THEORIA

Vol. 41/3 • September 2026

Established in 1952 by Miguel Sánchez-Mazas
Second Series

An International Journal for Theory, History and Foundations of Science



Ὁ Θεὸς ἀριθμετίζει

CALIJ
Centro de Análisis, Lógica e Informática Jurídica (CALIJ)

EHU

https://ojs.ehu.eus/index.php/THEORIA

THEORIA

REVISTA DE TEORÍA, HISTORIA Y FUNDAMENTOS DE LA CIENCIA

AN INTERNATIONAL JOURNAL FOR THEORY, HISTORY AND FOUNDATIONS OF SCIENCE

ESTABLISHED in 1952 by MIGUEL SÁNCHEZ-MAZAS Second Series

EDITORIAL BOARD

Editor-in-chief: Andoni IBARRA (University of the Basque Country, EHU)
Editors: María CAAMAÑO (Universidad de Valladolid), Michele PALMIRA (Universidad Complutense de Madrid)

Metaphysics: Delia BELLERI (CSIC)
Logic and philosophy of logic and mathematics: Eduardo A. BARRIO (Universidad de Buenos Aires)
Philosophy of language: Andrea ONOFRI (Universidad Autónoma de San Luis Potosí)
Philosophy of mind, cognition, and action: Marta JORBA (Universitat Pompeu Fabra)
Epistemology: Michele PALMIRA (Universidad Complutense de Madrid)
Philosophy of science: José DÍEZ (Universitat de Barcelona)
History of science & Science and Technology Studies: Miguel GARCÍA-SANCHO (University of Edinburgh)

Consulting Editor: Cristina CORREDOR (Universidad de Valladolid)
Managing Editor: Mario SANTOS-SOUSA (University College London)
Editorial Assistant: Jorge GUARDIOLA BLEDA (Universitat de València)

ADVISORY BOARD

Juan José ACERO (†) (Universidad de Granada), Ignazio ANGELELLI (†) (University of Texas at Austin), Manuel ATIENZA (Universitat d'Alacant), Salvador BARBERÀ (Universitat Autònoma de Barcelona), Nancy CARTWRIGHT (Durham University), Newton C.A. da COSTA (†) (Universidade de São Paulo), Joseph DAUBEN (City University of New York), Elías DÍAZ (Universidad Autónoma de Madrid), Javier ECHEVERRIA (Ikerbasque & EHU), Josep Maria FONT (Universitat de Barcelona), Manuel GARCÍA-CARPINTERO (Universitat de Barcelona), Philip KITCHER (Columbia University), Bruno LATOUR (†) (Sciences Po, Paris), Larry LAUDAN (†) (UNAM), Javier DE LORENZO (Universidad de Valladolid), Thomas MORMANN (EHU), C. Ulises MOULINES (Ludwig-Maxilians-Universität München), Carlos MOYA (Universitat de València), Javier MUGUERZA (†) (UNED), Roberto TORRETTI (†) (Universidad de Puerto Rico), Enric TRILLAS (Universidad Politécnica de Madrid), Nicanor URSUA (EHU), Bas C. VAN FRAASSEN (San Francisco State University)

EDITORIAL OFFICE

Editorial manager: Ainhoa FERNÁNDEZ (CALIJ, Donostia/San Sebastián, Spain)

Mailing address: CALIJ-THEORIA, Elhuyar plaza 2 (20018 Donostia/San Sebastián, Spain)
Tel.: (+34) 943 017 447. Fax: (+34) 943 015 990. E-mail: editor.theoria@ehu.eus

THEORIA is a four-monthly journal (issues in January, May and September). Accepted articles are published in Advanced Access as soon as they are ready, prior to their assignment to a specific issue. The contents of this journal are covered, among others, by the following abstracting and indexing services: Arts & Humanities Citation Index®, ISI Alerting Services, Current Contents® / Arts & Humanities, Bulletin Signalétique 519, DICE, ICYT, ISOC and RESH of the Spanish research center CSIC, Mathematical Reviews, Current Mathematical Publications, MathSci, Philosopher's Index, Repertoire bibliographique de la Philosophie, and SCOPUS.

THEORIA's website

http://www.ehu.eus/theoria

Contents

Volume 41/3, September 2026, pp. 345-469

ISSN 0495-4548
e-ISSN 2171-679X

Articles

Karl Gustav Bergman
Are proper functions normative? / *¿Son normativas las funciones propias?* 347

Esteban Withrington
Millikan on picturing and correspondence / *Millikan sobre figuración y correspondencia* 367

Santiago Arango-Muñoz, Sebastián Mejía-Rendón
Metacognitive feelings as mental affordances / *Sentimientos metacognitivos como affordances mentales* 378

Marco Facchin
Illusionism as a theory of mental content / *El ilusionismo como teoría del contenido mental* 396

Murali Ramachandran
The opacity of knowledge: An externalist's case for widespread KK-failure / *La opacidad del conocimiento: un argumento externista a favor del fracaso generalizado del principio KK* 434

Ken Aizawa
A case of entity realists using abductive confirmation / *Un caso de realistas de entidades que recurren a la confirmación abductiva* 445

Summary 462

Articles

Theoria, 2026, 41(3), 347-366
https://doi.org/10.1387/theoria.28092

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Are proper functions normative?

(¿Son normativas las funciones propias?)

Karl Gustav Bergman*
Universitat Pompeu Fabra
https://orcid.org/0000-0001-9635-6762

Keywords

Proper functions
Norms
Normativity
Oughts
David Papineau

ABSTRACT: Are proper functions normative? In this paper, I try to answer this question. Acknowledging that the notion of normativity is too ill-defined to afford a categorical yes/no answer, I proceed by attempting to identify arguments for an affirmative answer and trying to articulate what conception of normativity they rely on: what "normativity" has to be, in order for these arguments to work. I identify three arguments for the conclusion that proper functions are *bearer* normative, i.e., normative for the things that have the functions in question. I then propose that two of the arguments may be generalized to the *agentive* case, in defense of the conclusion that proper functions are also, at least in some cases, normative for agents.

Palabras clave

Funciones propias
Normas
Normatividad
Deber
David Papineau

RESUMEN: ¿Son normativas las funciones propias? En este artículo intento responder a esta pregunta. Reconociendo que el concepto de normatividad está demasiado mal definido como para ofrecer una respuesta categórica de sí o no, procedo a identificar los argumentos a favor de una respuesta afirmativa y a articular en qué concepción de la normatividad se basan: qué debe ser la «normatividad» para que estos argumentos sean válidos. Identifico tres argumentos en favor de la idea de que las funciones propias poseen *normatividad de dispositivo*. Es decir, que son normativas para las entidades que tienen las funciones en cuestión. A continuación, propongo que dos de los argumentos pueden generalizarse al caso *agente*, en defensa de la conclusión de que las funciones propias son también, al menos en algunos casos, normativas para los agentes.

Gako-hitzak

Funtzio propioak
Arauak
Normatibotasuna
Izanbeharrak
David Papineau

LABURPENA: Normatiboak al dira funtzio propioak? Artikulu honetan, galdera horri erantzuten saiatu naiz. Oharturik normatibotasunaren kontzeptua lausoegia dela eta, beraz, ezinezkoa dela bai/ez erako erantzun kategoriko bat ematea, saiatu naiz baiezko erantzunaren aldeko argudioak identifikatzen eta argudio horiek normatibotasunaren zer ikuskeratan oinarritzen diren artikulatzen: nolakoa izan behar du «normatibotasun»ak argudio horiek balekoak izan daitezen? Hiru argudio identifikatu ditut funtzio propioak haien eramaile diren entitateentzat normatiboak direla ondorioztatzearen alde. Ondoren, proposatu dut hiru argudio horietako bi agenteen kasura orokortu daitezkeela, funtzio propioak, kasu batzuetan behintzat, agenteentzat ere normatiboak direla defendatzeko.

* ***Correspondence to:*** Karl Gustav Bergman. Universitat Pompeu Fabra, Departament d'Humanitats, Edifici Jaume I, Ramón Trias Fargas, 25-27, 08005 Barcelona, Spain. – mail@karlbergman.com – https://orcid.org/0000-0001-9635-6762

How to cite: Bergman, Karl Gustav (2026). «Are proper functions normative?»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 41(3), 347-366. (https://doi.org/10.1387/theoria.28092).

Received: 01 December, 2025; Final version: 02 February, 2026.
ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press

This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. Introduction

Proper functions seem to be, in some sense, normative. To say of the heart that it has the function to pump blood is, it seems, *ipso facto* to say that it is *supposed* to pump blood, or that it *ought* to pump blood, or that a heart that fails to pump blood is a *bad* heart. Theorists of proper function often treat it as a desideratum on their theories that they be able to account for this normativity. At the same time, there are those who deny that proper functions are normative in more than a deflationary or metaphorical sense:

> Whatever norms are, I take it that they must involve some kind of prescription, some kind of implication about what ought to be done. [...] It is a vulgar, and indeed dangerous, error to infer, from the premise that X has been biologically designed to Y, that in some sense X ought to Y. My knuckles have arguably been biologically designed to hit people with, but it doesn't in any sense follow that I ought so to use them. (Papineau, 1999, p. 21, fn. 5)

Meanwhile, whether and how proper functions are normative is of concern not only for the philosophy of proper functions itself. The answer is at stake in several key debates within philosophy of language, philosophy of mind, and epistemology.

For instance, is meaning and content normative, as Saul Kripke (1984) influentially suggested? Some have powerfully argued that the answer is "no" (Glüer & Wikforss, 2009; Hattiangadi, 2006)—but others (e.g., Jarvis, 2012) argue that these authors have overlooked the normativity of proper functions and that once we take it seriously, their arguments crumble. Relatedly, there is growing interest in the *norms of belief*: Do evidence alone give reasons for belief, or can belief also be justified by practical reasons (as argued by, e.g., McCormick, 2014)? This literature often appeals to the "aim of belief" (Shah, 2003; Wedgwood, 2002; Williams, 2006); and though writers understand this idea differently, it is naturally understood in functional terms: The aim of belief is its function. Finally, there is much recent interest in the idea that concepts of epistemic normativity, like justification and warrant, can be analyzed in functional terms (Burge, 2003; Graham, 2012; Kelp, 2018; Millikan, 1984b; Olson, 2018; Plantinga, 1993; Simion, 2019).

So the titular question seems motivated. In this paper, I'll take a stab at answering it. But it must be acknowledged, right out the gate, that the question does not admit of any neat yes/no answer. We simply do not have the sort of firm independent grasp on the notion of normativity that this would require. On one hand, "normativity" is a term of art unaccompanied by strong intuitions of the kind formed by everyday use. On the other hand, there is no generally agreed-upon stipulative definition that we can simply bring to bear. We are not completely in the dark, of course. We have some sense of what normativity is, some ideas about which things are *definitely* normative if anything is (like morality) and which things are *definitely* not (like rocks). But this is of limited help in controversial cases, and proper functions constitute one of those.

Accordingly, I will not defend a categorical yes/no answer to the titular question. What I will do instead is to look for *possible* arguments for an affirmative answer and try to articulate what conception of normativity they rely on: what "normativity" has to be, in order for these arguments to work.

In keeping with this general aim, I will spell out three arguments for the normativity of proper functions: The Argument from teleology, the Argument from propensity, and the Argument from copy-producing power. Each argument proceeds from a particular feature that proper functions have been argued to possess. It adds a *general doctrine on the nature of normativity*, and concludes that proper functions are normative. Much of the paper will consist in laying the ground for these arguments. I will devote fewer words to *defending* the general doctrines on the nature of normativity that the arguments rely on: The point is mainly to pin down these doctrines, to make it clearer what commitments accompany an affirmative answer to the titular question. That said, the exercise would be pointless unless there were *some* constraints on what normativity could be, so I will try to give these doctrines at least initial plausibility.

In formulating the arguments, I have been guided by the idea that an affirmative answer to the titular question should be based on features to proper functions whose normativity-conferring status is *independently* ascertainable. In other words, I have sought to identify features such that proper functions may be held, in virtue of them, to share a nature with more uncontroversial cases of normativity. This approach is not mandatory. One may take the normativity of proper functions as a brute datum, one that any conception of normativity must accommodate. I hope, however, that even philosophers with this latter inclination may find my conclusions of some interest, as they help us see some of the connections that do exist between proper functions and other, more paradigmatically normative phenomena.

As we will see below, in discussing the putative normativity of proper functions, it will be important to distinguish between *bearer norms* and *agent norms*. The former are norms applying to the thing (the organ, artefact, behavior etc.) that has the function, whereas the latter are norms applying to agents such as human beings. My three arguments are, in the first instance, arguments for the *bearer* normativity of proper functions. Developing these arguments are the task of section 4. In section 5, I consider the prospects for extending these arguments to the agent case. These two sections are preceded by two that lay the conceptual groundwork. In section 2, I explain proper functions. In section 3, I characterize normativity itself: its metaphysics, its structure, and some of the forms it can take.

2. *Proper functions*

Intuitively, a proper function is what its bearer—a biological trait, an artifact, a behavior—is *supposed to* do, in contrast to the many other things it could or might do. The proper function of a heart is to pump blood. Hearts also typically make thudding sounds, but that is not their function. The proper function of a hammer is to drive nails. It can also serve as a paperweight or a doorstop, but those are not its proper functions.

The notion of proper function is, in the first instance, identified ostensively, by enumeration of examples: the heart pumping blood, the hammer driving nails. We intuit that these examples have something in common. The intuition is corroborated upon recognition that the examples share certain general features.

One such general feature that has attracted special attention is *malfunction*. To attribute a function to an entity is not to say what it *will* do or even what it *could* do, because entities can fail to perform their functions and indeed be incapable of so doing. An injured heart may fail to pump blood. A hammer whose head has come loose is useless for driving nails. Such failures of performance don't mean that the thing lacks the function. To the contrary, it is only because it possesses the function that it makes sense to say that it has failed to perform it. As we'll see in section 4.1, this possibility of malfunction has sometimes been identified as that wherein the normativity of proper functions resides.

Another general feature of proper functions is that they enter into a special kind of explanation of the existence and structure of their bearers. To the question "why do we have hearts?" we can answer "they are there in order to pump blood." To the question "why do hammers look the way they do?" we can answer "in order to drive nails." These explanations have the structure of *teleological* explanations—explanations by appeal to an end. Thus, proper functions can be regarded as *teloi* for their bearers. We'll return to this idea in section 4.3.

Philosophical theories of proper function seek to achieve extensional adequacy by attributing the intuitively right functions to the right things, while also accounting for the general features just described.

Such theories are often expected to satisfy some additional constraints. One such constraint is *naturalism* or *reductionism*, meaning that the conditions for possessing this-or-that proper function should be specifiable in terms that make reference neither to functions nor to other domains in need of reduction. Importantly for present purposes, the latter category includes the *normative* domain. If, as some have argued (e.g., Jarvis, 2012), the functional is reducible to the normative, then the titular question would be trivially answerable in the affirmative. So as not to make life too easy for ourselves, and in keeping with majority opinion, we'll take the reductionist assumption on board here.

Another commonly accepted constraint is *objectivity*. When something has a function, that's an objective fact about it—not, for instance, something projected upon it by the observer. This assumption has been challenged, e.g., by Searle (1995). Here, nevertheless, we will take it on board.

There are two main families of theories on the nature of proper function meeting the constraints of reductionism and objectivity. First, we have *etiological views* (Godfrey-Smith, 1994; Millikan, 1984a; Neander, 1991; Wright, 1973), according to which a proper function is, very roughly, a selected effect: something the bearer's ancestors have done, such that the propensity to do it is heritable and which helps explain the past persistence and proliferation of the function-bearer lineage.

Second, we have *propensity views*, which hold that functions are, roughly, effects such that if they were produced by the function-bearer, this would lead to some outcome—e.g., an increase in the bearer's fitness (Bigelow & Pargetter, 1987; Griffiths, 2016; Nanay, 2010) or maintenance of the structural integrity of the enclosing organism (e.g., Mossio *et al.*, 2009). Sometimes, the view is formulated in the indicative—a function is an effect that an entity produces and which leads to the relevant outcome—but the counterfactual formulation, made explicit by Nanay (2010), seems preferable, since it accommodates the possibility of malfunction. Unlike etiological views, propensity views make no particular demands on a function-bearer's history—indeed, the bearer may altogether lack a selective history.

It is fair to say that of these two broad views, the etiological theory is the leading contender. In section 4.4, we will take a closer look at the two and their respective implications for this essay's topic.

Both views are primarily motivated by reflection on *biological* functions. It is a further question whether they can be generalized to other seemingly function-bearing entities—in particular, to artifacts. It seems *prima facie* plausible that artifacts, like hammers, derive their functions neither from selection nor propensities, but from the intentions of their creators. Despite this, representatives of both camps have argued that their views generalize to artifacts, given due generalization of the operative notions of fitness, selection history, and so forth. The alternative is a *pluralism* about proper functions, according to which biological and artifactual functions require distinct theoretical accounts. Here, I do not take stands in this debate. I limit myself to highlighting, in what follows, those points at which a commitment to pluralism would impact the argument.

With this primer on proper functions now concluded, we'll turn next to consider the notion of normativity itself.

3. Normativity

What is normativity? As noted in the introduction, we don't possess any very informative definition of this notion. There are theories, but they are controversial and best not presupposed at the outset.

But we are not completely in the dark. We have some grasp, albeit a loose one, of what normativity is, and some sense of which things are called "normative." By gathering these ideas together, we can assemble a preliminary characterization of normativity suitable for our purposes. We can say what sort of thing normativity attaches to, characterize the structure of normative facts, and sketch a typology of different sorts of normativity. This I will do in, respectively, the three subsections to follow.

3.1. The normativity of what?

The word "normative" is used in many ways. We talk about

— Normative properties

— Normative facts

— Normative concepts

— Normative terms

and so on.

The question whether *proper functions* are normative sounds like a material-mode question, a question about properties and facts "out there" rather than about the representational devices—concepts and words—with which we think and talk about those properties and facts. This is also how the question will be understood here. It would be possible to conduct the investigation in the formal mode, examining whether the *concept* of proper functions or the *term* "proper function" are normative. However, we are assuming that proper-function discourse is factive and that "proper function" refers to a real, objective feature of things. On this assumption, the answer to the material-mode question bears directly on the formal-mode one. If it turns out that proper functions, the worldly phenomena, are normative, then it follows that the concept/term is also normative, in at least one important sense. *Mutatis mutandis* if it turns out that they're not. (Granted, even if proper functions themselves are not normative, there is a further sense in which the concept or word could still be; but I will not explore this possibility here.)

The question whether proper functions are normative, then, will be understood as a question about worldly phenomena. Which worldly phenomena? Well, proper functions, of course—but what are those? Viewed in the abstract, a proper function is just some possible effect an entity could have: pumping blood, driving nails, etc. But effects, as such, do not seem like the right sort of thing to be normative. When we talk about more familiar kinds of norms, like moral norms, it is not the *action* φ, the one Peter ought morally to perform, that is considered normative. Rather, what's normative is the *fact* that Peter morally ought to φ. By analogy, our question here is whether *proper-function facts*—facts of the form [bearer b has proper function ψ]—are normative.

Are *all* such facts normative if any of them are, or could it be that only some are? I take it that when we attribute normativity to proper-function facts, we typically imagine that there's something about the *proper-function relation*—the relation in which a bearer b stands to an effect ψ when ψ-ing is its proper function—that confers normativity on the fact that b has proper function ψ. If this is right, then, since this relation is a component of all proper-function facts, all of them are normative if any are. I will not presuppose that here, however. I'll leave open the possibility that only some proper-function facts are normative. In particular, if pluralism about proper functions is true (see end of last section), it would not be too surprising if it turned out that, e.g., biological functions are normative but artifactual ones not (or vice versa).

Below, I will often speak of facts about morality or rationality, thereby presupposing that there *are* indeed facts of these kinds. This, I recognize, is controversial. However, I think it will be clear that none of my arguments rely on these assumptions. My references to moral facts etc. can be considered hypothetical: If there are such facts, this is what we should say about them.

3.2. Structure of normative facts

Inspecting the aforementioned normative fact that Peter ought morally to φ, we see that it seems to involve two entities: a *type of conduct*, φ, and a *subject*, Peter, where Peter is linked to φ-ing by a special relation, the "morally-ought-to relation." In virtue of standing in this relation to φ, Peter is "called upon" to engage in the conduct in question and "precluded from" engaging in the complementary conduct, not φ-ing (Copp, 2024, p. 15). Some such structure seems to be present in most, if not all, of the facts we think of as normative. The subject need not be a specific individual: It may be a group or a kind, perhaps a very insubstantial kind such as rational beings generally.

In the case of the fact that Peter ought to φ, we can read this tripartite structure off of the syntactic structure of the sentence used to identify it, "Peter ought to φ," which contains terms referring to all three components. This is not al-

ways the case. Consider facts like *φ-ing is right*, whose identifying sentence has no term naming a subject (the sentence's *grammatical* subject instead names a conduct). But even in such cases, it is clear that the norm "calls upon" *somebody* to do the *φ*-ing, even if this somebody is just people, or rational beings, generally. Some subject is *involved* in the fact or presupposed by it, even if not named in the sentence that identifies it. Indeed, I find it hard to see what would be normative about a fact unless it, as it were, *placed demands* on someone (or something).

