanley, "Knowing (How)," 26

[58] Fridland provides compelling reasons to suspect that subpersonal automatic mechanisms could ever fulfil the role required by intellectualists (see "Problems with Intellectualism," 891). Furthermore, such scepticism may be supported by a radical confrontation with the representational cognitivist assumptions that underlie Stanley's approach, in the lines proposed by enactivists, such as Daniel D. Hutto and Erik Myin, *Radicalizing Enactivism. Basic Minds without Content* (Massachusetts: MIT, 2011).

contradiction? Once such assumption is discarded, the need to reply to the challenge disappears. In other words: Stanley does not have to take upon himself the task of finding out a mediator between static contemplation and mechanical action. There is a better way for him to go: to admit a notion of thought that is not alien to practice—a mission for which he could find in Ryle a good ally.

Furthermore, a more sympathetic reading of Ryle's text would show that there are moments where he seems to be preparing the ground for concepts that would later be introduced by Stanley and Williamson in order to understand the particular way in which rules must be grasped by agents in order to be effective in practice. I am referring to practical modes of presentation, which are the ones under which agents are supposed to grasp those regulative propositions that are, in their view, the *content* of know-how.[59] Many authors have claimed in this respect that the notion of practical modes of presentation is a surreptitious way of introducing the very idea that Stanley and Williamson's theory was supposed to explain, namely, know-how.[60] Remember: an agent knows how to perform a certain activity, according to the new intellectualists' theory, in virtue of her knowing a proposition about the way in which she could do it. But grasping that proposition in abstract is not enough: she would have to do it "under a practical mode of presentation," which implies certain dispositions to behave according to the rule. That is what, in Koethe's opinion, commits them to circularly: in order to *know that* that specific way is the right one, the agent would have to *know how* to apply the rule. This criticism is contested by Jeremy Fantl, who objects to Stanley and Williamson's reduction for different reasons.[61] In Fantl's opinion, there is no such circularity, and the problem is quite the opposite one: modes of presentation fall short of being *enough* to guarantee know-how. The fact that the proposition is grasped under a practical mode of presentation is compatible, in his opinion, with the agent being unable to apply the regulative proposition in particular occasions, and therefore it is not enough for her to really know how to do the thing.

I do not want, nor need, to take stance in this discussion. It may well be the case that practical modes of presentation imply spurious circularity, as Koethe claims, or perhaps they do not help solving the infinite regress argument, as Fantl holds. What is relevant for my point is that the very idea of practical modes of presentation is a feature of Stanley and Williamson's account that may be

[59] Stanley and Williamson, "Knowing How," 429

[60] See for instance: John Koethe, "Stanley and Williamson on Knowing How," *The Journal of Philosophy* 99, 6: 327; Katherine Hawley, "Testimony and Knowing How," *Studies in History and Philosophy of Science* 41, 4 (2010): 403.

[61] Jeremy Fantl, "Ryle's Regress Defended," *Philosophical Studies* 156, 1 (2011): 129.

considered in accordance with Ryle's central positive views. In other words: *practical modes of presentation are a Rylean seed at the core of new intellectualism*. The very idea is, in spirit, Rylean. This may sound odd, I concede, but an unprejudiced reading of Ryle would help defuse that sense of oddity. For instance, he holds that "even where efficient practice is the deliberate application of considered prescriptions, the intelligence involved in putting the prescriptions into practice is not identical with that involved in *intellectually* grasping the prescriptions."[62] Such a statement leaves the door open for other ways of grasping those same contents, which are more appropriate for that practical function—such as practical modes of presentation.

The fact that Stanley's positive conception on know-how is not so far from Ryleanism could make Stanley's views seem contentious from the point of view of mainstream cognitivism. That is so because some propositions, according to Stanley, would not even be grasped in the relevant way unless properly rooted in the behavioural patterns of the agent. The possibility, or even the necessity, of being disposed to engage in certain kinds of actions would be *constitutive* of the very understanding of their propositional content. Stanley exemplifies this following Gareth Evans, when he holds that the right comprehension of some *de se* thoughts requires the acquisition of some dispositional properties. The same would happen, in Stanley's opinion, with respect to know-how which, in his view, is a variety of *de se* thought. In Stanley's own words, the kind of intellectualism he intends to deploy is based on 'a view of at least some of the constituents of propositions according to which they can only be entertained if one possesses certain dispositions.'[63] They could thus not consist in pure, simple, theoretical *representations*.

I am not sure that such a view is consistent, but if it were correct, there would be processes of intelligence constituted by what happens, or may happen, 'on the outside,' on the body, on behaviour, at sight, and no narrow definition of such 'intellectual' processes could be restricted to what happens in the inner space of the mind. This may be understood as an attempt, by the part of Stanley, to leave HIDDEN behind, at least partially, in that it explicitly rejects its CONTEMPLATIONALIST horn, by defending 'that propositional knowledge is not behaviourally inert—indeed even entertaining certain thoughts is not behaviourally inert, but entails the possession of dispositions.'[64]

[62] Ryle, *The Concept of Mind*: 37 (my emphasis).

[63] Stanley, "Knowing (How)," 27.

[64] *Ibid*, 98.

In other words: even if Stanley's intellectualism claims that knowing how to do something is just a case of knowing that something is the case, it does not follow that know-how may become a purely 'intellectual' process, in the sense that Ryle found problematic: the body may have not merely a causal, but a constitutive role to play, and action itself would be part of the definition of those epistemic states, and not just their external, causal manifestations. Understanding those propositions from a practical perspective, which is, in Stanley's view, constitutive of know-how, would be something necessarily linked to actual performances and personal practice—all events that may happen in the public scene.

I believe that Stanley is so close to Ryle in this respect that one may even wonder whether their allegedly insurmountable dispute is based on any deep disagreement.[65] The introduction of those dispositional features in the very core of some propositional attitudes removes the grasping of those propositions from the realm of passive contemplation. Stanley and Ryle seem to be there on the same page, sharing the aim to take knowledge-that out of the contemplationalist limbo, which is a good part of Ryle's job against HIDDEN. From that perspective, Stanley's "reasonable intellectualist" owes much to Ryle's views—more that he is willing to confess. It may even be considered as a variation of Ryleanism more than as a reaction against it.

Let me finish with one general reflection that may help framing what is at stake in this confrontation: disconcertingly, outside the debate on knowledge-how—but still inside the field of epistemology—Stanley has defended a position that he himself dubs as 'anti-intellectualist.'[66] In that case, he is against the view that knowledge (in general, but he is focusing there on knowledge-that) is a purely epistemological notion. On the contrary, against this 'purist' position he defends, a variety of what would later be called 'pragmatic encroachment,' as the view that pragmatic factors belong to the core of our epistemic deliberations. He has been rebuked for labelling his own positions in such a misleading way, *viz.* as 'intellectualist' on the debate on knowledge-how, and as 'anti-intellectualist' in the debate on knowledge-that, a decision that apparently endangers the consistency of his general account.[67] Of course, it would be easy to dismiss this apparent

[65] One may wonder, for instance, if Stanley and Williamson's recent views on skill, as "a disposition to know" ("Skill," 715) may be understood as a restatement of Rylean views on know how under a different terminology. The view does seem quite similar to Kremer's unitarian interpretation of Ryle, discussed in section 4.

[66] Jason Stanley, *Knowledge and Practical Interests* (Oxford: Oxford University Press, 2005), 33.

[67] See for instance Stalnaker, "Intellectualism and the Objects of Knowledge," 754.

inconsistency as merely terminological—as Stanley himself does[68]—, but I believe there is a more remarkable moral to be earned from it: Stanley has set himself the general aim of bridging the divide between knowledge and practice, offering an account of the former that is constitutively linked to the latter. And he does so in those two different moments by confronting those views about knowledge-that which are, in his opinion, too intellectual (as in *Knowledge and Practical Interests*), *and* those views about knowledge-how which he finds too anti-intellectual (as in *Know How*). I believe this general project is perfectly consistent, just like it is reasonable to build a bridge by starting it from both sides of the river, which does not mean that one is working *against* oneself. The message I have intended to convey with this paper is that, just as Stanley himself may be found at different moments on different sides of the intellectualist divide, while still being coherent in his general aim, he could have been more alive to the fact that Gilbert Ryle's attack on intellectualism was an attempt to attain quite a similar goal. In that case, he would perhaps have found out that his attacks on Ryle's anti-intellectualism were an unfortunate case of friendly fire.[69]

[68] "Replies to Dickie, Schroeder and Stalnaker," *Philosophy and Phenomenological Research* 85, 3 (2012): 754.