Some *evaluative* facts, such as the fact that some state of affairs or outcome is good, are identified by sentences lacking names for either subjects or conducts. But even here, it seems, subjects and conducts are *involved*: The obtaining of such a fact means that some subject (which may, again, be people or rational beings generally) is called upon to engage in some conduct, such as striving to bring about or preserve the outcome or state of affairs in question, *modulo* other considerations.

One clear counterexample is facts about permission, such as the fact that Peter is morally permitted to *φ*. This fact does not call upon anyone to do anything—though it has implications for the structure of calling-for and precluding, by withholding from *φ*-ing the status of being precluded. It seems unlikely, however, that proper-function facts will turn out to be normative in the manner of permission facts, so we can safely bracket this counterexample for present purposes.

3.3. Degrees of normativity

It is widely held that normativity comes in two varieties: *strong* and *weak* (Schulte, 2020), or *robust* and *formal* (McPherson, 2011); or *reasons-implying* and *rules-implying* (Parfit, 2011). Though the distinction eludes neat definition (a recurring theme in this essay), strong normativity is typically understood to be the sort of normativity that is "genuinely action-guiding" or has "genuine normative clout," whereas weak normativity lacks those qualities. It is typically assumed that morality and rationality are strongly normative. Sometimes, as in (Parfit, 2011, , vol. 2, p. 268), strong normativity is defined in terms of *reasons*: It is the kind of normativity that entails reasons for action. Sometimes it is defined in terms of *oughts*: It is the kind of normativity that entails oughts.[1] Weak normativity, on the other hand, is "merely connected to judgments of correctness and incorrectness" (Schulte, 2020, p. 2284; cf. McPherson, 2011, p. 232). Standard examples of weak normativity come from the world of games, social conventions, and other man-made systems of rules.

It is a natural thought that insofar as proper functions are normative, their normativity is, at best, weak (e.g., Schulte, 2020, p. 2284). This may be so; but it's important to forestall a suspicion that may accompany this thought, namely, that there is something *trivial* about the claim that proper functions are weakly normative, so that if an affirmative answer to the titular question rested only on this claim, that answer would be obvious and therefore uninteresting. This suspicion, I believe, is groundless.

To explain why, I first need to introduce the notion of a *standard*. By a "standard," I mean a certain kind of formal object: roughly, a specification of the conditions under which the standard is met and those under which it is not. Standards, in this formal sense, are akin to the kind of abstract object that Paul Boghossian and others call a "rule" (Boghossian, 2015). They can be likened to propositional contents, especially when these are construed as sets of worlds.[2] Just like a proposition can characterize the actual beliefs of some real person, so a standard can characterize a substantive norm or body of rules that possesses concrete reality, e.g., Spanish law. But just like propositions are not *identical to* beliefs, but are just formal objects that can be used to characterize them, so the standard that characterizes Spanish law is

[1] Though not any ought will do. As we'll see, proper functions entail oughts, but this is presumably insufficient to render them strongly normative.

[2] Indeed, standards may be construed as normative contents à la Allan Gibbard (1990; 2003), i.e., as partitions of the set of all possible token actions into the "forbidden" and the "permitted."

not identical to Spanish law itself, considered as a concrete social-historical reality. And just as there is an abundance of propositions that do not characterize the beliefs of any real person, so there is an abundance of standards that do not characterize any norm or body of rules.

Now, the suspicion that there's something trivial about the claim that proper functions are weakly normative arises, I believe, from a certain view of what weak normativity consists in; namely, that it consists in *being characterizable by a standard*. If this were all that weak normativity consisted in, there would indeed be something trivial about it, since this property is very easy to come by. Given only a suitable *mapping* (i.e., a function in the mathematical sense—I use the word "mapping" here to avoid confusion) from standards to facts, *any* fact can be characterized by reference to a standard, as the standard's image under the mapping. For example, my being in the habit of drinking coffee in the mornings can be characterized by reference to the standard—call it Coffee—that requires drinking coffee in the morning: I'm in the habit of doing what Coffee requires. (Note that the word "requires" is idle here—it's just a label for the relevant partition.)

But for precisely this reason, weak normativity should *not* be identified with being characterizable by a standard. There is obviously nothing normative, not even weakly so, about the fact that I drink coffee in the morning. Or, if we insist that there is, then it is nevertheless clear that the standard examples of weak norms—the rules of chess, club bylaws, Spanish law—are normative in some *further*, more substantive sense.

Spanish law is characterizable by the Spanish-law standard, but this fact does not exhausts their relation. They relate in a characteristically normative way: Spanish law, as we might say, renders the Spanish-law standard *in force* for specific subjects (cf. Glüer & Pagin, 1998; García-Carpintero, 2022). Ana, a inhabitant of Spain, is subject to Spanish law—it is in force for her. She is not subject to "Shmanish law," which is like Spanish law except that it also requires you to walk backwards on Thursdays. During a game of chess, the rules of chess are in force for the players. The rules of *schmess*—like chess, but the knight can move diagonally (McPherson, 2011)—are not in force.

Some philosophers, like Manuel García-Carpintero (2022, p. 18), suggest that a standard's being in force is a matter of there being reasons to conform to it. If so, in-force-hood would seem to reduce to *strong* normativity. But surely, there exists a degree intermediate between strong normativity and mere standard-characterizability. There are systems of rules that, while not necessarily yielding reasons for action in the ethicist's sense, are nevertheless more than *just* things we can describe with reference to formal objects. Spanish law may not by itself give Ana reasons to act (though of course, she may have prudential and moral reasons to obey the law), but there is a clear sense in which it *exerts authority* or *wields clout* over her, in a way the fact that I'm in the habit of drinking coffee does not (over me or anyone else). Precisely what this rendering-in-force consists in is something to which I'll return later (section 4.4), but I take it that the intuition is clear.

The claim that proper functions are weakly normative, then, cannot be reduced to the claim that they are characterizable by reference to standards. They *are* characterizable by reference to standards, that much is true. The proper function to ψ can be characterized as the proper function to do what S_ψ requires, where S_ψ is the standard that requires ψ-ing. If this were all that we could say about them, this would indeed be a trivial and uninteresting fact about them. But to say that they are weakly normative is to say something more: It is to say that they render those standards *in force*, in the way Spanish laws renders the Spanish-law standard in force. This claim is nontrivial. In section 4.4, I will give reasons to think it is true.

If it is true, does that entail an affirmative answer to the titular question? Some may think that weak normativity is not really normativity at all. But this, I think, is wrong: Weak normativity, in the above sense, is really a kind of normativity. For one, it evinces the tripartite structure described in section 3.2. It involves a subject or subjects (Ana) and a type of conduct (obeying Spanish law); and by rendering in force a standard requiring that conduct for the subject, it would seem to "call upon" the subject to engage in the conduct, in some natural sense of that deliberately vague term.

But we need not settle this issue here. Our aim, as stated in the introduction, is to find possible arguments for an affirmative answer to the titular question and try to articulate what conception of normativity they rely on. Here, we have articulated one conception. The corresponding argument, as noted, will come in section 4.4.

4. Bearer normativity

We now have some sense of what normativity is and the different forms it takes. It is time to get to work and examine what can be said in favor of the view that proper functions possess the quality thus characterized.

Here's one place to start. If I was right in section 3.2, normative facts characteristically involve a *subject* for which the norm in question is a norm. We may then ask: If proper-function facts are normative, *who are they normative for*? The most natural answer, I believe, is that they are normative *for the functions' bearers*. In a word, proper functions are "bearer normative." In this section, we will explore this idea.

4.1. Bearer normativity: the very idea

To introduce our theme, let us return to the introductory quote from David Papineau:

> It is a vulgar, and indeed dangerous, error to infer, from the premise that X has been biologically designed to Y, that in some sense X ought to Y. My knuckles have arguably been biologically designed to hit people with, but it doesn't in any sense follow that I ought so to use them. (Papineau, 1999, p. 21, fn. 5)

(Papineau talks about "biological design" rather than "function," but I take it that to him, these amount to the same.)

The first thing to note is that, *contra* Papineau, proper functions *do* imply oughts (and their close linguistic cousins, shoulds).[3] We say things like, "the heart ought to pump blood," meaning: that is its function. The electrician says, "this circuit should give current to the bathroom lights," meaning: that is its function. Indeed, one way of saying that something functions properly is to say that it "works as it ought to" or "as it should."

Now, these are oughts applying to the *bearers* of said functions: It is the *heart* or the *circuit* that ought to do something. Papineau, meanwhile, is surely right that it would be a mistake to infer, from the fact that his knuckles are for punching, that *he*, the knuckles' proud owner, ought to punch.[4] But this only shows that proper functions are not normative for any *agents* who happen to be around. They could still very well be normative for their *bearers*. They could be bearer normative.

Does the notion of bearer normativity make sense? Many paradigm instances of normativity apply to agents. Norms of morality and practical rationality, for instance, are norms governing action: They prescribe conduct for agents. Bearers, however, are not agents, and so one may wonder if the notion of a norm is even applicable to them.[5]

---

[3] At least some of the time—cf. (Franssen, 2009, p. 113 ff.)

[4] It also sounds weird to say that Papineau's *knuckles* ought to punch. However, this is an artifact of the example. Presumably, the function of Papineau's knuckles is not in fact to punch, full stop, but rather to increase the pain and trauma inflicted upon his adversary *if and when* he punches. The function is *relational*, in Millikan's (1984a, p. 39 ff.) sense. And saying "Papineau's knuckles ought to increase the pain and trauma inflicted on his adversary if and when he punches" sounds much less weird.

[5] I assume, here and throughout, that organism—and so, *a fortiori*, agents—lack proper functions of their own. This, I take it, is the majority opinion; though see (Burge, 2003) for an opposing view.

Certainly, the idea of a *prescription* applying to a nonagent makes little sense, but it is not a given that all norms need be prescriptions, and seeming examples of nonagentive norms are not far to seek. Epistemic norms, for instance, apply to episodes of belief-formation and the like, which arguably aren't actions—at least not in the normal sense. They are, of course, things that happen in agents, but so is the pumping of hearts and the punching of knuckles. Such mere spatial proximity to an agent can hardly be enough. Indeed, it is presumably precisely because epistemic norms can plausibly be construed as nonagentive that many authors have found it reasonable to analyze them in functional terms (cf. Introduction). Other *prima facie* examples of nonagentive norms can also be found, such as norms of attributive goodness, i.e., norms for what constitutes a good instance of a kind (Geach, 1956). These, too, have been identified with functional norms by some authors (e.g., Kelp, 2018). We cannot reject the idea of bearer norms out of hand, then, so let's take the possibility seriously.

I have already suggested one model on which to think of bearer normativity. A normative fact, I suggested in section 3.2, has a tripartite structure involving a subject, a conduct or behavior, and a normative relation linking them. For a fact to be normative *for bearers*, I suggest, would be for a bearer to take the place of the subject in a structure of this kind. Now, a proper-function fact, [*b* has proper function ψ], *does* involve a bearer *b* linked to a "conduct" ψ (on a suitably broad construal of the term "conduct"), but this, in and of itself, is clearly pitifully insufficient for normativity. The fact that my heart does *not* have the function of driving nails links my heart to the behavior of driving nails, but there's nothing normative about *that* fact.

Some have perceived a sufficient condition for normativity in the *possibility of malfunction*, i.e., the fact that it is possible for a bearer to have proper function even if it doesn't actually perform that function and even if it is incapable of performing it (due to injury or congenital defects). This feature of proper functions has taken center stage in discussions about its putative normativity.[6] Some have, indeed, suggested that this is *all* it consist in. So Justin Garson writes that by "normativity," he simply means that "it is logically possible for a trait token to have a function that it cannot, in fact, perform" (Garson, 2019, p. 5). The reason for this, I believe, is that anything that aspires to any kind of "normativity" must meet some such condition. Compare moral norms. It would be a strange moral norm indeed that applied to you only if you obeyed it. Norms, by their nature, seem to permit both conformity and violation.

But though this condition may be necessary for normativity, it is clearly not sufficient. *Being indifferent about ψ-ing* is consistent with not ψ-ing and with being incapable of ψ-ing, and so is *being such that if one ψ-ed, one would explode*. But there is nothing even remotely normative about either of these properties.

What is missing from these counterexamples is, precisely, any sense that the thing possessing one of these properties is thereby *called upon* to ψ and *precluded from* not ψ-ing (or vice versa). If I'm indifferent about ψ-ing, that neither calls for nor precludes me from ψ-ing. That's just a fact about me. If I would explode if I ψ-ed, that would seem to preclude me from ψ-ing, but that's not the case if I am, for instance, a dynamite stick. The counterfactual relation itself neither calls for nor precludes anything; if any calling-for or precluding is being done, it is done by independent norms governing the subject, like a prudential norm telling people not to explode.

Do proper functions possess this further feature? When a bearer *b* has a proper function ψ, is the bearer thereby *called upon* to ψ? Well, it sure seems like *something* of the sort is going on. There's a quite intuitive sense in which functioning constitutes the "good case," malfunction the "bad case." This is marked by the language we use to talk about functions. We call failure to perform a function *dys*function or *mal*function;[7] and, as we saw above, the bearer "is supposed to" or "ought to" perform its function.

---

6 Particularly in relation to the so-called "disjunction problem" in the theory of mental content (Fodor, 1984; cf. Neander, 1996).

7 However, it may not in general be true that failure to perform a function amounts to dysfunction or malfunction. If a woman's Fallopian tubes are blocked, her reproductive system will fail to perform its function, and so will its parts, e.g., the ovaries, since *their* function is (*inter alia*) to contribute to the functioning of the system as a whole. But it sounds strange to say that the ovaries *malfunction* or are *dysfunctional* in this scenario. If anything is dysfunctional, that would be the Fallopian tubes (Neander, 1995).

4.2. Bearer oughts

This is a start, but not a very helpful one. If someone is skeptical that proper functions are normative for bearers, they should not let themselves be swayed by the observation that we sometimes use the normative-sounding "ought" to attribute functions. "Ought," "should," and other so-called *weak necessity modals* have non-normative readings. Some think these verbs have an *epistemic* reading, roughly synonymous with "it is probable that." Others think they have a "pseudoepistemic" reading (more on which soon). If the proper-function ought belongs to one of these non-normative kinds, then it presumably has no bearing on whether proper functions are normative, for bearers or for anything else.

Now, as a matter of fact, I don't think that the proper-function ought is either epistemic or pseudoepistemic. Start with the epistemic ought. In our example above, the electrician says, "this circuit ought to give current to the bathroom lights." We can very well imagine that he says this while being fully aware that *at present*, the circuit does *not* give current to the bathroom lights (that's the problem he has been hired to fix). Thus, his claim cannot be paraphrased as "it is probable that this circuit gives current to the bathroom lights." So the proper-function ought is not an epistemic ought.

What about the pseudoepistemic ought? This is an invention of Seth Yalcin (2016), who defends it using examples similar to that of our electrician. On the pseudoepistemic reading of "ought," ⌜*s* ought to ψ⌝ means roughly ⌜It would be normal for *s* to ψ⌝; an idea that Yalcin further elaborates in these terms:

> Roughly [the normality reading] expresses something about what one is, or would be, entitled to expect to be true. (Yalcin, 2016, p. 245) [8]

Is this an accurate paraphrase of the proper-function ought? It seems more promising than the epistemic reading. Even if the electrician knows that the cicuit does not carry current, he might well be entitled to expect it to: *Normally*, that is what the current would do (cf. Franssen, 2009). But it does not seem that the oughts licensed by proper-function attributions in general permit this paraphrase. The function of the sperm is to fertilize the ovum, and so the sperm *ought* to fertilize the ovum. But we are not entitled to expect, of any given sperm, that it will fertilize an ovum.[9]

So the proper-function ought is neither epistemic nor pseudoepistemic. But it is not therefore normative. It could constitute a yet further non-normative kind of ought. The very fact that "ought" possesses non-normative readings means that this possibility remains open, even if none of the specific non-normative readings proposed in the literature hitherto fit the bill.[10]

How are we to tell whether the proper-function ought is indeed normative? I despair of finding a way to do so that is *independent* of the larger question we are trying to answer, viz., whether proper functions themselves are normative. If

---

If the normative-sounding notions of dys- and malfunction fail to overlap with function failure, this might be thought to spell trouble for the view that proper functions are bearer normative. But this reasoning is far from conclusive. For one, some have argued that precisely because the ovaries *do not* dysfunction, neither should they be ascribed any function that fails to be performed in the scenario at hand: only *proximal* effects qualify as proper functions (Fagerberg & Garson, forthcoming). Second, "failure" itself has a normative ring, so already our willingness to say that the ovaries *fail* to perform their function may be taken to imply that a norm has been violated.

Third, we need not suppose that the norm constituted by a proper function ψ has the simple content: ψ! The relationship between a proper function and the requirement it makes, the conduct it favors, may be more indirect. For instance, it may be that the proper function to ψ constitutes a norm saying: ψ unless unable due to the malfunctioning of cooperative systems!

8 Interestingly enough, this means that the pseudoepistemic ought is, in a sense, normative—but normative, not for *b* itself, but for *us*: It governs what is legitimate for us to expect.

9 I owe this point to Thomas Raleigh.

10 Parallel considerations could, I believe, be advanced against an attempt to establish the normativity of proper functions on the basis of the "supposed to" locution, the "dysfunction" locution, etc.

they are, we may perhaps conclude that proper-function ought is, as well. But of course, once we have established the former, we have answered the titular question—and any further conclusions about "ought" is a mere bonus.

4.3. Teleology

Let's recap. We are evaluating the claim that proper functions are normative for their bearers. This, I suggested, amounts to determining whether the proper-function relation is indeed a *normative relation*, one in virtue of which the bearer is in some sense "called upon" to perform its functions. How can this be shown?

Above, I criticized the notion that the normativity of proper functions derives from the *possibility of malfunction*. This was one of the two "general features" of proper functions introduced in section 2. It would make sense, then, to now turn to the second of the two—*teleology*, the idea that proper functions are *teloi* or ends for its bearers—and see if a case for bearer normativity can be made on *its* basis.

An end can, arguably, give rise to norms. Given an end of ours, there is a sense in which we ought to do what conduces to that end. This "teleological ought" manifests in such everyday expressions as

(1) If Peter wants to rob the bank, he ought to buy a gun on the black market.

In this case, the "end" is the intention or desire of an agent, and functions, it may be thought, are not ends in *that* sense.[11] But perhaps this is inessential, and other kinds of ends, including proper-functional ones, can also generate teleological oughts.

Whether this is plausible depends, in part, on how teleological oughts arise. On one influential view, teleological ought-facts like (1) obtain in virtue of a normative principle along the lines of

> Instrumental rationality.
>
> One ought to take the means to one's ends.

where the "ought" is the ought of rationality. Function-bearers are presumably not subject to norms of rationality (at least not in general), so if this is how teleological oughts arise, there is no reason to think that they are generated by proper-functional ends in addition to intentional ones.

It may, perhaps, be argued that though function-bearers are not subject to norms of rationality, they are governed by normative principles of an analogous sort. But even if so, there is a further problem. If teleological oughts are generated by some overarching normative principle akin to Instrumental rationality, then there is, arguably, nothing normative about the ends in themselves: They are then just *normatively relevant factors* that combine with a general normative principle to produce specific normative facts.[12] If so, then even if proper functions are ends, it does not follow that they are normative.

Instrumental rationality faces the *bootstrapping problem*: It seemingly permits agents to subject themselves to rational norms simply by adopting certain ends, regardless of how ill-advised the ends themselves are. So Peter, by forming the ill-advised intention to rob the bank, would impose on himself a rational obligation to buy a gun on the black market. In response to this problem, some have suggested that the teleological ought is not an ought of instrumental ra-

[11] On the other hand, according to teleosemanticists like Millikan (1984a, p. 140), the ends defined by desires *are* proper-functional ends, since on this view, a desire has the proper function to bring about its own satisfaction.

[12] Compare: It may be the case that one ought morally to maximize happiness and that φ-ing would maximize happiness. If so, then one ought morally to φ. But the fact that φ-ing would maximize happiness is not therefore itself a normative fact.

tionality, but an *end-relative* ought generated by the end itself. Stephen Finlay has probably taken this idea furthest, arguing that *all* oughts are end-relative:

> Every "ought" presupposes some particular end as framing a normative perspective, and recommends a course of action from that particular point of view as serving that end. (Finlay, 2009, p. 171)

Finlay's theory primarily concerns normative *language*, and if it has any implications for normativity as a worldly feature, they are deflationary: For it to be the case that *S* ought to φ, relative to some end *E*, is just for φ-ing to be conducive to *E*. If that is all normativity is—indeed, as long as this is *sufficient* for normativity—there is no particular reason why facts generated by proper functions about what is conducive to performance of them could not also be considered normative.

Arguably, however, there still would be nothing normative about the fact that *b* has proper function ψ, just as, even if Finlay is right, there is arguably nothing normative about the fact that Peter wants to rob the bank. Though these facts would have the power to generate normative facts, i.e., facts about what is conducive to the ends in question, they themselves remain purely descriptive facts about *b* and Peter.

Could this last claim be denied? Perhaps end-facts, in virtue of having the envisioned norm-generating power, themselves deserve to be called "normative"—in at least a derivative sense. If this principle is defensible, we have indeed found one way to argue for an affirmative answer to the titular question; to wit:

Argument from teleology

(1) End facts give rise to normative facts.
(2) In virtue of (1), end facts are themselves (derivatively) normative.
(3) Proper-function facts are end facts.
(4) So, proper-function facts are (derivatively) normative.

This, then, constitutes the first of the three arguments promised in the introduction. As I indicated there, the argument combines a general doctrine on the nature of normativity (comprising premises 1 and 2) with a feature specific to proper functions (premise 3).

4.4. Bearer normativity reduced

We have now considered the proper-function ought as well as the two "general features" of proper functions described in section 2 —*possibility of malfunction* and *teleology*. We have thereby, I fear, exhausted what can be accomplished relying exclusively on the pre-theoretical notion of proper function. If we want to say more, we need some *theory* of proper functions, capable of yielding us a more fine-grained idea of what they consist in and what they entail.

Fortunately, we are already familiar, since section 2, with two such theories: the propensity view and the etiological view. Now, both these views are controversial (indeed, they are rivals), and it would be preferable not to rest our investigation on controversial assumptions. Ideally, we would do it the other way around: *first* determine whether proper functions are normative, and then use this as a *datum* by which to test rival theories of proper functions. However, I don't see any way to proceed without bringing *some* additional theoretical assumptions into the picture, and the extant reductive theories is the best we have in this respect. Fortunately, there are independent ways to evaluate these theories, so we need not treat them as brute assumptions. Thus, in this section, I propose to consider the two views in turn and examine whether an argument for the normativity of proper functions can be constructed on their basis.

The resulting arguments will only extend so far as the reductive theories themselves. In particular, if you're a pluralist about proper functions (cf. end of section 2) and believe that different classes of proper functions yield to different reductive accounts, then you may need a different argument for each such class.

4.4.1. The propensity view

According to the propensity view, recall, a bearer has the proper function to ψ iff ψ-ing would lead to some specified outcome, like increased fitness or maintenance of the organism's structural integrity. Does this counterfactual property have anything normative about it? Is there a sense in which it "calls upon" the bearer to ψ?

Clearly, the sought-for normativity cannot inhere in the counterfactual structure, as such, of the property. For any outcome *O*—e.g., an increase in global well-being or global paperclip supply—we can define counterfactual properties that stand to *O* as propensity functions stand to their defining outcome (fitness, structural integrity); viz., the property of being such that if you ψ, *O* comes about. Call such properties "*O*-functions." Obviously, for most *O*s, the corresponding *O*-functions have nothing normative at all about them, nor is there, accordingly, anything normative about the fact that some entity has an *O*-function. The property of *being such that if you're dropped, you get closer to the Earth's center of gravity* is shared by most terrestrial objects, including humble rocks, but nobody would say that having this property "calls upon" the rock to be dropped, in any interesting sense. Accordingly, if propensity functions are normative, then their normativity must derive from the specific outcome in question (fitness, structural integrity).

Can these outcomes carry this weight? That seems uncertain, at best. On one hand, it is very tempting to say that fitness or the organism's structural integrity constitute "goods," in a way a rock's proximity to the Earth's center of gravity does not. It is, accordingly, tempting to say that performing one's propensity function guarantees that things go "well," while failure guarantees that things go "badly." This, one may think, amounts to a more substantial notion of normativity: When a bearer has a proper function, it is "called upon" to perform it in the sense that *if* it performs it, things go well.

Even if this is granted, it remains to be shown that possessing one of the corresponding counterfactual properties—being such that if you were to ψ, one of these goods would result—actually subjects the bearer to a norm. Now, on at least some normative views, at least some norms are such that if you were to conform to them, goods would result. For instance, according to utilitarianism (specifically, objective act utilitarianism), conformity to moral norms would increase global well-being, simply because morality, on this view, says to perform acts that would have this effect. Granted, utilitarianism is a first-order view, lacking direct implications for wherein the normativity of moral norms consists. Utilitarianism is consistent with it simply being a *normatively relevant* fact that an action would maximize global well-being, one that combines with independent moral principles to produce oughts (cf. fn. 12, above). Nevertheless, we can imagine a view according to which the fact that *S* ought morally to φ *just is* the fact that *S*'s φ-ing would maximize global well-being. On such a view, moral normativity *reduces* to the counterfactual propensity to yield a good.[13] Once this is granted, it may as well be granted that any counterfactual propensity to yield a good constitutes a normative fact in like manner. Given this, the propensity theorist has an argument for an affirmative answer to the titular question, the second of the three arguments promised in the introduction. It goes as follows:

Argument from propensity

(1) Proper-function facts are facts of the form: [If *b* were to ψ, (fitness/structural integrity/etc.) would result], (*The propensity view*)

(2) Any fact of the form [If *S* were to φ, a good would result] constitutes a normative fact whereby *S* is called upon to φ. (*Reductive thesis about normativity*)

[13] This view of normativity is a cousin of the end-relative view described in the previous section, but with *goods* in the place of *ends*.

(3) (Fitness/structural integrity/etc.) is a good.

(4) A proper-function fact constitutes a normative facts whereby *b* is called upon to ψ.

Now, the *reductive thesis about normativity* gets whatever initial plausibility it has from the case of global well-being. Global well-being is a *universal* or *objective* good—but fitness, structural integrity, etc. are, at most, goods for the bearer itself or, perhaps, for the enclosing organism. Clearly, if such relative goods are admitted unrestrictedly into *reductive thesis about normativity*, the result is wild overgeneration of norms: Any good for *anyone* will impose a norm on *anyone else*. We should thus restrict the reductive thesis so that a relative good imposes a norm only on those things *for which* it is a good.

The question is then whether fitness, structural integrity, etc. are goods for bearers. This seems questionable. Indeed, it is unclear whether bearers, who are not agents or experiential subjects, *have* goods to begin with. Defending this claim is the argumentative burden for whoever wishes to pursue the Argument from propensity.

4.4.2. The etiological view

The etiological view, recall, identifies proper functions with ancestral, fitness-conducive, heritable effects. There is no guarantee that what was fitness-conducive to past generations will continue to be so, nor, indeed, that it will have any particular consequences at all, good or bad. The Argument from propensity is therefore not available to the friend of the etiological view.

But the latter can look elsewhere for the would-be normativity of proper functions. For a bearer *b* to have proper function ψ is, according to the etiological theory, for ancestral episodes in which ψ was produced to contribute to *explaining the bearer's existence* as well as *its own propensity to* ψ (insofar as it has such a propensity)—since the bearer's ancestors have enjoyed preferential replication in virtue of these ancestral productions of ψ. In virtue of this, I believe, the etiological function stands to the bearer in a position quite similar to that in which at least some *social norms*, of the weakly normative kind discussed in section 3.3, stand to the people for whom they are in force.

In virtue of what does a social norm obtain? There may not be any *general* answer to this question; but for at least one important class of social norms, it seems their obtaining consists in part in the fact that some specific form of conduct has come to attain the status of a *model* on which further instances of the conduct have been patterned. The conduct has, in effect, exercised a *power to produce copies of itself*. Such copying has taken place in part due to a general human penchant for imitation, in part due to the effects of this behavior itself; e.g., in enabling coordination, in forestalling social aggression directed against nonconformists, etc. Whether this is true of Spanish law is arguable, but it seems an apt characterization of many less formal, unspoken social norms, like rules of etiquette or linguistic rules (cf. Millikan, 2005). Importantly, behaviors that merely have the potential to generate copies of themselves do not suffice for there to be social norm: it must be a historical fact that such copying has occurred.

Now, consider a specific subject, such as Ana. In virtue of what is a social norm of the above kind in force *for her*? There is a sense of "in force" in which a given social norm is in force for Ana only if it gives her (rational, moral) reasons to obey it, and another sense in which it is in force only if she reflectively endorses the norm and its authority over her, but these are not the senses of "in force" we are after. Suppose, for concreteness sake, that Ana is visiting a place where women, as a matter of custom, are required to cover their hair. Ana may not care about this norm (or even know about it), and it may not enjoy the backing of rationality or morality, but there is still a sense in which it is in force for her, simply because she is situated, as it were, within its "social force-field." What this seems to amount to is that Ana is within reach of the aforementioned power of the customary pattern of conduct, covering your hair as a woman, to produce copies of itself.[14] If Ana

[14] What it means to be "within reach" of this power is a matter of degree—as it should be, since whether a social norm is in force for a subject is intuitively also a matter of degree. The cover-your-hair norm is intuitively less in force for Ana than for a local woman, though Ana does not entirely escape its authority.

chooses nonconformity, she risks disapproval and consternation or, at best, being treated like a know-nothing tourist; while if she does conform, this stands to facilitate her interactions with the locals. This is true whether or not these factors give Ana genuine reasons to cover her hair, and whether or not she reflectively endorses them as reasons; and it will—*ceteris paribus*—increase the likelihood that she conforms, whether or not she ultimately does so. If she does choose to conform for any of these reasons, it will be due to the same mechanisms that were responsible for producing past instances of the normatively favored conduct and which continue to act on her in the present.

This structure has obvious parallels to etiological functions. When a heart has the etiological function to pump blood, it is because past blood-pumping hearts have exercised a power to produce copies of themselves—a power deriving from its superior fitness relative to other hearts lacking this capacity—just like Ana is under the influence of the hair-covering norm insofar as she is subject to *its* copy-producing power. In both cases, this copy-producing power may fail to manifest fully, producing an Ana with bare hair or a heart that doesn't pump blood. In Ana's case, this will typically trigger a process of social feedback which may lead future timeslices of Ana to have their hair covered, whereas in the heart's case, it will most likely just lead to its death[15]—but then, at least, it will leave room for other hearts that do pump blood. But none of these downstream effects are essential: Ana may in fact find her nonconformist behavior encouraged, and the non-beating heart could (through some contrived setup) turn out to be fitter than its beating cousins.

These parallels are suggestive, but is there reason to believe that they are *sufficient* to confer upon proper functions a normativity of the sort social norms enjoy? Though this is not the place for detailed argument, I believe there is indeed some grounds for optimism on this score. In essence, the idea would be that past copy-producing power licenses a pattern of *explanation* that seems characteristically normative. If Ana does end up covering her hair, a sufficient answer to the question *why* she did so seems to be *that it is a norm* to do so. If she doesn't, on the other hand, that requires special appeals: She didn't know about the norm, or didn't care. This explanatory pattern is in turn explained by the hypothesis that the norm *consists in* the historically copy-producing power of the hair-covering behavior: Current instances of the behavior exist because past instances have exercised that power. Meanwhile, nonconforming behaviors require separate explanation.

The explanatory pattern laid out above precisely parallels an explanatory pattern to which proper functions give rise: We can explain *why* an entity ψ-s by citing the fact that it has the function to ψ, while explaining why it doesn't ψ requires separate appeal. The etiological theory can readily account for this pattern.

This parallel provide a starting-point from which the etiologist could argue that proper functions are normative in something like the way social norms are. In abstract form, the argument might look like this:

> Argument from copy-producing power
>
> (1) For S to be under the influence of the copy-producing power of past instances of a conduct φ constitutes for S to be subject to a weak norm in favor of φ-ing. (*Reductive thesis about weak normativity*)
>
> (2) For b to have proper function ψ is for b to be subject to the past copy-producing power of past instances of b's ψ-ing. (*Implication of the etiological view*)
>
> (3) So, for b to have proper function ψ is for b to be subject to a weak norm in favor of ψ-ing.

This, then, is the last one of the three arguments promised in the introduction. It is impressionistic, but I believe it provides a promising starting point. There is nevertheless much work left to do. Paradigm cases of weak norms have further features for which there exist no parallel in the proper-function case. For instance, social norms are typically *endorsed* by

[15] This could also happen to Ana, provided that the norm in question is backed by sufficiently draconian policing.

many of those subject to them, reinforced by explicit codes and rules, and so on. An argument is needed that these further features are inessential to their normativity. Hopefully, the present discussion can inspire continued inquiry.

5. Agent norms

In the previous section, I looked at ways to argue that proper functions are bearer normative. Even if one of these arguments works, it may seem disappointing if bearer normativity is the *only* normativity that proper functions possess. Much of the interest philosophers have taken in the putative normativity of proper functions is due to the hope one may thereby find means to account for semantic norms, speech-act norms, and epistemic norms—all of which seem to apply to *agents*, *qua* speakers and knowers. One could argue that such norms are not really agentive at all, that they do not impose any kind of ought, reason, or obligation on agents but are themselves bearer norms governing, e.g., the language faculty or the belief-forming apparatus (e.g. Olson, 2018). But this type of view would itself be deflationary of the relevant kind of normativity and is bound to strike many as disappointing, if not implausible.

In the introductory quote, Papineau insists that proper functions do not entail any oughts for agents. This seems correct, at least if we have in mind oughts expressive of recognized forms of "strong" normativity, such as morality or rationality. There is room to quibble here, especially about the exact meaning of "entail." Clearly, proper functions can be morally or prudentially *relevant* to agents in various ways, as argued, e.g., by Christoph Kelp (2018). In other words, they can have implications for what agents ought morally or rationally to do, given a background of contingent facts and general normative principles. However, this presumably does not suffice to confer normativity on the proper functions themselves, since almost any fact can, under the right circumstances, be normatively relevant in this way. Let us therefore grant Papineau that proper functions do not "entail" moral or rational oughts in any sense strong enough to confer normativity on the former.

But is this the only way to establish that proper functions are normative for agents? When we, in the preceding section, discussed the *bearer* normativity of proper functions, we didn't approach the issue in terms of some independently characterized family of bearer-normative facts that would confer their normativity upon proper functions by being "entailed" by them. Rather, we directly interrogated the possible normativity of these proper-function facts themselves. In doing so, we did not assume that this would-be normativity had to reduce to moral, rational, or any other form of antecedently recognized normativity. We should extend to agents the same courtesy: If proper functions subject them to normativity, it may well be a *sui generis* form of functional normativity.

If this is so, then it seems proper functions would also entail oughts for agents—*sui generis* functional oughts, of the form discussed in section 4.2. For "ought" is a flexible word which lends itself readily to relativization to norms of all kinds, not only to morality or rationality, but also to etiquette, conventions, club rules, and so forth. If proper functions impose *sui generis* functional norms on agents, then we would be able to relativize "ought" also to these norms. But there is no reason to assume that those oughts must be independently ascertainable, i.e., discernible *before* we have established that proper functions subject agents to functional normativity. Compare: We learn of etiquette oughts *by* learning that agents are subject to etiquette norms.

If proper functions are normative for agents, we should expect them to involve the sort of tripartite structure discussed in section 3.2, with the agent related to a conduct by a normative relation whereby it is "called upon" to engage in that conduct. Now, in our standard format for referring to proper-function facts—"[b has proper function ψ]"—there figures no name of any agent that could be called upon to do anything. But even if agents do not themselves have functions, it is conceivable that proper-function facts sometimes situate agents in the tripartite structure of section 3.2. As I argued there, a fact may *involve* a subject and a conduct in the way characteristic of normativity even if that subject and conduct are not named in the sentence standardly used to identify the fact. Perhaps proper-function facts sometimes involve agents in this way.

Here's one way to argue that they do. There are many sub-agential function-bearers who perform their functions, in part, by prompting certain behaviors on the part of the enclosing agent. Certain hormones may have the function of making us respond aggressively to threats: They perform their function by prompting aggressive action in the face of threats. In such cases, though the agent is not the proper function's bearer, the proper function nevertheless consists in the production of an agent-level behavior. It is an *agent-mediated function*.

Even if proper functions are normative, it does not follow that agent-mediated functions are normative *for the mediating agent*: It could be that the normativity targets only the bearer, with no normative pressure on the agent to cooperate and mediate the function's performance. Some additional premise is needed.

The requisite extra premise might be found in the reasoning in section 4.4. First, suppose the Argument from propensity is right, and the bearer normativity of proper functions consists in their counterfactual propensity to yield goods (fitness, structural integrity). In the case of agent-mediated functions, the agent behavior constituting the mediation will itself evince the relevant counterfactual propensity and thus, by the *reductive thesis about normativity*, a norm will be imposed on the agent as well.

Second, suppose that the Argument from copy-producing power is right, and the bearer normativity of proper functions consists in bearers being within reach of the copy-producing power of ancestral function-performing bearers. In the case of agent-mediated functions, the mediating agent will also, and for the same reason, be within reach of the copy-producing power of ancestral mediating agents; and thus, by the *reductive thesis about weak normativity*, the agent will itself be subject to a weak norm.

Thus, an argument to the effect that proper functions are sometimes normative for agents may be in the cards, contingent on the success of the arguments in section 4.4.

6. Conclusion

In this paper, I have sought to formulate arguments in favor of the view that proper functions are normative. I have identified three potential such arguments: the Argument from teleology, the Argument from propensity, and the Argument from copy-producing power. These arguments each represents a possible way forward for whoever wishes to defend an affirmative answer to the titular question on the basis of an independently evaluable conception of normativity. Even those who hold that no such argument is needed may find interest in these arguments, insofar as they suggest ways in which proper functions resemble other, more paradigmatically normative phenomena.

If I personally had to pick one of the three arguments as most promising, it would be the Argument from copy-producing power. The other two arguments rely on reductive views about the nature of normativity in general, whereas this one relies only on a reductive view about weak social normativity—a *prima facie* more promising candidate for reduction. Again, however, I stress that the purpose of the present exercise has not been to conclusively defend these arguments, but to take preliminary stock of them and to identify their accompanying commitments as regards the nature of normativity.

There are some topics that I, for lack of space, have had to pass over in silence. One of these is evaluativity and the claim, defended by Christoph Kelp (2018), that proper functions ground norms of attributive goodness. I hope to return to this topic in future publications.

To close, some words about implications. I believe my discussion can contribute clarity to the debate on the normativity of proper functions. Of particular importance in this regard is the distinction between bearer normativity and agentive normativity, which I believe must be kept firmly in view in any future discussion of this topic. We have also seen that some "easy" arguments for the normativity of proper functions, like the appeal to the possibility of failure, are in-

adequate, and that what remains carries with it commitments of a much more substantial sort. This may in turn help us advance those debates in epistemology and philosophy of language where appeal to proper-functional normativity is made, by making clearer what work such an appeal can and cannot perform. Finally, by spelling out the general structure of normativity as such, as well as by identifying specific views on its nature and tracing their implications, my discussion can hopefully contribute to a better understanding of this fraught topic.

Acknowledgements

I owe thanks to Fabian Hundertmark, Stephen Mann, Manolo Martínez, Peter Schulte, and Abel Suñé; to audiences at the BIAP workshop in epistemic normativity, the Barcelona mental content workshop, and the Functions and Norms conference at the University of Luxembourg; as well as to two anonymous referees for Theoria.

REFERENCES

Bigelow, J., & Pargetter, R. (1987). Functions. *Journal of Philosophy*, *84*(4), 181-196. https://doi.org/10.2307/2027157

Boghossian, P. (2015). Rules, Norms and Principles: A Conceptual Framework. In M. Araszkiewicz, P. Banaś, T. Gizbert-Studnicki, & K. Płeszka (Eds.), *Problems of Normativity, Rules and Rule-Following* (pp. 3-11). Springer. https://doi.org/10.1007/978-3-319-09375-8_1

Burge, T. (2003). Perceptual Entitlement. *Philosophy and Phenomenological Research*, *67*(3), 503-548. https://doi.org/10.1111/j.1933-1592.2003.tb00307.x

Copp, D. (2024). *Ethical naturalism and the problem of normativity*. Oxford University Press. https://doi.org/10.1093/oso/9780197601587.001.0001

Fagerberg, H., & Garson, J. (forthcoming). Proper Functions Are Proximal Functions. *British Journal for the Philosophy of Science*. Forthcoming. https://doi.org/10.1086/731869

Finlay, S. (2009). Against All Reason? Scepticism about the Instrumental Norm. In C. R. Pigden (Ed.), *Hume on Motivation and Virtue* (pp. 155-178). Palgrave Macmillan. https://doi.org/10.1057/9780230281158_8

Fodor, J. (1984). Semantics, Wisconsin style. *Synthese*, *59*(3), 231-250. https://doi.org/10.1007/bf00869335

Franssen, M. (2009). The Inherent Normativity of Functions in Biology and Technology. In U. Krohs & P. Kroes (Eds.), *Functions in Biological and Artificial Worlds: Comparative Philosophical Perspectives*. MIT Press.

García-Carpintero, M. (2022). How to Understand Rule-Constituted Kinds. *Review of Philosophy and Psychology*, *13*(1), 7-27. https://doi.org/10.1007/s13164-021-00576-z

Garson, J. (2019). *What biological functions are and why they matter*. Cambridge University Press.

Geach, P. T. (1956). Good and Evil. *Analysis*, *17*(2), 33-42. https://doi.org/10.2307/3326442

Gibbard, A. (1990). *Wise choices, apt feelings: A theory of normative judgment*. Clarendon Press.

Gibbard, A. (2003). *Thinking how to live*. Harvard University Press.

Glüer, K., & Pagin, P. (1998). Rules of Meaning and Practical Reasoning. *Synthese*, *117*(2), 207-227. https://doi.org/10.1023/A:1005162503125

Glüer, K., & Wikforss, Å. (2009). Against Content Normativity. *Mind*, *118*(469), 31-70. https://doi.org/10.1093/mind/fzn154

Godfrey-Smith, P. (1994). A Modern History Theory of Functions. *Noûs*, *28*(3), 344-362. https://doi.org/10.2307/2216063

Graham, P. J. (2012). Epistemic Entitlement. *Noûs*, *46*(3), 449-482. https://doi.org/10.1111/j.1468-0068.2010.00815.x

Griffiths, P. E. (2016). Proximate and Ultimate Information in Biology. In M. Couch & J. Pfeifer (Eds.), *The Philosophy of Philip Kitcher* (pp. 74-97). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199381357.003.0004

Hattiangadi, A. (2006). Is Meaning Normative? *Mind & Language*, *21*(2), 220-240. https://doi.org/10.1111/j.0268-1064.2006.00312.x

Jarvis, B. W. (2012). Norms of intentionality: Norms that don't guide. *Philosophical Studies*, *157*(1), 1-25. https://doi.org/10.1007/s11098-010-9610-4

Kelp, C. (2018). Assertion: A Function First Account. *Noûs*, *52*(2), 411-442. https://doi.org/10.1111/nous.12153

Kripke, S. (1984). *Wittgenstein on rules and private language: An elementary exposition*. Harvard University Press.