[69] A very incipient version of this paper was presented at 24th SIUCC Workshop at the University of Valencia in 2015. Thanks to that audience for stimulating discussion and helpful feedback, and particularly to Jason Stanley for his openness and friendly reception of my talk. Drafts of the paper have been read by Teresa Bejarano, J. Adam Carter, José Carlos Cañizares-Gaztelu, Boleslaw Charnecki and Britt Harrison who made insightful comments. I wrote part of this work as an Academic Visitor at the Eidyn Research Centre of the University of Edinburgh in two occasions during 2015 and 2016, and I am also very grateful to its director, Duncan Pritchard. Research for this paper was supported by the Spanish MINECO (Ministerio de Economía y Competitividad) via research Grants FFI2014-51811-P and FFI2015-67569-C2-1-P, and the V Plan Propio de Investigación of the Universidad de Sevilla. Their support is gratefully acknowledged.

# DISCUSSION NOTES/ DEBATE

# REPLY TO FORRAI: NO REPRIEVE FOR GETTIER "BELIEFS"

John BIRO

ABSTRACT: In a recent paper in this journal, Gabor Forrai offers ways to resist my argument that in so-called Gettier cases the belief condition is not, as is commonly assumed, satisfied. He argues that I am mistaken in taking someone's reluctance to assert a proposition he knows follows from a justified belief on finding the latter false as evidence that he does not believe it, as such reluctance may be explained in other ways. While this may be true, I show that it does not affect my central claim which does not turn on considerations special to assertion.

KEYWORDS: Gettier, belief, assertion, inference

In a recent paper in this journal, Gabor Forrai[1] offers ways to resist my claim[2] that in so-called Gettier cases the belief condition is not, as is commonly assumed, satisfied. My reason for rejecting the common assumption was that the belief the subject in those cases is supposed to have and which happens, fortuitously, to be true is a belief in a merely pickwickian sense. I contrasted such "beliefs" with what I called serious beliefs, those one is prepared to own and on the basis of which one is prepared to act. I argued that having a merely pickwickian belief is not enough for one to satisfy the belief condition of the justified-true-belief account of knowledge and that therefore that account is left untouched by the supposed Gettier-style counterexamples. Thus in the first Gettier case, involving existential generalization, while Smith believes that someone, namely Jones, has ten coins in his pocket, he only "believes" that someone *or other* has ten coins in his pocket, which is the proposition made true by *his* happening to have ten coins in his pocket. This is shown by the fact that he is not prepared to assert that if Jones does not, someone else does. In the second case, Smith does not seriously believe the disjunctions he is said to have "constructed" (Gettier's word), one of which is made true by the second disjunct's happening to be true, since he is not prepared to say that if the first disjunct is false, the second must be true.

[1] Gabor Forrai, "Gettiered Beliefs Are Genuine Beliefs: A Reply to Gaultier and Biro," *Logos & Episteme* X, 2 (2019): 217-224

[2] In John Biro, "Non-Pickwickian Belief and 'the Gettier Problem'," *Logos & Episteme* VIII, 3 (2017): 47-69.

Forrai challenges this line of argument in two ways. First, he claims that someone's unwillingness to assert something may be explained in ways other than by denying that he believes it. He describes a number of such ways, but I shall not take these up in detail, as I think that even if he is right, the fact that there are other explanations of the unwillingness to assert does not show that there in no conceptual connection between serious belief and assertion of the sort I posited. Take Forrai's twist on the well-known Havit/Nogot case:

> Suppose I want to buy a used Ford and believe that Havit's Ford is up for sale. It would then be perfectly rational to talk to him about buying it. However, if I *also* believe that Havit would not sell me his car for twice the market price because he hates my guts, I will not talk to him. The reason I do not talk to him is not that I do not seriously believe that his car is up for sale is up for sale but that I also believe something else.