McCormick, M. S. (2014). *Believing against the evidence: Agency and the ethics of belief*. Routledge.

McPherson, T. (2011). Against quietist normative realism. *Philosophical Studies*, *154*(2), 223-240. https://doi.org/10.1007/s11098-010-9535-y

Millikan, R. G. (1984a). *Language, thought, and other biological categories: New foundations for realism*. MIT Press.

Millikan, R. G. (1984b). Naturalist Reflections on Knowledge. *Pacific Philosophical Quarterly*, *65*(4), 315-334. https://doi.org/10.1111/j.1468-0114.1984.tb00233.x

Millikan, R. G. (2005). Language Conventions Made Simple. In *Language: A Biological Model* (pp. 1-23). Oxford University Press.

Mossio, M., Saborido, C., & Moreno, A. (2009). An Organizational Account of Biological Functions. *The British Journal for the Philosophy of Science*, *60*(4), 813-841. https://doi.org/10.1093/bjps/axp036

Nanay, B. (2010). A Modal Theory of Function. *The Journal of Philosophy*, *107*(8), 412–431. https://doi.org/10.5840/jphil2010107834

Neander, K. (1991). The Teleological Notion of 'Function'. *Australasian Journal of Philosophy*, *69*(4), 454-468. https://doi.org/10.1080/00048409112344881

Neander, K. (1995). Misrepresenting & Malfunctioning. *Philosophical Studies*, *79*, 109-141. https://doi.org/10.1007/BF00989706

Neander, K. (1996). Swampman Meets Swampcow. *Mind & Language*, *11*(1), 118-129. https://doi.org/10.1111/j.1468-0017.1996.tb00036.x

Olson, J. (2018). Moral and Epistemic Error Theory: The Parity Premise Reconsidered. In C. McHugh, J. Way, & D. Whiting (Eds.), *Metaepistemology* (pp. 107-121). Oxford University Press. https://doi.org/10.1093/oso/9780198805366.003.0007

Papineau, D. (1999). Normativity and judgment. *Aristotelian Society Supplementary Volume*, *73*(1), 17-43. https://doi.org/10.1111/1467-8349.00047

Parfit, D. (2011). *On what matters*. Oxford University Press.

Plantinga, A. (1993). *Warrant and proper function*. Oxford University Press.

Schulte, P. (2020). Why mental content is not like water: Reconsidering the reductive claims of teleosemantics. *Synthese*, *197*(5), 2271-2290. https://doi.org/10.1007/s11229-018-1808-6

Searle, J. R. (1995). *The Construction of Social Reality*. Free Press.

Shah, N. (2003). How Truth Governs Belief. *The Philosophical Review*, *112*(4), 447-482. https://doi.org/10.1215/00318108-112-4-447

Simion, M. (2019). Knowledge-first functionalism. *Philosophical Issues*, *29*(1), 254-267. https://doi.org/10.1111/phis.12152

Wedgwood, R. (2002). The Aim of Belief. *Philosophical Perspectives*, *16*, 267-297. https://doi.org/10.1111/1468-0068.36.s16.10

Williams, B. (2006). Deciding to believe. In *Problems of the self: Philosophical papers 1956-1972* (pp. 136-151). Cambridge University Press.

Wright, L. (1973). Functions. *The Philosophical Review*, *82*(2), 139. https://doi.org/10.2307/2183766

Yalcin, S. (2016). Modalities of normality. In N. Charlow & M. Chrisman (Eds.), *Deontic Modality* (pp. 230-255). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198717928.003.0009

Karl Bergman is currently a postdoc within the *Thought Sharing in Action* project at Pompeu Fabra University. His research interests include proper functions, mental content, meaning and communication, and metaethics.

Address: Universitat Pompeu Fabra, Departament d'Humanitats, Edifici Jaume I, Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain. E-mail: mail@karlbergman.com – ORCID: 0000-0001-9635-6762

Theoria, 2026, 41(3), 367-377
https://doi.org/10.1387/theoria.28172

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Millikan on picturing and correspondence

(Millikan sobre figuración y correspondencia)

Esteban Withrington*
UNAM
https://orcid.org/0000-0002-5350-1402

Keywords

Ruth Millikan
Correspondence theory of truth
Picturing
Isomorphism
Teleosemantics
Sentential truth

ABSTRACT: Ruth Millikan characterizes her account of truth as a pictorial theory inspired by Tractarian ideas of mapping and isomorphism. Given the well-known difficulties facing pictorial accounts of sentential truth, this characterization is puzzling. I argue that Millikan's account is best developed without requiring relations of structural resemblance between sentences and reality. Focusing on her notion of semantic mapping functions and her later distinction between projected and substitutional correspondence, I suggest that the relations relevant to sentential truth are best characterized in terms of systematic correlation rather than resemblance. While some representational systems may involve genuinely projective relations, Millikan's own treatment of natural-language sentences as governed primarily by substitutional rules points, despite her Tractarian rhetoric, toward a non-pictorial form of correspondence.

Palabras clave

Ruth Millikan
Teoría correspondentista de la verdad
Figuración
Isomorfismo
Teleosemántica
Verdad de las oraciones

RESUMEN: Ruth Millikan caracteriza su teoría de la verdad como una teoría pictórica inspirada en ideas tractarianas de mapeo e isomorfismo. Dadas las conocidas dificultades que enfrentan las teorías pictóricas de la verdad de las oraciones, esta caracterización resulta desconcertante. Sostengo que la teoría de Millikan puede elaborarse mejor sin exigir relaciones de semejanza estructural entre las oraciones y la realidad. Centrándome en su noción de funciones semánticas de mapeo y en su distinción posterior entre correspondencia proyectada y correspondencia sustitucional, sugiero que las relaciones relevantes para la verdad de las oraciones se caracterizan mejor en términos de correlación sistemática que de semejanza. Si bien algunos sistemas representacionales pueden involucrar relaciones genuinamente proyectivas, el propio tratamiento que Millikan da a las oraciones de los lenguajes naturales como gobernadas principalmente por reglas sustitucionales apunta, pese a su retórica tractariana, hacia una forma no pictórica de correspondencia.

Gako-hitzak

Ruth Millikan
Egiaren korrespondentziaren teoria
Figurazioa
Isomorfismoa
Teleosemantika
Esaldien egia

LABURPENA: Ruth Millikanek mapatzearen eta isomorfismoaren ideia traktarioetan inspiratutako teoria piktoriko gisa karakterizatzen du egiari buruzko bere teoria. Esaldien egiaren teoria piktorikoek dituzten zailtasun ezagunak kontuan hartuta, karakterizazio hori harrigarria da. Nik defendatzen dut Millikanen teoria hobeto garatzen dela esaldien eta errealitatearen arteko egiturazko antzekotasunaren beharrik gabe. Millikanen mapatze-funtzio semantikoen noziotik eta berak gerora korrespondentzia proiektatuaren eta ordezkoaren artean egindako bereizketatik abiatuta, proposatzen dut esaldien egiari dagozkion erlazio esanguratsuak hobeto karakterizatzen direla korrelazio sistematikoaren bidez, antzekotasunaren bidez baino. Irudikapen-sistema batzuetan erlazio benetan proiektiboak egon badaitezke ere, Millikanek mintzaira naturaletako esaldiei ematen dien tratamenduak berak, esaldiok nagusiki ordezkatze-arauek gobernatuak direla iritzirik, korrespondentzia-molde ez-piktoriko bat seinalatzen du, bere erretorika traktarista gorabehera.

* ***Correspondence to:*** Esteban Withrington. Institute for Philosophical Research, National Autonomous University of Mexico, Circuito Mario de la Cueva s/n, Ciudad Universitaria, Mexico City, 04510, Mexico. – ewithrington@filosoficas.unam.mx – https://orcid.org/0000-0002-5350-1402

How to cite: Withrington, Esteban (2026). «Millikan on picturing and correspondence»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 41(3), 367-377. (https://doi.org/10.1387/theoria.28172).

Received: 10 January, 2026; Final version: 15 August, 2026.
ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press

This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. Introduction

The canonical version of the correspondence theory of truth has been largely abandoned for good reasons. This version relied on alleged relations of *resemblance* between truth-bearers and facts. One of the boldest attempts to clarify what such relations involved was perhaps Wittgenstein's *Tractatus Logico-Philosophicus* (1922)—where the correspondence relation was portrayed as involving (1) a relation of structural resemblance between a true sentence and a fact, and (2) referential relations between the words of the sentence and the components of the fact. The serious difficulties faced by this view have been widely discussed in the literature and account for why it has been largely abandoned. However, Tractarian themes have resurfaced in the literature. Sellars, for example, endorsed a picture theory according to which "the manner in which the names occur in the picture is a projection, in accordance with a fantastically complex system of rules of projection, of the manner in which the objects occur in the world" (1963, p. 215). More recently, Millikan has pursued, as she puts it, "the picturing themes from the *Tractatus* that were carried through in Sellars' discussions" (2005, p. 77). According to Millikan, "what makes a sentence true is that there is something in the world onto which it maps in accordance with certain mapping functions" (1984, p. 9). She clarifies that what she means by 'mapping' is, basically, "as Wittgenstein put it, 'picturing'" (2005, p. 63). "The value of a system of representation", she claims, "depends... on there being some kind of isomorphism... between the domain of the signs and the domain of their signifieds" (2004a, p. 84). Given the well-known problems faced by the pictorial account of sentential truth, these confident claims are puzzling. If taken strictly, Millikan's Tractarian rhetoric commits her to a problematic pictorial account of such truth; if assessed in terms of the work her semantic mapping functions are supposed to do, however, her account can be developed in terms of systematic correlation rather than resemblance. Millikan does not provide detailed mapping rules for sentences of any specific form, so it remains unclear whether her semantic mapping rules themselves require genuinely pictorial relations. In this paper, I argue that Millikan's pictorial presentation of her view is misleading, at least in this domain, and that the relevant relations are best characterized within the framework of a contemporary correlation-based account of truth. The relations between true representations and reality that a plausible development of Millikan's account requires are, at least for sentences and similarly structured representations, of the sort that follows from a Tarski-based characterization of the truth of a sentence in terms of its syntactic structure and the referential properties of its words, combined with a reductive account of reference in terms of substantive relations between words and parts of reality. Millikan's teleosemantic framework is one of the main contemporary contenders for explaining the nature of such substantive relations, by grounding the reference-determining links between representations and reality in the functions that representations have acquired through natural selection, learning, or cultural inheritance—an approach she explicitly applies to account for the referential properties of words. I will set aside other objections to Millikan's theory in order to focus exclusively on her claim that it involves picturing relations.

My argument is related to, but distinct from, objections raised by Godfrey-Smith (1996) and Shea (2013) concerning the explanatory role of isomorphism in Millikan's theory. Their discussions focus on relatively simple representational systems lacking the logico-syntactic structure of sentences—such as inner states representing mates or predators, or bee dances representing the location of nectar—rather than on sentential truth. Moreover, their concern is primarily that isomorphism is too liberal to do the work Millikan assigns to it. By contrast, my concern is that, for sentential truth, genuine structural resemblance would be too demanding, while a Tarski-based correlation account can do the needed work.

2. Correspondence, Picturing, and Correlation

This section is concerned with the nature of correspondence relations appealed to in accounts of truth, focusing on the contrast between correspondence understood as structural resemblance and correspondence understood as systematic correlation. What today is regarded as the canonical version of the correspondence view is the claim that a truth-bearer is true if and only if it *corresponds to a fact*. This idea was advanced in the early twentieth century by Moore, Russell and

Wittgenstein. The canonical view requires an ontology of facts. Moore, Russell and Wittgenstein argued that facts are entities in their own right, existing over and above the things that compose them. The basic idea is that facts are structured combinations of objects with properties or relations. For example, the fact that a certain tree is an oak allegedly consists of an object, the tree, combined with the property of being an oak (Moore, 1953, p. 311), while the fact that Desdemona loves Cassio allegedly is a "complex unity" of two objects, Desdemona and Cassio, connected by the loving relation which "binds [them] together" in a particular "order" (Russell, 1912, p. 128). The canonical view also needs an account of the correspondence relation. Wittgenstein and Russell proposed that a true sentence "pictures" or "mirrors" a fact in virtue of (1) the *referential* relations between its words and the constituents of the fact, and (2) a relation of *structural resemblance* or *isomorphism* between the sentence and the fact (Wittgenstein, 1922, 2.131 and 3.21; Russell, 1918/2010, pp. 24-25). For example, 'Socrates is wise' is true if there is a fact consisting of the object and property referred to by the name and predicate, respectively, which has what is taken to be the same monadic structure, Φx, as the sentence. Similarly, 'Napoleon hates Wellington' is true if there is a fact that has what is taken to be the same dyadic form as the sentence, Φxy, and which has Napoleon and Wellington and the hating relation as constituents.

While very influential in its time, the canonical version of correspondence faces serious difficulties. As Austin warned, the idea that portions of reality are *isomorphic* to the true sentences that represent them seems to commit "the error of reading back into the world the features of language" (1950, p. 119). Austin proposed that only correlations (not isomorphisms) between sentences and portions of reality are required for correspondence (1950, pp. 115-120). Moreover, it turns out that sentences are *not* isomorphic to what they represent, as Russell himself noticed:

> In the phrase "A is to the left of B," even if we treat "is-to-the-left-of" as one word, we have a fact consisting of *three* terms with a *triadic* relation, not two terms with a dyadic relation. The linguistic symbol for a relation is not itself a relation, but a term as solid as the other words of the sentence. (1919, p. 38)

The fact represented by a true sentence *'Rab'* allegedly is a unity of two objects, a and b, bound by a dyadic relation *R*. But the sentence is instead a unity of three words, *'R'*, *'a'* and *'b'*, bound by a triadic logico-syntactic relation. If there is such a thing as the fact that *Rab*, it is not isomorphic to the sentence *'Rab'*.

Given the serious problems faced by the canonical version of the correspondence theory, contemporary theorists have articulated a version of the view that relies neither on facts nor on resemblance relations. Instead, it relies on (1) an account of truth in terms of the logical structures of sentences and the referential properties of their parts—based on Tarski's (1944, 1956) influential work—coupled with (2) a reductive account of reference in terms of one or another kind of causal relation between words and objects—a combination which, Field (1972) argued, is necessary to provide a robust account of truth. While the canonical view regards correspondence as a resemblance-to-facts relation, the contemporary view treats correspondence as a correlation-to-objects relation.

Consider a simple sentence *'Kr'*. Following the procedure proposed by Tarski, we can specify the truth-conditions of *'Kr'* as follows:

> *'Kr'* is true ↔ $\exists x$(*'r' designates x* & *'K' applies to x*).

In plain English: *'Kr'* is true if and only if there is an object that the name *'r'* designates and the predicate *'K'* applies to. Notice that the truth-conditions of the sentence are characterized in terms of its structure and the referential relations—application and designation—between its component words and an object in the world. Sentential structure and referential relations were also components of the canonical view. But here "facts" and "structural resemblances" *play no role*. These are correlation-to-objects relations instead of resemblance-to-facts relations. Given the difficulties faced by the canonical view, related to "structural resemblance", this is a clear advantage.

Many authors have expressed skepticism about the existence of facts or states of affairs, and have instead favored object-based formulations of correspondence. While this is an important metaphysical issue, nothing in the present discussion turns on

whether correspondence is formulated in object-based or fact-based terms. The crucial contrast, for present purposes, is between correspondence understood as structural resemblance and correspondence understood as systematic correlation. Millikan herself endorses a fact- or state-of-affairs-based framework, but her account of truth can be developed without requiring resemblance relations. While Tarskian truth-rules are standardly formulated in object-based terms, an entirely parallel, non-pictorial formulation can be given in fact-based terms, as the rule *SP* discussed below illustrates. Conversely, a Millikanian who wished to avoid commitment to facts could equally adopt strictly object-based rules of the sort favored by Field (1972), without affecting the argument that follows. I contend that, despite its pictorial rhetoric, Millikan's account of truth is best developed without requiring relations of structural resemblance—at least for sentences and similarly structured representations—and can be reconstructed, in this domain, as invoking systematic correlation between representations and reality.

3. *Millikan's Pictorial Account and Teleology*

Millikan combines what she presents as a pictorial account of truth with a teleological account of the functions performed by representations. On her view, mapping relations or isomorphisms between representations and reality are the core features that make representations true, while representational mechanisms have functions acquired through natural selection, learning, or cultural inheritance that explain why representations are supposed, but may fail, to stand in those relations (Godfrey-Smith, 1996, pp. 184-185; Shea, 2013, p. 63). The same teleological framework is also meant to explain how the parts of representations acquire referential properties. Her naturalist account of reference is among the main contemporary candidates for reducing referential properties to substantial non-semantic relations. Earlier pictorial theories typically assumed referential links between words and worldly items without providing a naturalistic account of them. Millikan's view is complex and the role of picturing in her account is often overlooked, because most of the interest has been on her theory of reference.

Notable exceptions are Godfrey-Smith (1996) and Shea (2013), who are sympathetic to some aspects of Millikan's program but skeptical about the role of isomorphism in her account, at least as Millikan articulates it. Godfrey-Smith raises the general worry that *mapping* is too unconstrained to serve as a proto-semantic relation or substantive constraint antecedent to the teleological story. Shea develops a related point in terms of the multiplicity of possible isomorphisms: since many isomorphisms can be found between representations and represented affairs, isomorphism by itself cannot serve as an independent substantive constraint on content. For Shea, however, this is not an objection to Millikan's teleosemantics, since the relevant isomorphism can be *delivered* by the teleosemantic story rather than fixed independently of it. Their shared concern is that isomorphism is too liberal to play the independent explanatory role Millikan seems to assign to it (Godfrey-Smith, 1996, pp. 184-187; Shea, 2013, pp. 63-80).

My concern here is different. I am not concerned with whether the teleological part of the theory relies on an antecedent mapping relation or instead delivers the relevant mapping. My question is what sort of relation this mapping is supposed to involve. If it is understood as genuine structural resemblance, then it is too demanding for sentential truth; if it is understood as systematic pairing, then it is a form of correlation rather than picturing. Millikan intends her theory to apply to all sorts of representations, from human representations to the most basic cases of non-human representation. Whereas Godfrey-Smith and Shea focus primarily on simpler representational systems, I focus primarily on sentences of natural human languages, while drawing on non-sentential cases where they help to clarify the issues at stake.

4. *Mapping Rules and Correlation*

According to Millikan, sentences "correspond to states of affairs in accordance with semantic-mapping functions" (2005, p. 63). The notion of "function" here is the mathematical one (2005, p. 53), rather than the teleological no-

tion that she deploys in other parts of her theory. Semantic-mapping functions, she argues, are rules that determine the truth-conditions of sentences:

> The semantic-mapping function is given by rules according to which significant transformations of the sentence that conserve its syntactic form yield different truth- or satisfaction-conditions... 'It's raining' contrasts with 'It's snowing', 'It's hailing', 'It's sleeting', and so forth. All display the same syntactic form, the transformations that substitute 'snow', 'hail', and 'sleet' for 'rain' determining different satisfaction-conditions in a systematic way. (2005, pp. 63-64)

Substitutions of component parts of a sentence of a specific syntactic form, she claims, change its truth-conditions in a systematic way. So, for example, *'Fa'* can be transformed into *'Fb'*, *'Ga'*, *'Gb'*, etc., which have the same syntactic structure but represent different states of affairs. The difference regarding what portions of reality each of these sentences "maps" onto is determined exclusively by the different referential properties of their components. To the extent that mapping rules determine what states of affairs are represented by *transformations* of sentences of the *same* form—which is what Millikan suggests in the quoted passage—the "mapping" determined by these rules has nothing to do with the traditional notion of structural resemblance between sentences and facts. Of course, the mapping rules that determine the truth-conditions of *'Rab'* and its transformations—*'Rba'*, *'Scd'*, *'Sdc'*, etc.—will be different from those governing 'Fa' and its transformations. It is clear that Millikan thinks that mapping functions relate each sentence to a state of affairs that would make it true, and that this relation depends on (1) the syntactic structure of the sentence and (2) the referential properties of its words. These two features are shared by pictorial theories and by contemporary non-pictorial accounts of correspondence. The latter do not typically posit facts or states of affairs, but a variant of them may nonetheless do so without relying on picturing relations. The questions are whether Millikan's "mapping" rules are additionally *intended* to relate sentences to states of affairs that have the very same structures as the sentences, and whether this additional commitment is actually *required* by her account of truth-conditions in terms of "mapping" rules.

Millikan does not explicitly formulate mapping rules for sentences, so it is hard to assess whether her account requires specifically pictorial commitments. Consider the following candidate non-pictorial mapping rule for simple subject-predicate sentences:

> *SP: The sentence ⌜Φx⌝ is true if and only if the state of affairs in which the object referred to by ⌜x⌝ has the property referred to by ⌜Φ⌝ obtains.*

The rule *SP* is compatible with Millikan's claim that the *relata* of sentences are states of affairs and her professed Aristotelian realism about properties (1984, Chs. 15 and 16). Also, *SP* covers all the substitution instances of ⌜*Φx*⌝—i.e., *'Fa'*, *'Fb'*, *'Ga'*, etc.—assigning a specific state of affairs as the truth-maker for each instance. It yields in a *systematic way* the truth-conditions for all the *transformations* of *'Fa'*. Assuming that these corresponding states of affairs have the same ontological structure—that of an object having a monadic property—*SP pairs* or *correlates* sentences that have a specific syntactic structure with states of affairs that have a specific ontological structure. *SP* delivers what Millikan calls "a correspondence by a given rule or function between form and some structure in the world" (2005, p. 98). Everything that Millikan seems to require of a "semantic-mapping rule" is met by *SP*. The crucial point is that *SP* does not require the combination of a name with a monadic predicate to structurally resemble the combination of an object with a monadic property. If *SP* is the kind of "semantic-mapping rule" that Millikan has in mind, then hers is a correlation-based rather than a resemblance-based correspondence view. What is puzzling is Millikan's talk of "mapping" as a Tractarian form of "picturing" involving an "isomorphism" between representations and reality, which may suggest that she intends something stronger than a merely correlational rule such as *SP*. A correlation account may preserve talk of correspondence, but it does not need the stronger claim that representations picture what they represent.

5. *Bee Dances, Isomorphism and Projection*

Millikan's favorite example of representation is the figure-eight waggle dance made by honey bees to communicate to each other the location of nectar—which she often compares to human sentences. The ethologist von Frisch discovered that waggle dances communicate the location of nectar and that they do this by representing its direction and distance from the hive: the angle of the waggle dance with respect to the vertical axis of the hive represents the angle from the line between the hive and the Sun where nectar is located, while the duration of the dance represents how far away from the hive the nectar is (1953, Ch. 11). According to Millikan, the bee dance "is a tiny map" that "maps the location of nectar by a certain rule of projection" (2005, p. 83). "Variations in possible bee dances", she points out, "correspond one-to-one to variations in possible locations of nectar" and "the principle involved", she claims, "is mathematical isomorphism" (2005, p. 97). The way she describes how the "bee language"—as von Frisch called it—represents the location of nectar has a distinctly Tractarian flavor. It is striking that bee dances do represent the location of nectar by means of mathematical rules of projection. A waggle dance x represents the presence of nectar in direction y at distance z from the hive. There is a direct mathematical projection between the angle of x and the direction y, and another one between the duration of x and the distance z. Mathematical rules of projection determine the values of y and z. Whether this makes the whole dance "isomorphic" to the location of nectar is not so clear and I will not discuss it. But what is this supposed to tell us about representations in general and about human sentences in particular?

Bee dances are Millikan's favorite example because she thinks that they illustrate basic features of representations in general. Here is how she explains what "mapping functions" do in any system of representation: "Semantic mapping functions define isomorphisms between the set of possible signs in a certain sign domain and the set of their possible signifieds" (2004a, p. 49). This is meant to apply also to the linguistic and mental representations of human beings, although she warns that we should "expect the... mapping functions to be much more abstract" in the human case (2005, p. 102). Furthermore, she insists that she uses the term 'isomorphism' in "the abstract mathematical sense" (2004a, p. 84). In their pictorial period, Wittgenstein and Russell also intended true sentences to be isomorphic to facts only in an abstract sense. They intended only the underlying logical form of a true sentence to resemble the structure of what it represents, independently of its superficial structure which may be different. If this were the kind of abstract isomorphism that Millikan expects to obtain between true sentences and what they "map" onto, however, her theory would be undermined by the problem discovered by Russell: that a sentence *'Rab'* would represent a fact containing two objects united by a dyadic relation, while the sentence contains three words united by a triadic logico-syntactic relation. Despite her talk of "picturing", "mapping", "rules of projection" and "isomorphisms", however, Millikan does not say enough to determine whether, in the case of sentences, her mapping rules require genuine structural resemblance rather than merely systematic correlation, since she provides no explicit mapping rule for sentences of any particular form.

Recall Millikan's claim that, in human languages, the mapping rules systematically determine for different sentences different states of affairs as truth-makers (2005, pp. 63-64). This is compatible with a correlation-based view of correspondence. Recall her claim that mapping rules determine a correspondence between the form of a true sentence and some *structure* in the world (2005, p. 98). This claim also falls short of being genuinely pictorial. As I argued above, the rule *SP* I suggested would determine a correspondence between a sentence of the *form* $\ulcorner \Phi x \urcorner$ and a state of affairs with a specific structure—that of an object having a monadic property—regardless of whether the sentence's form matches that structure. This independence from structural resemblance is an advantage, since, as Russell showed, the sentence's form does not match the structure of the corresponding state of affairs.

With this in mind, consider the following passage, where Millikan hints at a way of understanding mapping rules for sentences and similarly structured representations that would not require their structural resemblance to facts:

> We can think of beliefs as... mental sentences... and think of... their truth makers... as determined by some sort of Tarskian mapping... Certain substitutions of sentence parts correspond to substitutions in aspects of world affairs mapped... That is, there is a functional isomorphism between

> the whole set of possible mental sentences... and the set of world affairs to which these sentences would correspond... (2004b, pp. 231-232)

This passage should be read with caution. Millikan speaks only of "some sort of Tarskian mapping", and she does not provide an explicit mapping rule for any sentence. Nor does she further develop this Tarskian suggestion here or elsewhere or explain whether such a rule would be non-pictorial or would instead somehow involve structural resemblance between sentences and facts. The passage is nevertheless revealing, because it gestures toward a way of developing mapping rules for sentential representations that would not require such resemblance. The rule *SP* I formulated above is a candidate development of this line of thought: a "sort of Tarskian mapping" for simple predicative sentences that is compatible with Millikan's realism about properties and structured states of affairs (or facts), while not requiring sentences to picture their truth-makers. Analogous rules would be required to cover other syntactic structures. What those rules would determine is a *systematic pairing* between possible sentences and the states of affairs that would make them true. This seems to be what Millikan in the passage calls—rather confusingly, in my opinion—a "functional isomorphism". Much, though not all, of Millikan's talk of isomorphism seems to concern this system-level pairing between possible representations and possible represenateds. Such a pairing may amount to systematic correlation rather than structural resemblance. In any case, system-level isomorphism does not by itself entail that individual sentences structurally resemble the states of affairs that make them true. Millikan's explicit appeal to Tractarian picturing nevertheless invites that stronger reading, which is why the issue remains unsettled. For a fact-based view of correspondence as correlation, what matters is that each sentence is paired by such rules with the state of affairs that would make it true. That is less than what a genuinely pictorial theory requires. Incidentally, notice that the set of all instances of what is now commonly called the "equivalence schema"—*'p'* is true ↔ *p*—also seems to provide a systematic pairing between sentences and what they represent (*'p'* is true ↔ *p*, *'q'* is true ↔ *q*, *'r'* is true ↔ r, etc.). But this "pairing" does *not* itself state a relation of correspondence.[1] A correlation-based *correspondence* theory requires the systematic pairing to be determined by substantial relations between sentences and truth-makers.

6. *Projected vs. Substitutional Correspondence*

All of this, however, leaves open the possibility that Millikan's talk of relations of "picturing" between sentences and their truth-makers may require actual structural resemblance. The issue is not settled by the fact that her mapping rules need not be read pictorially, since her Tractarian talk and her comparison between sentences and bee dances suggest that she may intend something stronger. However, Millikan's later distinction between "projected correspondence" and "substitutional correspondence" (2013, pp. 84-86) helps to clarify this issue. This distinction is introduced in response to an objection by Shea. I will provide some context before explaining the distinction and discussing what it shows about whether Millikan's account requires true sentences to structurally resemble states of affairs.

In his objection to Millikan's isomorphism requirement, Shea suggests an interesting hypothetical scenario. Suppose that the bee rules for representing distance were as follows: *1 waggle = 75 m*, *2 waggles = 300 m*, *3 waggles = 60 m*, *4 waggles = 150 m, etc.* (2013, p. 74). Of course, these are very different from the actual rules, discovered by von Frisch, relating bee dances to the distance of nectar. But in principle—while very unlikely—some species or subspecies of bee could have evolved to have dances which represent by Shea's rules, provided their foraging range was limited, so that a finite list

[1] As Douglas Patterson points out, many philosophers—notably Tarski (1944, pp. 342-343)—have held the view that instances of the schema "state correspondences between the sentences they mention and something else" (2003, p. 421). But instances of the schema do not state relations of correspondence, since they do not state any relations whatsoever: their basic form is *A* ↔ *B*, rather than *Rab*. Indeed, deflationists argue that there is no need to regard the word 'truth' as standing for a relational property because instances of the equivalence schema are all that is required to account for its meaning and utility (Field, 1994; Horwich, 1998).

of rules could be used to represent all the required distances. I will call these creatures "bees*". One of the features of Shea's bee* rules is that they are list-like: there is a different rule for each distance represented. Another feature is that the list is arbitrary: there is no general principle involved, such as more waggles representing more distance. The only principle is that each number of waggles represents a particular distance. Bee* dances are not "isomorphic" in any minimally strict sense of the word to the distance of nectar they represent. Yet, as Shea points out, his arbitrary rules seem to determine what qualifies as a "functional isomorphism" in Millikan's sense (2013, p. 75). As I indicated above, a systematic one-to-one pairing of representations with states of affairs seems indeed to be what she calls a "functional isomorphism". Shea's hypothetical bee* rules deliver such a pairing. The example echoes and supports Godfrey-Smith's objection that Millikan's notion of isomorphism is too liberal—that "nearly any relation can constitute a «mapping» in Millikan's theory" (1996, p. 185). Shea and Godfrey-Smith make these points in the context of discussing the relation between Millikan's isomorphism requirement and her teleological theory of content. As far as Shea is concerned, Millikan's teleosemantic theory has the resources to explain his hypothetical bee* rules in terms of evolutionarily grounded substantive relations.[2] He is not objecting to Millikan's teleosemantics. What Shea intends the example to show is that the isomorphism requirement does not play an independent substantive role in the theory.

Millikan draws the distinction between projected and substitutional correspondence in her response to Shea. The language of actual bees, she suggests, is a clear example of mapping by means of projected correspondence. There are mathematical functions or projection rules relating features of different bee dances to different locations of nectar. She offers the following illustration of what the rule for representing distance may look like: "For example: *add one waggle to the dance = add 1,000 yards to the distance of nectar*. If this equation applies to every possible bee dance, it describes a projected correspondence between dances and distances" (2013, p. 84). While also imaginary, this rule does capture the basic principle involved in actual bee dances: a longer dance represents a longer distance of nectar from the hive. About Shea's bee* rules, Millikan says: "Shea suggests for bee dances a list giving arbitrary correspondences between numbers of waggles and distances. This would be a list of correlated substitution transformations..." (2013, p. 84). Her point is that there is no mathematical equation or *projection* rule relating the number of waggles to the distance of nectar. Instead, she claims, there are *substitutional* rules that arbitrarily relate changes in the number of waggles to changes in the distance of nectar from the hive represented—e.g., replacing a two-waggle dance with a three-waggle dance transforms the distance represented from 300 m to 60 m (2013, pp. 84-85). Even though they are arbitrary, she acknowledges, bee* rules do determine a systematic one-to-one pairing of representations and what they represent (2013, p. 85). Setting aside the representation of direction, her claim is that a bee dance is true when it has a *projected correspondence* to the distance of nectar, while a bee* dance is true when it has a *substitutional correspondence* to the distance of nectar.[3]

Millikan further illustrates the distinction with another example where, she says, both kinds of correspondence are at play: a human-made map of a county where lines represent "the shape and placement of roads" and "different colors... represent different road surfaces, such as dirt, gravel, paved, and so forth" (2013, p. 86). This is basically a hybrid case where, if the map is accurate, the shape and placement of the lines have a projected correspondence to the actual shape and placement of the roads, while the colors of the lines have a substitutional correspondence to the actual surfaces of the roads.

Millikan's distinction can be reconstructed as a distinction between pictorial and non-pictorial forms of correspondence. Millikan herself, however, does not draw the contrast in these terms. On the contrary, she seems to regard projected and substitutional correspondences as two broad kinds of correspondence within the same framework of isomorphism (2013, pp. 83-84). Ryder *et al.* interpret her this way: "There is a *natural* isomorphism even when the correspondence is merely substitutional" (2013, p. 8). My suggestion is that this obscures an important distinction.

---

[2] The list-like nature of bee* rules may raise concerns about whether they determine substantive correspondences, since deflationists offer precisely list-like characterizations which, they claim, involve no substantive relation. Presumably the reply to this concern is that the bee* dances are substantially related to locations of nectar by evolutionary history.

[3] Millikan now claims that the teleological part of her theory is what grounds any such correspondences on real relations (2013, pp. 82-83).

What Millikan calls a "projected correspondence" involves something close to structural resemblance or isomorphism between representation and reality, while what she calls a "substitutional correspondence" involves merely a correlation determined by some substantial relation other than projection. While this is partly a terminological issue, I see no point in applying the term 'isomorphism' to non-projective correspondences. This would be misleading, since it would obscure the standard distinction made in the literature on the nature of truth between correspondence as congruence, structural resemblance or isomorphism, on one hand, and correspondence as mere correlation, on the other (Kirkham, 1992). Once a theory admits both non-projective and projective correspondences, it is better characterized as a correlation theory, since it claims that the common feature of all true representations is that they are paired to their truthmakers by one or another substantive relation—projective relations being merely one such kind of relation that some representational systems (e.g., bee dances) may rely on.

7. Sentential Truth without Picturing

Now we can address Millikan's position regarding sentences in human languages. Drawing on her distinction between projected and substitutional correspondence, she says that "almost all of the correspondences found in natural language are substitutional" (2013, p. 84) and claims that "this is the main difference between simpler representations and the representations of human language" (2013, p. 86). I am skeptical about the second claim, to which I return briefly in the next section. But let me start by commenting on the first one. Millikan has repeatedly argued that what sets apart sentences from other representations is their syntactic structure: they have subject-predicate form and can be negated (1984, pp. 308-309). She has pointed out, for example, that "bee dances are not sentences, for they have no subject terms, and a rabbit's danger thumps are not sentences, for they never predicate of a time and place anything other than danger... [while] the alarm calls of vervet monkeys are not sentences because... [they] are never negated" (1993, p. 118). She has also previously characterized the "mapping" rules of sentences in mainly substitutional terms, as shown by some passages I have already quoted. Here is an earlier example: "In the case of *sentences* that map, the significant transformations... are mostly substitution transformations..." (1984, p. 108). However, only in her response to Shea does she explicitly distinguish substitutional rules from projective rules. This distinction provides resources for a non-pictorial reconstruction of her account of sentential truth. If almost all natural-language correspondences are substitutional rather than projected, then the truth of sentences need not involve anything like the projective correspondence of accurate bee dances. On the reconstruction I have been defending, the correspondence involved in sentential truth is therefore best characterized as correlation, rather than as congruence, structural resemblance or isomorphism. While Millikan herself does not present the distinction as a rejection of her earlier Tractarian rhetoric or commitments, the distinction helps show how her account of sentential truth can be developed without requiring true sentences to structurally resemble states of affairs.

8. Representation without Picturing

I close this paper with two critical remarks. The first is substantive and the second is terminological. My first remark is against Millikan's claim that the distinction between substitutional and projected correspondence is what sets apart human sentences from *simpler* representations. (I think that her syntax-based distinction is more appropriate, although we may find syntax also in non-human languages of thought: whether syntax is uniquely human is an empirical question.) Notice that *some* very sophisticated representations (e.g., computer-generated models) may rely at least partly on projection rules. More importantly, many simple non-human representations seem to have "substitutional" correspondences. Vervet monkey alarm calls are an obvious example: they are merely correlated with the presence of the appropriate predators when accurate. Indeed, plenty of extremely simple "detection" systems represent reality without relying on any projection

rules or mathematical isomorphisms. The trigger hairs of the Venus flytrap are a clear example, if they are treated as representational at all. These mechanisms would *only* represent—when successful—that an insect has landed, so there are neither projected nor substitutional "transformations" involved. The distinction does not apply across the board. It is likely that some *perceptual* representations do rely on some forms of isomorphism (Millikan, 2005, p. 102). But correspondence as correlation seems to be ubiquitous and not merely a sophisticated human phenomenon.

My second point is that once correspondence by means of mathematical isomorphism is no longer treated as a necessary condition for representational truth, there is no motivation for regarding true representations in general as picturing what they represent, or for using talk of "mapping" in anything stronger than the thin sense of systematic pairing. Perhaps bee dances and many other representations literally picture what they represent. But this cannot be true of representation in general. Consider simple human sentences like 'Socrates is wise' or 'Snow is white'. Saying that these sentences *picture* something seems clearly metaphorical. What could that possibly mean other than that they *represent* something? Talk of "mapping" may still be retained in the thin sense of systematic pairing, but it should not be confused with picturing.

9. Conclusion

In this paper I have argued that Millikan's "pictorial" account, if taken at face value, seems to require genuinely pictorial relations between sentences and the states of affairs that make them true. This is why her Tractarian rhetoric is puzzling: structural resemblance is too demanding for sentential truth, and Millikan does not provide detailed mapping rules showing how true sentences could picture their truth-makers. The work that her mapping rules are supposed to do, however, can be done by something less than picturing. As the rule *SP* suggested above illustrates, a mapping rule may systematically pair sentences with the states of affairs that make them true without requiring the former to structurally resemble the latter. Millikan's own work provides resources for this non-pictorial reconstruction: her earlier suggestion of "some sort of Tarskian mapping" (2004b, pp. 231-232) and her later distinction between projected and substitutional correspondence (2013, pp. 83-86) both point toward a correlation-based account of sentential truth. Millikan does not explicitly present these resources as a rejection of her Tractarian commitments; indeed, in later work she continues to present representation in explicitly Tractarian terms: "Representations 'picture' what they represent, just as Wittgenstein said in *Tractatus logico philosophicus*" (2024, p. 58). My conclusion is therefore not that Millikan officially abandons picturing, but that the most plausible development of her account of sentential truth is a correlation account based on substantive relations between sentences and reality.

Acknowledgements

I am grateful to Michael Devitt, David Rosenthal, and Michael Levin. I also thank the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

Austin, J. L. (1950). Truth. *Proceedings of the Aristotelian Society, Supplementary Volumes, 24*, 111-128.

Field, H. (1972). Tarski's theory of truth. *The Journal of Philosophy, 69*(13), 347-375.

Field, H. (1994). Deflationist views of meaning and content. *Mind, 103*(411), 249-285.

Godfrey-Smith, P. (1996). *Complexity and the function of mind in nature*. Cambridge University Press.

Horwich, P. (1998). *Truth* (2nd ed.). Oxford University Press.

Kirkham, R. (1992). *Theories of truth: A critical introduction*. The MIT Press.

Millikan, R. G. (1984). *Language, thought, and other biological categories*. The MIT Press.

Millikan, R. G. (1993). *White queen psychology and other essays for Alice*. The MIT Press.

Millikan, R. G. (2004a). *Varieties of meaning: The 2002 Jean Nicod lectures*. The MIT Press.

Millikan, R. G. (2004b). Existence proof for a viable externalism. In R. Schantz (Ed.), *The externalist challenge* (pp. 227-238). De Gruyter.

Millikan, R. G. (2005). *Language: A biological model*. Oxford University Press.

Millikan, R. G. (2013). Reply to Shea. In D. Ryder, J. Kingsbury, & K. Williford (Eds.), *Millikan and her critics* (pp. 81-86). Wiley-Blackwell.

Millikan, R. G. (2024). Teleosemantics and the frogs. *Mind & Language, 39*(1), 52-60.

Moore, G. E. (1953). *Some main problems of philosophy*. George Allen & Unwin.

Patterson, D. (2003). What is a correspondence theory of truth? *Synthese*, 137, 421-444.

Russell, B. (1912). *The problems of philosophy*. Oxford University Press.

Russell, B. (1919). On propositions: What they are and how they mean. *Proceedings of the Aristotelian Society, Supplementary Volumes, 2*, 1-43.

Russell, B. (2010). The philosophy of logical atomism. In B. Russell, *The philosophy of logical atomism* (pp. 1-125). Routledge. (Original work published 1918).

Ryder, D., Kingsbury, J., & Williford, K. (2013). Introduction. In D. Ryder, J. Kingsbury, & K. Williford (Eds.), *Millikan and her critics* (pp. 1-20). Wiley-Blackwell.

Sellars, W. (1963). *Science, perception and reality*. Humanities Press.

Shea, N. (2013). Millikan's isomorphism requirement. In D. Ryder, J. Kingsbury, & K. Williford (Eds.), *Millikan and her critics* (pp. 63-80). Wiley-Blackwell.

Tarski, A. (1944). The semantic conception of truth. *Philosophy and Phenomenological Research*, 4(3), 341-376.

Tarski, A. (1956). The concept of truth in formalized languages. In *Logic, semantics, metamathematics* (pp. 152-278). Oxford University Press.

von Frisch, K. (1953). *The dancing bees: An account of the life and senses of the honey bee* (D. Ilse, Trans.). Harvest Books.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus* (C. K. Ogden, Trans.). Kegan Paul.