All this shows, though, is that, unsurprisingly, the connection holds only *ceteris paribus*. What Forrai's example brings out is that in a particular instance someone who would normally be prepared to assert something may have reason not to assert it. He is right that for this reason his not asserting it is not sufficient evidence that he does not believe it. But the point of insisting on the link between serious belief and willingness to assert was not epistemological.

Forrai says that "[b]elieving that 'Someone or other in the building owns the Ford' amounts [to] believing that 'Someone in the building owns the Ford' and not believing anything concerning who that person might be…" I agree. The question is, can one believe *this* if one believes that someone, namely, Nogot owns a Ford? It is to this question to which I urged a negative answer. Forrai's formulation in fact makes vivid that that must be the right answer: it cannot be the case both that I believe and do not believe something concerning who the owner is.

Others have also wondered about whether tying the seriousness of one's belief to what one is prepared to assert, as I did, is as illuminating as I claimed. Consider lies. Suppose little Timmy says he did not break the window, even though he did. Little Timmy does not seriously believe he did not break the window (he knows he did!), but he is willing to assert that he did not do it. Or, while each gladiator is willing to claim to be Spartacus in order to protect his leader, obviously, none of them seriously believes that he is Spartacus.

However, I offered being prepared to assert as a necessary condition on seriousness of belief, not as a sufficient one. Indeed, Timmy's willingness to deny breaking the window does not show that he believes that he did not, nor does a gladiator's willingness to claim to be Spartacus show that he believes that he is Spartacus. I claimed only that it is a mistake to think that Timmy believes that he did not break the window if he is *not* prepared (*ceteris paribus*) to say that he did not; similarly, it would be a mistake to think that a gladiator believed that he was Spartacus if he were not prepared (*ceteris paribus*) to say that he was. The fact that a *ceteris paribus* clause is needed does not affect the point. *Of course*, one can have reason not to be prepared to assert something one believes (or not to be prepared to act in a certain way). The conceptual connection I have suggested holds between serious belief and preparedness to assert or to act is not thereby compromised.

It is important to emphasize, though, that the main argument against counting Gettier "beliefs" as serious does not rest solely, or even primarily, on considerations having to do with assertion. In fact such considerations are not essential to the argument, as Forrai seems to assume. That this is so can be shown by examples that do not involve assertion at all.

Having just seen our neighbour enter his house, I believe that he is in the house, and, of course, that there is *someone* in the house; I will bet you that there is if you claim otherwise. But I do not believe that there is *someone or other* in the house – let us go and see if it is our neighbour! Of course, even having seen him enter, I could have reason to believe that there is no-one in the house – say, hearing the motorcycle he keeps by the back door start up, its sound gradually fading. Now imagine that my neighbour did leave by the back door, but quietly, on foot. However, he did not lock the door, and a burglar has snuck in. The reason why this is not a Gettier case is that believing that there is someone, namely, my neighbour, in the house is incompatible with believing (though not, of course, with "believing") that there is someone or other in the house, the first proposition's entailing the second and my knowing that it does notwithstanding. If serious, the two beliefs would be based on different evidence and would prompt different actions. Seeing my neighbour enter his house is one thing, seeing the light go on in the living room is another: the former may prompt me to walk over to ask how he enjoyed his trip, the latter, to call the police if I believe him to be still away. That I do not do both shows that I do not believe both that my neighbour is in the house and that someone or other is in the house. (If so, the fact that it is true that there is someone – the burglar – in the house does not show that I have a justified true belief but no knowledge.)

While such cases show that the argument does not turn on considerations special to assertion, they do allow for a gap, similar to that between being prepared to assert and actually asserting, between being prepared to act in a certain way and in fact acting in that way. While actions may speak louder than words, they, too, are not an infallible guide to serious belief. Thus positing a link between action and belief is subject to a *ceteris paribus* clause no less than the link between assertion and belief. But these cases do show that the Gricean considerations Forrai appeals to do not go to the heart of the matter.