Esteban Withrington is an Associate Researcher at the Institute for Philosophical Research, National Autonomous University of Mexico (UNAM). His work focuses on the philosophy of mind, the philosophy of language, and the philosophy of cognitive science, especially teleosemantics and other naturalistic theories of mental and linguistic content.

Address: Institute for Philosophical Research, National Autonomous University of Mexico, Circuito Mario de la Cueva s/n, Ciudad Universitaria, Mexico City, 04510, Mexico. Email: ewithrington@filosoficas.unam.mx – ORCID: 0000-0002-5350-1402

Theoria, 2026, 41(3), 378-395
https://doi.org/10.1387/theoria.27706

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Metacognitive feelings as mental affordances

(Sentimientos metacognitivos como affordances mentales)

Santiago Arango-Muñoz*
Universidad de Antioquia
https://orcid.org/0000-0003-3761-3487

Sebastián Mejía-Rendón**
Universidad Nacional de Córdoba
https://orcid.org/0000-0001-7356-1181

Keywords

Metacognitive feelings
Mental affordances
Mental action
Remembering
Reasoning
Attention

ABSTRACT: This paper explores the role of metacognitive feelings as mental affordances. Drawing on McClelland's Mental Affordance Hypothesis (MAH), we argue that metacognitive feelings are mental affordances because they motivate agents from within and guide their mental behavior. More specifically, we claim that metacognitive feelings satisfy McClelland's sensitivity and potentiation requirements: they make agents sensitive to their own cognitive states and potentiate specific mental actions. After presenting and assessing the MAH, we examine the motivational features of metacognitive feelings, focusing especially on valence and intensity. We then analyze metacognitive feelings related to remembering, reasoning, and attention. Finally, we argue that the MAH should be expanded to incorporate these feelings, since they function as internal affective signals that potentiate and modulate mental actions in ways consistent with the hypothesis.

Palabras clave

Sentimientos metacognitivos
Affordances mentales
Acción mental
Recordar
Razonar
Atención

RESUMEN: Este artículo explora el papel de los sentimientos metacognitivos como *affordances* mentales. Basándonos en la Hipótesis de las *Affordances* Mentales (Mental Affordance Hypothesis, MAH) de McClelland, sostenemos que los sentimientos metacognitivos constituyen *affordances* mentales porque motivan a los agentes desde el interior y orientan su conducta mental. Más específicamente, argumentamos que los sentimientos metacognitivos satisfacen los requisitos de sensibilidad y potenciación propuestos por McClelland: hacen que los agentes sean sensibles a sus propios estados cognitivos y potencian acciones mentales específicas. Tras presentar y evaluar la MAH, exponemos las características motivacionales de los sentimientos metacognitivos, centrándonos especialmente en su valencia e intensidad. A continuación, analizamos los sentimientos metacognitivos relacionados con la memoria, el razonamiento y la atención. Finalmente, sostenemos que la MAH debería ampliarse para incorporar estos sentimientos, ya que funcionan como señales afectivas internas que potencian y modulan las acciones mentales de un modo coherente con dicha hipótesis.

Gako-hitzak

Sentimendu metakognitiboak
Affordance mentalak
Ekintza mentala
Gogoratzea
Arrazoitzea
Arreta

LABURPENA: Artikulu honetan sentimendu metakognitiboek *affordance* mental gisa duten eginkizuna aztertzen da. McClelland-en *affordance* mentalen hipotesia (Mental Affordance Hypothesis, MAH) oinarritzat hartuta, sentimendu metakognitiboak *affordance* mentalak direla argudiatzen dugu, agenteak barrutik motibatu eta haien portaera mentala gidatzen dutelako. Zehatzago esanda, baieztatzen dugu sentimendu metakognitiboek McClellandek proposatutako sentikortasun- eta potentziazio-baldintzak betetzen dituztela: agenteak beren egoera kognitiboekiko sentikor bihurtzen dituzte eta ekintza mental espezifikoak potentziatzen dituzte. MAHa aurkeztu eta ebaluatu ondoren, sentimendu metakognitiboen motibaziozko ezaugarriak aztertzen ditugu, bereziki balentzian eta intentsitatean arreta jarriz. Ondoren, gogoratzearekin, arrazoitzearekin eta arretarekin lotutako sentimendu metakognitiboak aztertzen ditugu. Azkenik, argudiatzen dugu MAHa zabaldu behar litzatekeela sentimendu horiek barne hartzeko; izan ere, ekintza mentalak hipotesiarekin bat datozen moduetan potentziatzen eta modulatzen dituzten barne-seinale afektibo gisa funtzionatzen dute sentimenduok.

* ***Correspondence to:*** Santiago Arango-Muñoz. Instituto de Filosofía, Universidad de Antioquia, Ciudad Universitaria, Block 12, Office 408, Calle 70 No. 52-21, Medellín, Colombia. – santiago.arango5@udea.edu.co – https://orcid.org/0000-0003-3761-3487

** ***Correspondence to:*** Sebastián Mejía-Rendón. Institute of Humanities (IDH), CONICET – Universidad Nacional de Córdoba, Faculty of Philosophy and Humanities, Córdoba, Argentina. – joan.mejia@mi.unc.edu.ar – https://orcid.org/0000-0001-7356-1181

How to cite: Arango-Muñoz, Santiago; Mejía-Rendón, Sebastián (2026). «Metacognitive feelings as mental affordances»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 41(3), 378-395. (https://doi.org/10.1387/theoria.27706).

Received: 14 July, 2025; Final version: 13 July, 2026.
ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press

This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Introduction

McClelland recently put forward the Mental Affordance Hypothesis (henceforth MAH), which posits the existence of mental affordances. According to McClelland, mental affordances are perceptions of external stimuli that invite certain mental actions such as attending, imagining, or calculating (McClelland, 2020). An amplified voice invites us to pay attention, a fantasy novel presents the opportunity to imagine fictional characters, and a jar full of marbles affords the act of counting them.

While we endorse McClelland's MAH, we argue that it should be expanded to include other types of stimuli that afford mental action. Based on Proust (2013, 2015a, 2015b, 2024), we claim that metacognitive feelings and epistemic emotions are mental affordances because they motivate us to engage in mental action. Metacognitive feelings are first-order affective experiences related to cognitive processes, dispositions, and mental states. Paradigmatic examples of metacognitive feelings are the tip-of-the-tongue phenomenon, the feeling of knowing, the feeling of error, the feeling of forgetting, among others. For instance, the feeling of knowing and the tip-of-the-tongue phenomenon afford the mental action of memory retrieval. The feeling of error affords the revision of the outcome of the mental process, while the feeling of forgetting affords the deployment of alternative mnemonic strategies when memory retrieval is not possible.

Recently, Proust (2024) analyzed how metacognitive feelings can be understood as mental affordances—a significant contribution to the literature on mental agency. While our perspective aligns with Proust's earlier work (2013, 2015a, 2015b, 2024), we diverge at a key point. Proust proposes an "informational model" in which metacognitive feelings serve to detect cognitive affordances (opportunities for mental action), treating the two as distinct entities. In contrast, we argue that metacognitive feelings are not merely detectors of affordances—they *are* mental affordances themselves. This proposal avoids ontological duplication: there is no need to posit both a detector and a detected affordance. Metacognitive feelings are already experienced as opportunities for action; hence, they fulfill the role of affordances directly.[1] To support this claim, we propose an expanded definition of the Mental Affordance Hypothesis (MAH) that explicitly includes such metacognitive states. We offer two key reasons for this extension.

First, adopting an account of affordances that incorporates feelings and phenomenal states provides a richer understanding of how mental affordances are experienced. By acknowledging this experiential dimension, we can better explain why certain mental affordances are more compelling than others and how metacognitive feelings, such as curiosity, the feeling of error, or cognitive effort, shape the way we engage in mental action. Second, incorporating internal mental states into the MAH allows us to account for additional dimensions of mental action that have been largely overlooked in the literature: namely, the internal modulation of mental action, including shifts in strategy, revision of outcomes, and even the decision to disengage from a mental process altogether.

The paper proceeds as follows. In section 1, we present the MAH. In section 2, we assess the MAH. In section 3, we outline the motivational dimension of metacognitive feelings. In section 4, we analyze metacognitive feelings related to remembering, reasoning, and attention. Finally, we defend the proposed extension of the MAH.

1. The Mental Affordance Hypothesis

The term "affordance" was coined by Gibson in the 20th century, and it has since led to an enormous literature that employs it in various ways. The fields of neuroscience, the arts, ecological psychology, and tool-use studies have all made

[1] Before proceeding, it is important to clarify that identifying metacognitive feelings with mental affordances does not require attributing rich representational content to those feelings. Metacognitive feelings are non-conceptual, first-order affective experiences (see §3.1). The cognitive sophistication lies in the underlying cognitive system (memory, reasoning, attention mechanisms); the feeling is the subjective manifestation of the relational property between that system and the task. Thus, we avoid both over-intellectualizing the feeling and thinning the affordance.

use of this concept.[2] Gibson introduced this concept to emphasize that the environment provides opportunities for action relative to an animal's bodily and sensory capacities (Gibson, 1979). For example, a stick is useful for a chimpanzee in hunting termites, but not for a dolphin. In this case, the affordance is animal-relative, meaning that affordances are relative to an individual's potential for action:[3]

> The *affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*, either for good or ill. The verb to *afford* is found in the dictionary, but the noun affordance is not. I have made it up. I mean by it something that refers to both the environment and the animal in a way that no existing term does (Gibson, 1979, p. 127; emphasis in the original).

For decades, the term "affordance" has been used across a wide range of disciplines and has consequently become conceptually diffuse. Given the variety of applications of the term, it is often difficult to determine its precise reference when used in different contexts (Osiurak & Badets, 2017). Nevertheless, despite the diversity of uses of this term, an affordance can simply be defined as an opportunity for action in a situation in which it is possible for a subject to deploy some ability she possesses. In a nutshell, affordances can be understood as possibilities for action that arise from the relation between an agent and her environment. According to McClelland, an affordance can be defined as follows:

AFFORDANCE: x affords ϕ-ing for S if and only if x offers an opportunity for S to ϕ (McClelland, 2020, p. 405).

The concept of affordance has been introduced in the literature to explain agency. Gibson introduced the concept to explain how perceived environmental features guide bodily action (Zipoli Caiani, 2014). In his original proposal, Gibson mainly focused on bodily affordances and how the perception of an object offers, solicits or motivates[4] some bodily actions of an agent, such as grasping a glass or climbing a tree. We perceive affordances directly without the mediation of any other mental representation; that is, we do not infer them from the perceptual content, but we perceive opportunities for action directly in the environment. Building on Gibson's proposal—while adapting it to his representational framework—, McClelland argues that we are perceptually sensitive to both bodily and mental affordances. He distinguishes between these two types of affordances as follows. On the one hand, a bodily affordance is characterized by the perception of a stimulus that presents the possibility of performing a bodily action. For example, seeing a chair affords the action of sitting; seeing a glass of water affords the actions of grasping it and drinking the water. On the other hand, mental affordances are perceptions of external stimuli that present an opportunity to perform a mental action. A mental action is a mental event that is under our control, so that one can initiate, modulate or stop it (Peacocke, 2021; Proust, 2001), and does not necessarily involve any bodily movement for achieving its goal.[5] McClelland illustrates this idea with the case of a jar of marbles that affords the mental action of counting (McClelland, 2020, p. 422). In this case,

---

2 In theory of action, affordances have been understood as (1) action opportunities, (2) links between visual attributes and motor knowledge, (3) online motor control, (4) sensorimotor knowledge, or (5) semantic knowledge. In tool-use reflection, affordances have been understood as (1) familiar tool use, (2) unfamiliar tool use, (3) novel tool use, or (4) non-tool use actions. In neuroscience, affordances can be associated with (1) a hand-centered, (2) a tool-centered frame of reference, or (3) both. Or, affordances can be regarded as relying on (1) the dorso-dorsal system, (2) the dorsal system, (3) the ventrodorsal system or (4) the ventral system (Osiurak & Badets, 2017).

3 Here we endorse Rietveld and Kiverstein's (2014) view on affordances. According to these authors, and against the traditional Gibsonian view of affordances, to perceive an affordance is animal-relative (not species-relative) and depends on the skills and abilities of the agent. Thus, an agent can only perceive an external stimulus as affording an action if the agent has the ability to perform such an action.

4 It is important to note that while a given affordance invites a particular action, it does not necessarily motivate it. For example, one might perceive a door as openable but feel no inclination to do so after reading a sign that says, "Keep the door closed." However, when an affordance is perceived and the corresponding action follows, we can consider that the affordance played a role—among other factors—in motivating the action. In this paper, we focus on cases where affordances contribute to motivation, as these are crucial for explaining agency. However, this does not mean that individuals are always compelled to act on affordances; resisting their pull is also possible. The phenomenon of resisting affordances will be the subject of another project.

5 This does not mean that mental actions cannot be accompanied by or realized through bodily movements and external interventions, as Clark and Chalmers (1998) have shown.

we perceive an object that potentiates the arithmetical action of counting. McClelland, in turn, introduced the concept of *mental affordance* to enrich Gibson's concept so that it can also explain our capacity to mentally act, that is, our mental agency.

Bodily and mental affordances share two key requirements: sensitivity and potentiation. The first requirement means that the agent is sensitive to affordances. For example, the agent sees a cake *as edible*, a bike *as rideable*, etc. The second feature, potentiation, refers to the automatic preparation of motor patterns responsible for performing the afforded action. The same can be said for mental affordances. In McClelland's example concerning stepping stones across a river, we perceive the situation as sometimes affording a bodily action, such as the next step in the easy cases, or affording the mental action of imagining. According to McClelland, in the difficult cases, that is, when it is not evident where one should step, the act of imagining a viable route for the next step is afforded. Thus, the stepping stone case shows how the bodily affordances (those that motivate a bodily action, e.g., executing the next step) and the mental affordances (those that motivate a mental action, e.g., imagining a possible way through the stones) are intertwined and are both triggered by the perception of environmental stimuli.

However, our point is that human mental agency is more complex than the Gibsonian view suggests, and it is not always bound to the perception of external stimuli. Unlike other animals, whose actions are tightly coupled with their immediate environment, humans are representational agents that can entertain thoughts about distal objects and properties (e.g., the center of the Earth, the Andromeda galaxy, etc.), non-existent beings (e.g., Sherlock Holmes, the fountain of youth, unicorns, etc.), and even about their own mental states and affective experiences (e.g., my trust in science, my feeling of anxiety, and my desire to formulate a strong argument). Furthermore, mental actions cannot be reduced merely to mental affordances associated with external stimuli. In contrast, Metzinger (2017) and Jorba (2020) have proposed accounts of mental agency that recognize the existence of mental affordances while allowing them to be triggered by elements internal to the mind. Thus, we argue that it is important to broaden the concept to encompass other types of affording elements, namely metacognitive feelings and emotions.

2. *Assessing the MAH*

The MAH posits the existence of mental affordances, understood as perceptions of external stimuli that invite certain mental actions such as attending, imagining, or calculating (McClelland, 2020). In this view, mental affordances are not merely triggered by external stimuli; rather, the perception of these stimuli potentiates particular mental actions. While McClelland's framework is informed by ecological psychology, which emphasizes a direct relationship between organisms and their environment, he adopts a representational view of affordances. That is, perceiving an affordance involves forming a perceptual representation of its object. This contrasts with the non-representational view in ecological psychology, which holds that affordance perception does not require a mental representation as intermediary. Our main critique of McClelland's approach is that he does not sufficiently distance himself from the core principles of ecological psychology. Even though McClelland claims to endorse a representational view of affordances in line with Siegel (2014), his emphasis on the perception of external stimuli ties affordances too narrowly to external stimuli and the environment, overlooking other potential sources of affordance—particularly internal mental states. In contrast to this way of thinking about agency, Metzinger (2017) and Jorba (2020) have proposed two related accounts of mental agency that acknowledge the existence of mental affordances, but also allow them to be triggered by internal elements of our mental environment. We will describe both views in the next paragraphs.

On the one hand, Metzinger puts forward an account of mental action in relation to predictive processing. According to his view, the brain generates a range of potential mental actions beyond what the external environment offers to us as possibilities of bodily action. These possible actions compete with each other until one of them becomes available in the stream of consciousness. Metzinger describes this as "a constant flow of possible mental ac-

tions, a dynamic inner environment constituted by non-sensory events which need to be predicted and controlled" (Metzinger, 2017, p. 11). This constant flow creates a rich inner repertoire of possible mental actions that are referred to as "affordances for cognitive agency". Therefore, according to Metzinger, our inner world is not made up of static mental objects, but rather a dynamic competition of affordances for potential cognitive actions (Metzinger, 2017, p. 12).

On the other hand, Jorba (2020) returns to Husserl's notion of "horizon" understood as an implicit feature of experience, and suggests that experience goes beyond the perceptual domain, namely, to thinking. Based on Husserl's work, she distinguishes between the perceptual horizon and the cognitive horizon. While the perceptual horizon consists of a set of possibilities opened by the experience of external objects, the cognitive horizon can be understood as a rational dimension of the cognitive experiences that can be motivating reasons for action (Jorba, 2020, p. 859). According to her, the cognitive horizon is an inner experience that involves a free play dimension of inferential relations and associations. Thus, the cognitive horizon shapes the experience of thinking which consists in "seeing" what you can do with your thoughts; that is, some thoughts in your stream of conscious thought might lead to physical or mental actions. This sense of the possibility of a mental action is what Jorba calls "cognitive affordance" (Jorba, 2020, p. 856).[6]

Despite the evident differences between the two accounts, Metzinger's "inner environment" and Jorba's "cognitive horizon" both underscore a key point we have emphasized: many of our mental actions are initiated, guided, and motivated by interactions among internal mental states. For Metzinger, this involves a dynamic competition of potential cognitive actions, of which the subject is aware as "cognitive affordances". In Jorba's view, our cognitive experience presents us with associative and inferential relations that motivate different courses of action or thought. However, in both cases it remains unclear how the subject becomes aware of these affordances or the relational dynamics that drive action and thinking.

Although we endorse McClelland's Mental Affordance Hypothesis to a certain extent, following Metzinger (2017) and Jorba's (2020) suggestions, we argue that the MAH should be expanded to encompass metacognitive feelings and emotions. Two considerations motivate this extension. First, mental affordances are experienced, not just perceived: incorporating feelings and phenomenal states into the concept of affordance provides a richer understanding of how mental affordances are felt. Mental activity is often accompanied by affective tones—such as ease, difficulty, urgency, or reluctance— that shape our mental actions. For example, the urges to focus on a thought, recall a memory, or entertain an imaginative scenario are influenced by how those thoughts feel to us. A more comprehensive theory of mental affordances should include these experiential dimensions. Second, mental affordances are not exclusively triggered by external stimuli. More specifically, not all invitations to think, imagine, or attend originate from the external environment; some are generated by internal mental states. In this sense, some philosophers of mind have introduced the 'mental environment' metaphor to characterize thought as a dynamic, self-sustaining space where a continuous flow of mental content shapes potential mental actions. Within this rich cognitive landscape, affordances emerge from our own thoughts, doubts, memories, or emotions. For example, an idea we have may afford reflection (Jorba, 2020; Metzinger, 2017), a doubt may afford the suspension of judgment, an imagined scenario may motivate daydreaming, and a sudden memory may afford reminiscing –much like the famous madeleine episode in *In Search of Lost Time* by Marcel Proust. Thus, a complete account of mental agency must recognize that mental affordances have an affective dimension that guides mental action, and they can arise from both external stimuli and internal mental states.

In the following section, we propose that metacognitive feelings play a crucial role in this awareness and guide mental actions. These experiences provide guidance for selecting and regulating mental actions, bridging the gap between internal mental affordances and conscious mental agency.

---

[6] For present purposes, we treat Jorba's notion of cognitive affordance as broadly equivalent to what we call a mental affordance.

3. Motivational properties and mental agency: Metacognitive feelings as mental affordances

As argued in the previous section, McClelland's framework provides a valuable starting point but remains limited in two respects. First, it ties mental affordances primarily to external stimuli (e.g., perceiving a jar of marbles invites counting). Second, even if he acknowledges that there can be internal triggers (e.g., thoughts, emotions, memories), he does not explain how we can be sensitive to internal mental affordances. To elaborate on this final point, we propose that a first step in this direction is to identify specific instances of mental states that function as mental affordances, but are not necessarily tied to the perception of external stimuli. Among such mental states, we suggest that metacognitive feelings are particularly good candidates for fulfilling this role. Paradigmatic examples of metacognitive feelings are the tip-of-the-tongue phenomenon, the feeling of knowing, the feeling of error, the feeling of forgetting, among others.

3.1. Metacognitive feelings

We acknowledge that broadening the notion of affordance to encompass all forms of mental experience may be problematic and risks diluting the concept. Nonetheless, a natural first step is to explore the motivational role of metacognitive feelings. Although various terms have been proposed to describe these experiences —such as cognitive feelings, epistemic feelings (Arango-Muñoz, 2014; de Sousa, 2009), noetic feelings (Dokic, 2012), and metacognitive feelings (Norman *et al.,* 2010; Arango-Muñoz, 2019)— they share common features. Despite terminological differences, these notions largely belong to the same family of experiences. Metacognitive feelings are first-order, conscious, affective (Loev, 2022), non-conceptual experiences (Proust, 2015a) that emerge during cognitive activity (Koriat, 2000), functioning as embodied evaluations of one's own mental processes (Dorsch, 2025), including memory retrieval, reasoning, and decision-making (Arango-Muñoz, 2014). We examine each of these features below.

As Proust (2015a, 2015b) theorizes, these feelings function as affordance sensings—implicit detections of mental-action opportunities that bypass conceptual inference: "Affordances are neither perceived nor inferred; they are 'sensed'" (Proust, 2015b, p. 1). Their conscious character is evident in phenomena like the tip-of-the-tongue (TOT) state, where the subject is acutely aware of the temporary inability to recall a word (Koriat, 2000). The non-conceptual character of metacognitive feelings means that they do not represent propositions, procedures, or complex structures. They signal "recallability" or "error" in a direct, affectively laden manner, much like a bodily sensation signals "loss of balance" without describing the physics of falling. This non-conceptual character is precisely what allows them to function as affordances: they present opportunities for action without specifying the means, leaving the cognitive system to supply the detailed procedure.

Metacognitive feelings are constituted by transient, context-specific positive or negative valence, that is, the felt positive or negative affects attached to certain experiences (Teroni, 2018). These affective experiences are commonly understood either hedonically, as sensations of pleasantness or unpleasantness, or evaluatively, as conveying a sense of worth or lack thereof. Importantly, valence is not merely a descriptive feature but is closely tied to the motivational role of emotional experiences. That is, emotional and metacognitive experiences exert a motivational force that can prompt specific behaviors or guide cognitive actions (Carruthers, 2018; Corns, 2014). This motivational power is not binary but gradable —some feelings may drive immediate action or revision, while others exert weaker influence, allowing for sustained deliberation or disengagement (Kozuch, 2020). In this way, the valenced character of metacognitive feelings not only registers internal evaluations but also helps shape the direction and intensity of subsequent thought and behavior (Loev, 2022).

As embodied evaluations (Dorsch, 2025), metacognitive feelings bridge cognition and emotion by translating mental uncertainty into action-guiding signals (Fernández Velasco & Loev, 2025). According to Dorsch (2025), these feelings arise from interoceptive awareness of bodily changes, particularly cardiovascular signals, which encode and convey metacognitive information. Experimental evidence suggests that bodily afferents not only correlate with but causally influence the intensity and valence of noetic feelings. This supports a "strong embodied" view: bodily states do not merely

accompany cognition but actively shape it by translating cognitive processing fluency into affective signals that guide thought. Thus, these findings suggest that metacognitive processes depend in part on extracerebral bodily processes.

Finally, metacognitive feelings are non-conceptual (Proust, 2015a): they register bodily and contextual cues through affect without representing explicit causal relations. As Proust notes, "Epistemic feelings express affordances in a nonconceptual way" (Proust, 2013, p. 269). Their non-conceptual character distinguishes them from explicit metacognitive judgments: they encode bodily dynamics and predictive expectations (e.g., a feeling of knowing as a "sensed" opportunity to recall) rather than conceptual content (Proust, 2015b, p. 2). In a recent paper, Proust (2024) develops an informational model according to which metacognitive experiences detect cognitive affordances, inform agents about the viability of cognitive tasks, and guide action through three control stages: detection, adjustment, and evaluation. She argues that cognitive affordances are informational opportunities sensed through specialized affective appraisals. Accordingly, affordance sensitivity occurs at control junctures involving the detection of an opportunity, the adjustment of ongoing behavior, and the assessment of the final outcome (Goupil & Proust, 2023). On this account, metacognitive feelings and cognitive affordances are distinct entities: the former function as detectors of the latter. By contrast, we argue that metacognitive feelings are themselves cognitive affordances. Rather than detecting independently existing affordances, they constitute the very opportunities for mental action that they make available to the agent. This proposal avoids an unnecessary multiplication of entities while preserving the explanatory role that Proust assigns to metacognitive feelings.

A potential concern is that identifying metacognitive feelings with affordances either over-intellectualizes the feeling or thins the affordance. We do not claim that a feeling, such as the feeling of familiarity, must represent the entire structural complexity of the memory system. Rather, the feeling is the relational manifestation through which the latent capacities of the cognitive system become available to the agent. In this view, the feeling is the non-conceptual signal that translates sub-personal complexity into an action-guiding opportunity. By identifying the two, we avoid the ontological duplication of a 'detector' and a 'detected' entity, acknowledging that the affordance's functional role is only realized through the feeling's motivational presence. Thus, our approach differs slightly from Proust's, insofar as we avoid positing a duplication of entities. As a result, our account of metacognitive feelings is more parsimonious.

3.2. Metacognitive feelings as affordances

In this section, we propose that metacognitive feelings should be understood as affordances because they internally motivate the agent to act in a certain way and thus guide her mental behavior. In this sense, they function as internal motivation, not by representing possibilities from a detached standpoint, but by guiding cognitive activity. When we say that a metacognitive feeling is a mental affordance, we mean that the feeling itself constitutes the affordance for mental action – not a signal that detects an independent affordance. As such, it (a) presents a possible mental action (e.g., "try to recall"), (b) motivates initiating or sustaining it, and (c) dynamically structures its execution by enabling online adjustments (shifting strategy, persisting, disengaging) as the cognitive process unfolds (see the three stages in this section). These three functions are not additional features of the feeling; they constitute its affordance character. Taken together, these considerations support the inclusion of metacognitive feelings among mental affordance experiences, given that they meet both the sensitivity to internal cognitive states and the potentiation of action requirements outlined by McClelland. To establish this claim, we examine each requirement in turn. We first argue that metacognitive feelings satisfy the sensitivity requirement because they make agents aware of features of their own cognitive activity. We then show that they satisfy the potentiation requirement because they motivate and regulate mental action.

Concerning the sensitivity requirement, metacognitive feelings are felt experiences and thus have qualitative properties. Although they are not directly caused by perception of external stimuli, as in McClelland's examples, there is something it is like to undergo a feeling of knowing different from what it is like to have a feeling of error or a feeling of forgetting. That is why Proust claims: "an affordance is neither directly perceived nor inferred. It is directly sensed, which requires a different kind of ability" (Proust, 2015b, p. 3). This phenomenological accessibility is precisely what allows metacog-

nitive feelings to satisfy the sensitivity requirement. However, merely making an agent aware of a cognitive situation is not sufficient for something to count as a mental affordance. Mental affordances must also motivate or facilitate action. We must therefore examine whether metacognitive feelings satisfy McClelland's potentiation requirement.

Concerning the potentiation requirement, it seems that metacognitive feelings also potentiate certain mental actions. Metacognitive feelings, as argued by de Sousa (2009), not only shape our beliefs and decisions, but also play a crucial role in guiding our cognition. This role is displayed thanks to the involvement of valence. Metacognitive feelings can have either positive or negative affect (Arango-Muñoz, 2014). The feeling of knowing seems to be in part constituted by a positive valence concerning the possibility of retrieving a given piece of information, whereas the feeling of forgetting seems to be constituted by a negative valence. The same applies to curiosity and boredom. Moreover, it has also been highlighted that metacognitive feelings have intensity: "[they] can be felt as weak or strong, according to the activity in their dedicated somatic markers" (Proust, 2013, p. 136). Accordingly, some researchers have suggested that metacognitive feelings are embodied experiences that the agent can use to monitor and control her own mental dispositions (Arango-Muñoz, 2019; Dokic, 2012; Dorsch, 2025). In this respect, metacognitive feelings resemble paradigmatic affordances: they do not merely reveal possibilities for action but actively invite their realization.

Moreover, the potentiation of certain cognitive processes is explained by the functional characterization of the origin of feelings and the different stages in cognition where they take place. Metacognitive monitoring involves three stages where metacognition assesses the mental processes and triggers metacognitive feelings accordingly: *self-prediction, intermediate assessment, and post-evaluation* (Arango-Muñoz, 2014; Goupil & Proust, 2023; Proust, 2024). Each stage is associated with characteristic metacognitive feelings that guide mental action in different ways. Goal-related *self-prediction* assesses whether a potential or future cognitive process can attain a particular goal: for example, whether by trying to remember one can retrieve a particular piece of information, or whether by mentally calculating one can solve a math problem. In the *intermediate assessment*, metacognition evaluates the unfolding of the mental process and whether it is leading to the attainment of the desired goal, so that the subject can adjust her cognitive strategy online. For example, during a philosophy exam about Hegel's *Phenomenology*, the subject may try to report the main ideas of the book. In trying to remember Hegel's ideas, the subject may notice that she is only barely able to retrieve some words and concepts, but not to reconstruct his main ideas or arguments. Finally, in the *post-evaluation*, metacognition evaluates the outcome of the cognitive process and determines whether the cognitive process was successful. For example, after calculating a sum, the subject may feel that she has committed an error in the procedure. All these assessments generate positive or negative metacognitive feelings that inform the subject about the dynamics of her mental processing and thus motivate her to carry out certain actions such as selecting a mental strategy, changing the strategy, and/or accepting the result. The feeling of knowing potentiates remembering, the feeling of forgetting potentiates using mnemonic tools to remember, and the feeling of rightness potentiates accepting a conclusion of a reasoning process.

While sensitivity ensures that the subject has access to her own cognitive states through felt experiences, potentiation refers to their action-guiding character (shaped by valenced experiences), which we identify as essentially motivational features. It is important to distinguish between a property being merely 'accessible' and it becoming 'operative.' While an informational model might suggest that an affordance exists independently and is simply sensed, we argue that for an affordance to be truly mental, it must be operative—meaning it must function as the mechanism of online control. This is precisely where our view departs from informational models that treat metacognitive feelings merely as detectors of independently existing affordances. By "operative" we mean three interrelated functions that go beyond mere access: (a) the feeling presents the agent with a possible mental action Y (e.g., "try to recall"), (b) it motivates the initiation or continuation of Y, and (c) it dynamically structures the execution of Y by enabling online adjustments—such as persisting, shifting strategy, or disengaging—depending on the unfolding of the cognitive process. These adjustments correspond to the three stages of metacognitive monitoring—self-prediction, intermediate assessment, and post-evaluation—through which metacognitive feelings regulate ongoing cognition. A feeling of knowing does more than provide access to the fact that a memory exists; it provides the bridge that motivates the agent to initiate the reconstruction

process and guides the deployment of mnemonic resources. What makes a feeling "operative," then, is not merely that it makes an affordance available, but that it participates in its temporal unfolding as a guide for action.

Taken together, these features suggest that metacognitive feelings are not just feedback signals, but constitutive components of mental agency. They motivate the initiation, suspension, or revision of cognitive processes in a sensitive manner. By identifying metacognitive feelings (MFs) with the affordances that emerge in these stages, we avoid the ontological duplication of a separate 'detector' and 'detected' entity. The feeling does not need to 'represent' the entire structural complexity of the memory system to be an affordance; rather, it is the relational signal through which the system's latent capacities become available to the agentive self. Therefore, their role cannot be reduced to the detection of affordances—as in Proust's model—but should be understood as affordances themselves: they motivate mental action.

To conclude this section, we must address the concern that identifying metacognitive feelings with affordances overlooks the procedural and structural differences compared to physical affordances. We contend that these disanalogies—procedural, structural, and feedback-related—stem from an over-reliance on visual models of action. Metacognitive feelings present procedures in a non-conceptual, affective format. Furthermore, our three-stage model—self-prediction, intermediate assessment, and post-evaluation—demonstrates that these feelings are not static outputs but dynamic, online signals that provide the feedback and structure necessary for mental control. By identifying the feeling as the affordance itself, we avoid the ontological duplication inherent in the Proustian model, acknowledging that a mental affordance only becomes operative for the agent when it is felt as a motivating opportunity for action.

One might object that metacognitive feelings cannot themselves structure cognitive activity because the detailed procedures involved in remembering, reasoning, or attending are implemented by subpersonal mechanisms.[7] We agree that metacognitive feelings do not encode or execute these procedures. However, this objection assumes that affordances must contain procedural information. Bodily affordances do not always specify detailed motor instructions; rather, they make certain actions available and guide their ongoing regulation. Similarly, metacognitive feelings need not determine the specific cognitive operations involved in retrieval, inference, or attention. Their role is to present these activities as available courses of action and to regulate their continuation, modification, or termination through ongoing affective feedback. In this sense, they function as affordances without duplicating the underlying cognitive mechanisms.

In the following section, we examine how these general claims apply to particular cognitive domains. We focus on remembering, reasoning, and attention, arguing that metacognitive feelings guide and regulate mental action in each of these cases.

4. The analysis of three cases

4.1. Metacognitive feelings related to remembering

Remembering can sometimes be considered as a mental action in the sense that the agent controls her mnemic construction (Arango-Muñoz & Bermúdez, 2018), and the process can be improved with practice and the agent's performance can be assessed for excellence (Goldwasser, 2023). Remembering is especially relevant for our purposes because it provides one of the clearest cases in which metacognitive feelings both disclose cognitive possibilities and guide subsequent mental action. Remembering is not just the passive process of encountering some mental contents that one has stored in memory, but the active reconstruction of a past episode (in the case of episodic memory) or some data one has learned (in the case of semantic memory) (De Brigard, 2014; Michaelian, 2016).

This active reconstruction of past episodes is monitored and controlled by metacognition throughout the three stages that we have presented in the previous subsection: *self-prediction, intermediate assessment, and post-evaluation* (Arango-Muñoz

[7] We are grateful to an anonymous reviewer for raising this concern, which helped us refine our account of how metacognitive feelings function as mental affordances.

& Bermúdez, 2018). There are some metacognitive feelings related to these stages in remembering. In particular, the feeling of knowing precedes some instances of remembering (*self-prediction*). Psychologists have found that often, when someone is confronted with a question that requires memory retrieval, the subject experiences a feeling of knowing. This feeling precedes memory retrieval or reconstruction and often predicts the subject's ability to succeed in this mental action (Koriat, 2000; Reder, 1996; Reder & Ritter, 1992). For instance, after being asked about the name of the capital of Peru, the subject may undergo a feeling of knowing that affords remembering the name (Dokic, 2012). Importantly, the feeling of knowing does not merely predict successful retrieval. By presenting the memory task as potentially solvable, it motivates the agent to initiate or persist in the retrieval attempt. In this sense, it functions as an affordance for remembering. Thus, the feeling of knowing potentiates the mental action of trying to reconstruct a past episode.

In the *intermediate assessment*, that is, during the attempt to reconstruct a piece of information, the subject may feel that the remembering process is easier or harder. For example, when trying to remember a poem (say, Borge's "*Labyrinth"*), the subject may have an experience of ease of retrieval of each verse that affords attempting to remember the verses and the whole poem.[8] According to Koriat's accessibility model (1993, 2000), the accessibility of partial or contextual information relevant to the memory target gives rise to this feeling. In contrast, one may also have the feeling of difficulty when the information needed for the reconstruction is not easily accessible. These feelings are more apparent when the reconstruction involves reconstructing longer or more complex bits of information, such as trying to remember a poem or all the places visited on a road trip, and not just a specific name. These feelings afford mental actions related to the process of remembering: the feeling of the ease of retrieval affords trying to remember, whereas the feeling of difficulty in retrieval affords looking for alternative mnemonic strategies.

In the tip-of-the-tongue (TOT) phenomenon, the subject feels that she can retrieve a piece of information that she cannot actually retrieve at present. This experience is puzzling because the prediction of the possibility of memory reconstruction (*self-prediction*) and the actual reconstruction *(intermediate and post-evaluation*) come apart. In other words, the metacognitive system predicts, by a feeling of knowing, that the subject can reconstruct a piece of information but, then, she realizes that she cannot, "demonstrating that the processes of retrieval are not identical to the processes that elicit the TOTs" (Schwartz, 2002, p. 17). From our perspective, the significance of TOT experiences lies in their motivational profile. Even when retrieval fails, the feeling continues to present the information as accessible and thereby motivates continued search. The affordance remains operative despite the temporary failure of retrieval.

Another feeling that we often encounter in our daily life, during the *intermediate and post-evaluations*, is the feeling of forgetting. Consider the case in which one is leaving one's house: in this context, one often has the feeling that one is forgetting something. This feeling affords the revision of one's memory (what am I forgetting?). Ordinary experience suggests that this feeling is often worth heeding, given that one does not want to forget something important (say the keys or the credit card). Moreover, there is some empirical evidence that this feeling is reliable (Halamish *et al.*, 2011). In other contexts, such as when I'm writing a paper and need to remember a reference, the feeling of forgetting affords the use of external mnemonic tools such as a notebook, a reminder list, or just the internet (Arango-Muñoz, 2013).

Across the different stages of remembering, metacognitive feelings play a central role in structuring mental action. Feelings of knowing characterize self-prediction by presenting a memory task as potentially solvable and motivating retrieval attempts. Feelings of ease and difficulty regulate the ongoing process of reconstruction, guiding persistence or strategic adjustment during intermediate assessment. Feelings of forgetting, in turn, often contribute to post-evaluation by motivating the revision of the retrieval process and the deployment of alternative mnemonic resources. These experiences do more than provide information about memory performance: they present possibilities for mental action, motivate the selection

[8] "No habrá nunca una puerta. Estás adentro / y el alcázar abarca el universo / y no tiene anverso ni reverso / ni externo muro ni secreto centro..." ("There'll never be a door. You are inside / and the fortress contains the universe / and has no other side nor any back / nor any outer wall nor secret core...").

of mnemonic strategies, and regulate ongoing retrieval attempts. In this respect, they satisfy both the sensitivity and potentiation requirements of the MAH and therefore constitute genuine affordances for cognitive agency.

4.2. Metacognitive feelings related to reasoning

The notion of inference has been widely discussed in recent decades. According to the classical literature, inferences are rule-based transitions between thoughts. For example, my perception that the street is wet moves me to judge that it rained last night. Mental transitions of thought are guided by rules of inference, for example *modus ponens* or *modus tollens,* where the subject infers q from p with the goal of acquiring a true belief about q. Metacognitive feelings also play a crucial role in the way we perform inferences and other reasoning processes. De Sousa (2009) claims that epistemic feelings play an important role in shaping our beliefs and decisions by providing emotional guidance of our inferential processes.

During a reasoning process we can also distinguish the three functional stages: *self-prediction, intermediate assessment* and *post-evaluation*. In reasoning, each stage is associated with a different type of metacognitive feeling: the feeling of inferential ability evaluates our ability to generate new inferences or solve reasoning problems (*self-prediction*). The feeling of reliability monitors the performance of inference and determines whether it is on track to reach the desired goal, allowing for adjustments to the cognitive strategy if necessary (*intermediate assessment*). Finally, the feelings of confidence, rightness, and endorsement evaluate the result of the reasoning episode and afford accepting the conclusion of a reasoning episode (*post-evaluation*). In the following paragraphs, we will analyze each of these types of feelings related to reasoning.

According to cognitive psychology, reasoning ability develops through the acquisition of inferential skills, including the capacity to recognize logical relations between propositions, distinguish validity from empirical truth, and reason from hypothetical premises (Markovits & Vachon, 1989; Moshman & Franks, 1986). As these abilities develop, agents become increasingly sensitive to their own capacity to solve reasoning problems. We suggest that this sensitivity is often experienced as a feeling of inferential ability: a metacognitive feeling through which agents evaluate, prior to engaging in a reasoning task, whether they are likely to succeed in generating an inference or reaching a solution. This feeling does more than provide information about one's cognitive capacities. By presenting a reasoning problem as manageable or solvable, it motivates engagement with the task and guides the initiation of inferential activity. In this respect, the feeling of inferential ability functions as an affordance for reasoning.

The feeling of inferential ability may be understood by appealing to Chudnoff's notion of "inferential seeming" (Chudnoff, 2024). Chudnoff argues that there is a type of phenomenal experience (an inferential seeming) that occurs in the context of drawing a conclusion from some premises. According to him, there is "a felt difference between following an argument to its conclusion and keeping up with an argument in your judgments while failing to see how its conclusion follows from its premises" (Chudnoff, 2024, p. 315). Thus, we propose that when confronted with a familiar inferential problem, one may undergo the feeling of inferential ability that affords trying to solve the problem. One feels that one knows how to address and solve the problem. In contrast, when confronted with an unfamiliar type of reasoning problem, one may feel confused about how to even start solving the problem. For example, the student might experience a feeling of confusion when faced with the problem $x^3 \cdot x^2 = x^5$ if she has not been trained to multiply powers with the same base by adding their exponents (x^{3+2}). The feeling of confusion may be discouraging in some cases, but it also affords more thorough information processing or looking for alternative strategies to solve the problem (either in a book, on the internet, or by asking the teacher) (Vazard & Audrin, 2022). This feeling does not merely signal the absence of understanding. It motivates specific cognitive responses, such as slowing down, searching for additional information, or adopting alternative problem-solving strategies. In this sense, it functions as a mental affordance.

During the unfolding of the reasoning process (that is, during the intermediate assessment), individuals may experience feelings of reliability or unreliability. These feelings pertain to the ongoing monitoring of a reasoning process. If an individual feels that their reasoning is reliable, this will afford continuing to use the chosen argumentative strategy. Oth-

erwise, if an individual feels that their reasoning is unreliable, it will afford reevaluating the argumentative strategy and any mental transitions that are taking place at that moment.