In fact, the main point does not turn even on the link between serious belief and action. There is a simple and direct way to make it. Take, again, Gettier's first case, and ask, would Smith believe that the man who will get the job has ten coins in his pocket if he did not believe that Jones will and does? Or would Smith believe that someone in the office owns a Ford if he did not believe that Nogot does? If the answer is, no – as it surely is – does that not show that he does not believe what turns out to be true, namely, that someone else – Smith himself – has ten coins in his pocket or that someone else – Havit – owns a Ford? In the same way, ask if in Gettier's second case I would believe the disjunction that turns out to be true if I did not believe the first disjunct. If the answer is, no – as it surely is – does that not show that even though believing the first disjunct is sufficient for "believing" the disjunction (that is, recognizing that it is entailed by the first disjunct), it is not sufficient for believing it.

But wait! If you are right, we never come to seriously believe something by inferring it from something we believe? A fine pickle! But that is, of course, not what I am suggesting. The inferences in the Gettier examples each have special features that set them apart from the normal case. In the first, Smith's inference needs to be from his belief about Jones to a belief about *someone or other, I have no idea who*, if Smith is to have a belief his getting the job and having ten coins in his pocket can make true. But that belief is incompatible with the belief from which it is supposed to be inferred. I can have it only by ceasing to believe that *Jones* has ten coins in his pocket. In the second example, while the disjunctions supposedly inferred ("constructed," as Gettier tellingly puts it) are made true by the truth of the second disjunct, to believe them seriously requires believing that if the first disjunct is false, the second must be true. Inferring the disjunctions by addition gives one no reason at all to think this.

I close by offering a definition of what I have called serious, non-pickwickian belief:

> For any set of propositions such that one knows that one of them follows from the others but could be true even if those others were not, one believes the entailed proposition if and only if one would believe it even if one did not believe (all) the entailing ones.

This makes room for the idea that one can recognize that it follows from *Fa* that *E(x) Fx* without believing the latter as usually understood, viz., as containing no information about what instantiates *x*. But such recognition is not enough for one to believe the existential generalization so understood, if what one believes is only *Fa.* Believing *Fa* is tantamount to believing that something, *namely a*, is *F.* Someone's believing *that* gives us no reason to think that he believes that if *a* is not *F*, something else is. But that is the belief Smith must have if he is to have a belief that his getting the job and having ten coins in his pocket makes true, and that is the belief he must have if he is to have a belief that Havit's owning a Ford makes true. Similarly, someone's believing *p* and recognizing that *p* entails *p v q* is not enough for one to believe that *p v else q* (that is, that ~*p* → *q*), which is the belief one must have if one is to have a belief that *q*'s being true makes true. Someone who does not believe that *p v else q* believes *p v q* in only a pickwickian sense.

Thus to say, as is said in the typical formulations of Gettier cases that their subjects *infer* the proposition that turns out to be true is misleading in two ways. First, to recognize that a proposition follows from some other(s) is not to infer the first from the second. There is more to inferring than recognizing logical relations. Second, if inferring amounts to coming to believe, the propositions supposedly inferred in Gettier cases (and which turns out to be true) are not ones their subjects infer, even if they see that they follow from propositions they believe.[3]

[3] My thanks to Rodrigo Borges, James Gillespie, Greg Ray and James Simpson.

# FACTIVITY OR GROUNDS? COMMENT ON MIZRAHI

Howard SANKEY

ABSTRACT: This note is a comment on a recent paper in this journal by Moti Mizrahi. Mizrahi claims that the factivity of knowledge entails that knowledge requires epistemic certainty. But the argument that Mizrahi presents does not proceed from factivity to certainty. Instead, it proceeds from a premise about the relationship between grounds and knowledge to the conclusion about certainty.

KEYWORDS: Moti Mizrahi, factivity, epistemic certainty, fallibilism, knowledge

In "You Can't Handle the Truth: Knowledge = Epistemic Certainty," Moti Mizrahi presents an argument for an infallibilist theory of knowledge.[1] Mizrahi claims that the factivity of knowledge entails that knowledge is epistemic certainty. But the argument that Mizrahi presents does not in fact proceed from the factivity of knowledge to knowledge being epistemic certainty. Rather, the argument proceeds from an assumption about the relation between grounds and knowledge to the conclusion about epistemic certainty.

Mizrahi's argument is as follows:

1) If S knows that p on the grounds that e, then p cannot be false given e.

2) If p cannot be false given e, then e makes p epistemically certain.

3) Therefore, if S knows that p on the grounds that e, then e makes p epistemically certain.[2]

As indicated, this argument begins with a premise about the grounds on which the knowing subject knows a proposition. But this is quite different from the claim that knowledge is factive. It is a claim about the relation between grounds (or evidence) and knowing.

More specifically, Mizrahi explains that: "To say that knowledge is factive is to say that, if S knows that p, then p is true." In other words, knowledge is factive in the sense that knowledge requires truth. It is not possible to know a proposition if that proposition is false. Another way of stating the point is perhaps to say that

[1] *Logos & Episteme* X, 2 (2019): 225-227.

[2] Mizrahi, "You Can't Handle the Truth," 225.

knowledge is sensitive to the facts. If what one purports to know gets the facts wrong, then one does not know.

Now it is important to notice that the claim that knowledge is factive says nothing about a relation between grounds and knowledge. All that is required for knowledge to be factive is that the item of knowledge in question be true. There is no mention here of grounds or evidence. The only thing relevant to factivity is truth.

This may only be a small point. But it does seem to show that it is not quite right to claim that the factivity of knowledge entails that knowledge is epistemic certainty. The work is being done, not by the factivity of knowledge, but by the relation between grounds and knowledge.

# NOTES ON THE CONTRIBUTORS

**John Biro** teaches at the University of Florida. His interests are in the philosophy of language, the philosophy of mind, epistemology, metaphysics and the history of philosophy, especially in the seventeenth and eighteenth centuries. He has published papers in these areas in journals such as *The Monist, Australasian Journal of Philosophy, Philosophy and Phenomenological Research, Philosophical Studies, History of Philosophy Quarterly, The Review of Metaphysics and Hume Studies.* He is also co-editor of two collections of papers on Spinoza and one on Frege. Contact: jbiro@ufl.edu.

**Víctor Fernández Castro** is a current post-doctoral research fellow at LAAS-CNRS (Université de Toulouse) and the Institut Jean Nicod (Ecole Normale Superiure & PSL Research University). Previously, he obtained his Ph.D. at the University of Granada in 2017, where he was also a post-doctoral research fellow (2017-2018) thanks to the project "Inner speech, Metacognition, and the Narrative View of Identity" (FFI2015-65953-P). His main areas of interest are the philosophy of language and the philosophy of mind and psychology. His work focuses on social cognition, joint action and inner speech. Contact: vfernandezcastro@gmail.com.

**James M. Joyce** is the C. H. Langford Collegiate Professor of Philosophy at the University of Michigan, Ann Arbor. He is the author of *The Foundations of Causal Decision Theory* (Cambridge, 1999). Contact: jjoyce@umich.edu.

**N. Gabriel Martin** is Visiting Assistant Professor of Philosophy at the Lebanese American University. His primary research focus is the epistemology and phenomenology of disagreement. He contends that the philosophical challenges posed by disagreement are distinct from those posed by doubt, but no less significant. Contact: ngabrielmartin@gmail.com.

**Jesús Navarro** is Associate Professor of philosophy at the University of Seville (Spain). He has published about early modern philosophy (mostly Montaigne), epistemology and pragmatics. He is author of *How to do philosophy with words: reflections on the Searle/Derrida debate* (John Benjamins, 2017). Some recent papers are "Luck and Risk: How To Tell Them Apart" (*Metaphilosophy*, 2019), "The Defeasibility of Knowledge-How" (with J. Adam Carter, *Philosophy and Phenomenological Research,* 2017), "Intention (Including Speech Acts)"

(*Routledge Handbook of Pragmatics*, 2017), and "No Achievement Beyond Intention. A New Defence of Robust Virtue Epistemology" (*Synthese*, 2015). Contact: jnr@us.es.

**Howard Sankey** is Associate Professor of Philosophy in the School of Historical and Philosophical Studies at the University of Melbourne (Australia). He teaches in epistemology and philosophy of science. He has published on the problem of incommensurability, epistemic relativism and scientific realism. He is the author of *Scientific Realism and the Rationality of Science* (Ashgate, 2008), *Theories of Scientific Method: An Introduction* (Acumen, 2007, with Robert Nola), *Rationality, Relativism and Incommensurability* (Ashgate, 1997) and *The Incommensurability Thesis* (Avebury, 1994). More information may be found at https://philpeople.org/profiles/howard-sankey Contact: chs@unimelb.edu.au.

**Brian Weatherson** is the Marshall M. Weinberg Professor of Philosophy at the University of Michigan, Ann Arbor. He is the author of *Normative Externalism* (Oxford, 2019). Contact: brian@weatherson.org.

# *LOGOS & EPISTEME*: AIMS & SCOPE

*Logos & Episteme* is a quarterly open-access international journal of epistemology that appears at the end of March, June, September, and December. Its fundamental mission is to support philosophical research on human knowledgein all its aspects, forms, types, dimensions or practices.

For this purpose, the journal publishes articles, reviews or discussion notes focused as well on problems concerning the general theory of knowledge, as on problems specific to the philosophy, methodology and ethics of science, philosophical logic, metaphilosophy, moral epistemology, epistemology of art, epistemology of religion, social or political epistemology, epistemology of communication. Studies in the history of science and of the philosophy of knowledge, or studies in the sociology of knowledge, cognitive psychology, and cognitive science are also welcome.

The journal promotes all methods, perspectives and traditions in the philosophical analysis of knowledge, from the normative to the naturalistic and experimental, and from the Anglo-American to the Continental or Eastern.

The journal accepts for publication texts in English, French and German, which satisfy the norms of clarity and rigour in exposition and argumentation.

*Logos & Episteme* is published and financed by the "Gheorghe Zane" Institute for Economic and Social Research of The Romanian Academy, Iasi Branch. The publication is free of any fees or charges.

For further information, please see the Notes to Contributors.

Contact: logosandepisteme@yahoo.com.

# NOTES TO CONTRIBUTORS

## 1. Accepted Submissions

The journal accepts for publication articles, discussion notes and book reviews.

Please submit your manuscripts electronically at: logosandepisteme@yahoo.com. Authors will receive an e-mail confirming the submission. All subsequent correspondence with the authors will be carried via e-mail. When a paper is co-written, only one author should be identified as the corresponding author.

There are no submission fees or page charges for our journal.

## 2. Publication Ethics

The journal accepts for publication papers submitted exclusively to *Logos & Episteme* and not published, in whole or substantial part, elsewhere. The submitted papers should be the author's own work. All (and only) persons who have a reasonable claim to authorship must be named as co-authors.

The papers suspected of plagiarism, self-plagiarism, redundant publications, unwarranted ('honorary') authorship, unwarranted citations, omitting relevant citations, citing sources that were not read, participation in citation groups (and/or other forms of scholarly misconduct) or the papers containing racist and sexist (or any other kind of offensive, abusive, defamatory, obscene or fraudulent) opinions will be rejected. The authors will be informed about the reasons of the rejection. The editors of *Logos & Episteme* reserve the right to take any other legitimate sanctions against the authors proven of scholarly misconduct (such as refusing all future submissions belonging to these authors).

## 3. Paper Size

The articles should normally not exceed 12000 words in length, including footnotes and references. Articles exceeding 12000 words will be accepted only occasionally and upon a reasonable justification from their authors. The discussion notes must be no longer than 3000 words and the book reviews must not exceed 4000 words, including footnotes and references. The editors reserve the right to ask the authors to shorten their texts when necessary.

### 4. Manuscript Format

Manuscripts should be formatted in Rich Text Format file (*rtf) or Microsoft Word document (*docx) and must be double-spaced, including quotes and footnotes, in 12 point Times New Roman font. Where manuscripts contain special symbols, characters and diagrams, the authors are advised to also submit their paper in PDF format. Each page must be numbered and footnotes should be numbered consecutively in the main body of the text and appear at footer of page. For all references authors must use the Humanities style, as it is presented in The Chicago Manual of Style, 15th edition. Large quotations should be set off clearly, by indenting the left margin of the manuscript or by using a smaller font size. Double quotation marks should be used for direct quotations and single quotation marks should be used for quotations within quotations and for words or phrases used in a special sense.

### 5. Official Languages

The official languages of the journal are: English, French and German. Authors who submit papers not written in their native language are advised to have the article checked for style and grammar by a native speaker. Articles which are not linguistically acceptable may be rejected.

### 6. Abstract

All submitted articles must have a short abstract not exceeding 200 words in English and 3 to 6 keywords. The abstract must not contain any undefined abbreviations or unspecified references. Authors are asked to compile their manuscripts in the following order: title; abstract; keywords; main text; appendices (as appropriate); references.

### 7. Author's CV

A short CV including the author`s affiliation and professional postal and email address must be sent in a separate file. All special acknowledgements on behalf of the authors must not appear in the submitted text and should be sent in the separate file. When the manuscript is accepted for publication in the journal, the special acknowledgement will be included in a footnote on the first page of the paper.

### 8. Review Process

The reason for these requests is that all articles which pass the editorial review, with the exception of articles from the invited contributors, will be subject to a strict double anonymous-review process. Therefore the authors should avoid in their manuscripts any mention to their previous work or use an impersonal or neutral form when referring to it.

The submissions will be sent to at least two reviewers recognized as specialists in their topics. The editors will take  the necessary measures to assure that no conflict of interest is involved in the review process.

The review process is intended to be as quick as possible and to take no more than three months. Authors not receiving any answer during the mentioned period are kindly asked to get in contact with the editors.

The authors will be notified by the editors via e-mail about the acceptance or rejection of their papers.

The editors reserve their right to ask the authors to revise their papers and the right to require reformatting of accepted manuscripts if they do not meet the norms of the journal.

### 9. Acceptance of the Papers

The editorial committee has the final decision on the acceptance of the papers. Papers accepted will be published, as far as possible, in the order in which they are received and they will appear in the journal in the alphabetical order of their authors.

### 10. Responsibilities

Authors bear full responsibility for the contents of their own contributions. The opinions expressed in the texts published do not necessarily express the views of the editors. It is the responsibility of the author to obtain written permission for quotations from unpublished material, or for all quotations that exceed the limits provided in the copyright regulations.

### 11. Checking Proofs

Authors should retain a copy of their paper against which to check proofs. The final proofs will be sent to the corresponding author in PDF format. The author must send an answer within 3 days. Only minor corrections are accepted and should be sent in a separate file as an e-mail attachment.

### 12. Reviews

Authors who wish to have their books reviewed in the journal should send them at the following address: Institutul de Cercetări Economice şi Sociale „Gh. Zane" Academia Română, FilialaIaşi, Str. Teodor Codrescu, Nr. 2, 700481, Iaşi, România.

The authors of the books are asked to give a valid e-mail address where they will be notified concerning the publishing of a review of their book in our journal. The editors do not guarantee that all the books sent will be reviewed in the journal. The books sent for reviews will not be returned.

### 13. Property & Royalties

Articles accepted for publication will become the property of *Logos & Episteme* and may not be reprinted or translated without the previous notification to the editors. No manuscripts will be returned to their authors. The journal does not pay royalties.

### 14. Permissions

Authors have the right to use their papers in whole and in part for non-commercial purposes. They do not need to ask permission to re-publish their papers but they are kindly asked to inform the Editorial Board of their intention and to provide acknowledgement of the original publication in *Logos & Episteme*, including the title of the article, the journal name, volume, issue number, page number and year of publication. All articles are free for anybody to read and download. They can also be distributed, copied and transmitted on the web, but only for non-commercial purposes, and provided that the journal copyright is acknowledged.

### 15. Electronic Archives

The journal is archived on the Romanian Academy, Iasi Branch web page. The electronic archives of *Logos & Episteme are* also freely available on Philosophy Documentation Center  web page.