A modified version of the "Rain Inference Example" (Boghossian, 2018) can be used to illustrate how reasoning can be calibrated through explicit evaluation of the inferences being made. For instance, someone infers that it rained last night from seeing that the street is wet:

1. The street is wet
2. If the street is wet, it rained last night
3. It rained last night

But imagine that before arriving at the conclusion in (3), the subject evaluates the if-then premise (2). Suppose she realizes that it could not have rained last night because it is summer, and the street must be wet for some other reason. Suddenly, the individual gets the feeling of unreliability about the mental transition that is taking place at that moment and modifies the reasoning strategy by including other propositions. For example, the subject might consider that her neighbor washed his car last night, which could explain the wetness of the street. The feeling of unreliability thus functions as an affordance for cognitive revision. Rather than merely informing the agent that something has gone wrong, it motivates the search for alternative premises and explanatory hypotheses.

1. The street is wet
2. If the street is wet, my neighbor washed his car last night
3. My neighbor washed his car last night

The ability to modify one's reasoning process is shaped by metacognitive feelings of reliability and unreliability. The feeling of reliability affords adherence to the premises and the current reasoning strategy, whereas the feeling of unreliability affords revising both the premises and the reasoning strategy. These feelings express the *intermediate assessment* of one's reasoning processes concerning whether they are on track to achieving a desired goal. In the previous example, the desired goal is to explain why the street is wet even though it is summer. Thus, both the feelings of reliability and unreliability allow the individual to adjust their cognitive strategy according to their monitoring of the reasoning episode. Moreover, according to the theory of dual processes of cognition, the feeling of unreliability may trigger a type 2 process where the agent deliberates about the logical or inferential rules she is following (Thompson, 2009).

Finally, there are three types of epistemic feelings that have been discussed in the literature and occur once one arrives at an inferential conclusion: feelings of confidence, rightness, and endorsement. Although these three types of epistemic feelings occur after the inference has been carried out and enhance the action of accepting the conclusion, each one differs in some respect. The feeling of confidence concerns the truth of the conclusion, the feeling of rightness concerns the validity of the inference, and the feeling of endorsement concerns the epistemic justification of both the conclusion and the inference. We characterize these feelings in detail in the following paragraphs.

The feeling of confidence pertains to the level of certainty one has in the truth of a conclusion.[9] If one does not feel confident in the truth of a conclusion, one may be less likely to rely on it. Moreover, without a strong feeling that one's conclusion is right, one won't accept it as true, and one will hesitate to take actions based on it. Instead, one would carefully revise the reasoning process and consider other options before making decisions.

The feeling of rightness concerns the validity of the inference. It is the feeling that our conclusion follows correctly from the premises and provides us with a sensation that we have done it in a proper manner (Thompson *et al.*, 2011).

[9] Here we are not talking about objective truth, but about the way a subject estimates that a given content of a conclusion is true. According to Reber and Unkelbach (2010), fluency or the ease of processing of some information inclines the subject to consider that it is true.

For instance, when solving a simple linear equation such as $2x + 3 = 7$, a student may experience a feeling of rightness in her answer when inferring that $x = 2$. This feeling affords terminating the calculation and accepting the result. Conversely, if the student experiences a feeling of error, it prompts her to revise her strategy and look for new ways to approach the math problem. As with the other metacognitive feelings, cognitive scientists have demonstrated that the feeling of error is correlated with error, and thus it is a reliable marker of error (Fernández-Cruz *et al.,* 2016). This feeling motivates the student to revise the calculation procedure or even to ask the teacher for help. Thus, the feeling of error affords the revision of the outcome of a reasoning process.

The feeling of endorsement plays a distinct epistemic role when a thinker considers the epistemic justification of a conclusion. Kornblith, for example, holds that the feeling of endorsement is at play when the thinker appreciates that his conclusion is correct and it follows from a set of premises (Kornblith, 2012, p. 143). Along the same lines, Chudnoff (2024) discusses normative characteristics of inferential processing that describe how premises support a conclusion which is justified by the premises.

Although all of these metacognitive feelings seem to occur during the same stage of cognitive processing (i.e., post-evaluation), it is unclear whether the different terms used to describe them refer to three different experiences or to the same underlying one. Some researchers have used one of these concepts while others have used different terms. As we have seen, Thompson has studied the f*eeling of rightness,* whereas Kornblith examines the *feeling of endorsement,* and de Sousa analyzes the *feeling of confidence*. The varied terminology used to describe the feelings elicited after a reasoning outcome—during the post-evaluation stage—reflects the specific interests of different researchers and highlights various aspects of the experience, whether cognitive or epistemological. It is possible that the character of these feelings (the way they are felt) depends on the context and goals of the cognitive task at hand. However, further research is needed to determine whether the differences in the naming reflect actual differences in the phenomenal experience or merely reflect different theoretical perspectives. In other words, it remains unclear whether these three feelings are reducible to a single experience.

Regardless of whether confidence, rightness, and endorsement correspond to distinct metacognitive feelings or different aspects of a single experience, they play a similar functional role in reasoning. They present the outcome of a reasoning process as acceptable, questionable, or in need of revision, thereby guiding subsequent cognitive action. Taken together, the examples discussed in this section illustrate how metacognitive feelings structure reasoning across all three stages of cognitive processing: they motivate engagement with a problem, regulate the unfolding of inferential activity, and guide the acceptance or revision of conclusions. In doing so, they satisfy both the sensitivity and potentiation requirements of the Mental Affordance Hypothesis and therefore constitute genuine mental affordances.

4.3. Metacognitive feelings related to attention

McClelland (2020) identifies attention as one of the mental actions that can be afforded by environmental stimuli. According to the affordance model of perceptual motivation (McClelland & Jorba, 2023), perceptual salience can motivate attentional shifts, making certain features of the environment function as attentional affordances. While this account captures an important aspect of attentional guidance, it focuses primarily on externally driven attention and leaves open the question of how attention is regulated from within.

Cognitive psychology suggests that attention can be modulated both externally and internally. While certain external stimuli possess properties that automatically capture attention ("exogenous attention"), agents can also voluntarily direct their attention toward specific targets ("endogenous attention") (Carrasco, 2011). This distinction is consistent with Wu's (2014) action-centered account of attention. According to Wu, attention is fundamentally a form of selection for action rather than merely the selection of information. On this view, attentional states are individuated by the role they play in coordinating subsequent action. A complete account of attentional affordances should explain not only how attention is guided by salient features of the environment, but also how it is regulated from within. Following Wu's action-centered conception of attention, we argue that metacognitive feelings constitute an important class of internal at-

tentional affordances. Feelings such as curiosity, surprise, interest, boredom, and disorientation can guide the initiation, maintenance, or redirection of attention by making certain informational goals appear more or less worth pursuing.

One useful framework for understanding the role of metacognitive feelings in attention is informational foraging, that is, the activity of seeking and acquiring information from the environment. According to several authors, informational foraging is modulated by feelings such as curiosity and boredom (Bench & Lench, 2013; Carruthers, 2024; Eastwood *et al.*, 2012; Goupil & Proust, 2023; Kurzban, 2016). Within this framework, curiosity is generally characterized as an intrinsic motivation to acquire information, whereas boredom is understood as an aversive state that motivates the pursuit of alternative informational goals. From our perspective, what matters is that both feelings function as internal guides for attentional regulation. Curiosity presents the environment as containing information worth acquiring and thereby affords sustained attention, exploration, and information-seeking. Boredom, by contrast, presents the current task as no longer worth pursuing and affords disengagement and the search for alternative sources of information. In this sense, both feelings function as internal attentional affordances that regulate attentional engagement.

Evidence from habituation studies supports this interpretation. Infants tend to attend longer to novel stimuli than to repeatedly presented ones (Day & McKenzie, 1973).[10] This pattern is commonly interpreted as reflecting the motivational influence of curiosity in the presence of novelty and boredom in the absence of new information. Even if this interpretation is not correct in every detail, these findings are compatible with an affective interpretation of attentional engagement and disengagement. Thus, curiosity and boredom do not merely accompany attentional processes; they actively guide the direction and persistence of attention.

Attention offers a particularly clear example of how metacognitive feelings function as mental affordances. Feelings such as curiosity, interest, boredom, surprise, and disorientation do not merely accompany attentional processes; they regulate them by making agents sensitive to their informational needs and motivating specific patterns of attentional engagement or disengagement. In doing so, they satisfy both the sensitivity and potentiation requirements of the MAH and therefore constitute genuine mental affordances.

Conclusion

We have argued that metacognitive feelings are best understood as mental affordances. Rather than merely informing agents about the state of their cognitive processes, these experiences motivate, guide, and regulate mental action. By making agents sensitive to features of their own cognitive activity while simultaneously motivating specific cognitive responses, metacognitive feelings satisfy the two core requirements of the MAH: sensitivity and potentiation.

Like bodily affordances, metacognitive feelings are relational properties. However, whereas bodily affordances emerge from the relation between an agent and her physical environment, mental affordances emerge from the relation between an agent and her own cognitive system. In our view, metacognitive feelings instantiate such relational properties. They reveal possibilities for mental action relative to the agent's abilities, goals, dispositions, and ongoing cognitive tasks. In this sense, metacognitive feelings map the agent's mental environment by making certain cognitive actions available, salient, and actionable.

A reasonable concern is that identifying metacognitive feelings with mental affordances might either over-intellectualize the feelings or, conversely, oversimplify the complexity of cognitive processes. Our response is that the complexity resides primarily in the subpersonal mechanisms of the cognitive system—memory stores, retrieval algorithms, comparison processes, attentional dynamics—not in the felt experience itself. The metacognitive feeling is not a representation

[10] Curiosity has been observed not only in monkeys but also in various other non-human species, including pigeons and starlings (Fortes *et al.*, 2016; Gipson *et al.*, 2009; Vasconcelos *et al.*, 2015). Many animals engage in activities that suggest exploratory or investigative behavior, using their energy to investigate and understand their environment (Berlyne, 1950, 1966). However, there are some debates about the nature of curiosity, including whether it is metacognitive (Carruthers, 2024; Goupil & Proust, 2023).

of that procedural complexity. Rather, it is the affective, non-conceptual manifestation through which the latent capacities of the cognitive system become available to the agent as opportunities for mental action. For this reason, our proposal offers a more parsimonious alternative to informational models that treat metacognitive feelings as detectors of independently existing cognitive affordances. On such views, the feeling and the affordance are distinct entities. By contrast, we have argued that the feeling itself constitutes the affordance. This avoids unnecessary ontological duplication while preserving the motivational and action-guiding functions that affordances are meant to explain.

Expanding the MAH in this way provides a richer account of mental agency—one that recognizes that opportunities for mental action are not only perceived in the external environment but are also felt from within through the dynamics of our own cognitive lives.

Acknowledgments

Earlier versions of this paper were presented at the VII ALFAn Conference of the Latin American Association for Analytic Philosophy (Universidad de la República, Uruguay, 2024), the online workshop *Cognitive Affordances: Theoretical Foundations and Developments* (UNAM, Mexico, 2024), the *Transcending Borders: International Philosophy Seminar of the Americas* (UNC/Arizona State University, 2024), and the *X Jornada de Filosofía de la mente* (Universidad de Antioquia). We thank the participants for their valuable comments and feedback, as well as the two anonymous reviewers for their insightful comments and suggestions.

Declaration of AI Use

The authors used ChatGPT (OpenAI, GPT-5.5) during the editorial revision of the manuscript as a writing and copy-editing assistant to improve clarity, grammar, style, and internal consistency of the English text. All substantive arguments, interpretations, and editorial decisions were made by the authors, who take full responsibility for the final version of the manuscript.

Funding Statement

This paper is a research output of the project "Mental agency: actions and omissions", funded by Comité para el Desarrollo de la Investigación (CODI) de la Universidad de Antioquia. Project code: 2022-52270.

REFERENCES

Arango-Muñoz, S. (2013). Scaffolded Memory and Metacognitive Feelings. *Review of Philosophy and Psychology*, *4*(1), 135-152. https://doi.org/10.1007/s13164-012-0124-1

Arango-Muñoz, S. (2014). The nature of epistemic feelings. *Philosophical Psychology*, *27*(2), 193-211. https://doi.org/10.1080/09515089.2012.732002

Arango-Muñoz, S. (2019). Cognitive phenomenology and metacognitive feelings. *Mind & Language*, *34*(2), 247-262. https://doi.org/10.1111/mila.12215

Arango-Muñoz, S., & Bermúdez, J. P. (2018). Remembering as a mental action. In K. Michaelian, D. Debus, & D. Perrin (Eds.), *New Directions in the Philosophy of Memory* (pp. 75-96). Routledge.

Bench, S. W., & Lench, H. C. (2013). On the Function of Boredom. *Behavioral Sciences*, *3*(3), 459-472. https://doi.org/10.3390/bs3030459

Berlyne, D. E. (1950). Novelty and curiosity as determinants of exploratory behaviour. *British Journal of Psychology. General Section*, *41*(1-2), 68-80. https://doi.org/10.1111/j.2044-8295.1950.tb00262.x

Berlyne, D. E. (1966). Curiosity and Exploration. *Science*, *153*(3731), 25-33. https://doi.org/10.1126/science.153.3731.25

Boghossian, P. (2018). Delimiting the Boundaries of Inference. *Philosophical Issues*, *28*(1), 55-69. https://doi.org/10.1111/phis.12115

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484-1525. https://doi.org/10.1016/j.visres.2011.04.012

Carruthers, P. (2018). Valence and Value. *Philosophy and Phenomenological Research, 97*(3), 658-680. https://doi.org/10.1111/phpr.12395

Carruthers, P. (2024). The Contents and Causes of Curiosity. *The British Journal for the Philosophy of Science*, *75*(4), 871-895. https://doi.org/10.1086/716609

Chudnoff, E. (2024). Inferential Seemings. In U. Kriegel (Ed.), *Oxford Studies in Philosophy of Mind* (Vol. 4, pp. 311-333). Oxford University Press. https://doi.org/10.1093/9780198924159.003.0009

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58*(1), 7-19. https://doi.org/10.1093/analys/58.1.7

Corns, J. (2014). Unpleasantness, Motivational *Oomph*, and Painfulness. *Mind & Language, 29*(2), 238-254. https://doi.org/10.1111/mila.12048

Day, R. H., & McKenzie, B. E. (1973). Perceptual Shape Constancy in Early Infancy. *Perception*, *2*(3), 315-320. https://doi.org/10.1068/p020315

de Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, *191*(2), 155-185. https://doi.org/10.1007/s11229-013-0247-7

De Sousa, R. (2009). Epistemic Feelings. *Mind and Matter*, *7*(2), 139-161.

Dokic, J. (2012). Seeds of self-knowledge: Noetic feelings and metacognition. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *The Foundations of Metacognition* (pp. 302-321). Oxford University Press.

Dorsch, J. (2025). Are Noetic Feelings Embodied? The Case for Embodied Metacognition. *Philosophical Psychology, 38*(2), 625-647. https://doi.org/10.1080/09515089.2023.2197937

Eastwood, J. D., Frischen, A., Fenske, M. J., & Smilek, D. (2012). The Unengaged Mind: Defining Boredom in Terms of Attention. *Perspectives on Psychological Science*, *7*(5), 482-495. https://doi.org/10.1177/1745691612456044

Fernández Velasco, P., & Loev, S. (2025). Metacognitive Feelings: A Predictive-Processing Perspective. *Perspectives on Psychological Science*, *20*(4), 691-713. https://doi.org/10.1177/17456916231221976

Fernández-Cruz, A. L., Arango-Muñoz, S., & Volz, K. (2016). Oops, scratch that! Monitoring one's own errors during mental calculation. *Cognition*, *146*, 110-120. https://doi.org/10.1016/j.cognition.2015.09.005

Fortes, I., Vasconcelos, M., & Machado, A. (2016). Testing the boundaries of "paradoxical" predictions: Pigeons do disregard bad news. *Journal of Experimental Psychology: Animal Learning and Cognition*, *42*(4), 336-346. https://doi.org/10.1037/xan0000114

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Psychology Press.

Gipson, C. D., Alessandri, J. J. D., Miller, H. C., & Zentall, T. R. (2009). Preference for 50% reinforcement over 75% reinforcement by pigeons. *Learning & Behavior*, *37*(4), 289-298. https://doi.org/10.3758/LB.37.4.289

Goldwasser, S. (2023). Memory as Skill. *Review of Philosophy and Psychology*, 14, 833-856. https://doi.org/10.1007/s13164-021-00605-x

Goupil, L., & Proust, J. (2023). Curiosity as a metacognitive feeling. *Cognition*, *231*, 105325. https://doi.org/10.1016/j.cognition.2022.105325

Halamish, V., McGillivray, S., & Castel, A. D. (2011). Monitoring One's Own Forgetting in Younger and Older Adults. *Psychology and Aging, 26*(3), 631-635. https://doi.org/10.1037/a0022852

Jorba, M. (2020). Husserlian horizons, cognitive affordances and motivating reasons for action. *Phenomenology and the Cognitive Sciences*, *19*(5), 847-868. https://doi.org/10.1007/s11097-019-09648-z

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609-639. https://doi.org/10.1037/0033-295X.100.4.609

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, *9*(2), 149-171. https://doi.org/10.1006/ccog.2000.0433

Kornblith, H. (2012). *On reflection* (1st ed). Oxford University Press.

Kozuch, B. (2020). No Pain, No Gain (in Darwinian Fitness): A Representational Account of Affective Experience. *Erkenntnis*, *85*(3), 693-714. https://doi.org/10.1007/s10670-018-0044-2

Kurzban, R. (2016). The sense of effort. *Current Opinion in Psychology*, *7*, 67-70. https://doi.org/10.1016/j.copsyc.2015.08.003

Loev, S. (2022). Affectivism about Intuitions. *Synthese*, *200*(4), 274. https://doi.org/10.1007/s11229-022-03749-0

Markovits, H., & Vachon, R. (1989). Reasoning with contrary-to-fact propositions. *Journal of Experimental Child Psychology*, *47*(3), 398-412. https://doi.org/10.1016/0022-0965(89)90021-0

McClelland, T. (2020). The Mental Affordance Hypothesis. *Mind*, *129*(514), 401-427. https://doi.org/10.1093/mind/fzz036

McClelland, T., & Jorba, M. (2023). Perceptual Motivation for Action. *Review of Philosophy and Psychology, 14*, 939-958. https://doi.org/10.1007/s13164-021-00608-8

Metzinger, T. (2017). The problem of mental action: Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing* (pp. 1-26). MIND Group. https://doi.org/10.15502/9783958573208

Michaelian, K. (2016). *Mental Time Travel: Episodic Memory and our Knowledge of the Personal Past*. The MIT Press.

Moshman, D., & Franks, B. A. (1986). Development of the Concept of Inferential Validity. *Child Development*, *57*(1), 153-165. https://doi.org/10.2307/1130647

Norman, E., Price, M. C., & Duff, S. C. (2010). Fringe consciousness: A useful framework for clarifying the nature of experience-based metacognitive feelings. In A. Efklides & P. Misailidi (Eds.), *Trends and prospects in metacognition research* (pp. 63-80). Springer. https://doi.org/10.1007/978-1-4419-6546-2_4

Osiurak, F., & Badets, A. (2017). Use of tools and misuse of embodied cognition: Reply to Buxbaum (2017). *Psychological Review*, *124*(3), 361-368. https://doi.org/10.1037/rev0000065

Peacocke, A. (2021). Mental action. *Philosophy Compass*, *16*(6), e12741. https://doi.org/10.1111/phc3.12741

Proust, J. (2001). A Plea for Mental Acts. *Synthese*, *129*, 105-128. https://doi.org/10.1023/A:1012651308747

Proust, J. (2013). *The Philosophy of Metacognition: Mental Agency and Self-Awareness*. Oxford University Press.

Proust, J. (2015a). The Representational Structure of Feelings. In T. Metzinger & J. M. Windt (Eds.), *Open MIND*, *31*, 1-26. https://doi.org/10.15502/9783958570047

Proust, J. (2015b). Feelings as evaluative indicators: A reply to Iuliia Pliushch. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (31[R], pp. 1-7). MIND Group. https://doi.org/10.15502/9783958570955

Proust, J. (2024). Affordances from a control viewpoint. *Philosophical Psychology,* 37(7), 1590-1614. https://doi.org/10.1080/09515089.2023.2274489

Reber, R., & Unkelbach, C. (2010). The Epistemic Status of Processing Fluency as Source for Judgments of Truth. *Review of Philosophy and Psychology*, *1*, 563-581. https://doi.org/10.1007/s13164-010-0039-7

Reder, L. M. (Ed.) (1996). *Implicit Memory and Metacognition*. Lawrence Erlbaum.

Reder, L. M., & Ritter, F. E. (1992). What Determines Initial Feeling of Knowing? Familiarity with Question Terms, Not with the Answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 435-451. https://doi.org/10.1037/0278-7393.18.3.435

Rietveld, E., & Kiverstein, J. (2014). A Rich Landscape of Affordances. *Ecological Psychology*, 26(4), 325-352. https://doi.org/10.1080/10407413.2014.958035

Schwartz, B. L. (2002). *Tip-of-the-tongue states: phenomenology, mechanism and lexical retrieval*. Lawrence Erlbaum Associates.

Siegel, S. (2014). Affordances and the contents of perception. In B. Brogaard (Ed.), *Does perception have content?* (pp. 39-76). Oxford University Press.

Teroni, F. (2018). Emotionally Charged—The Puzzle of Affective Valence. In C. Tappolet, F. Teroni, & K. Ziv (Eds.), *Shadows of the Soul: Philosophical Perspectives on Negative Emotions* (pp. 10-19). Routledge.

Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In J. St. B. T. Evans & K. Frankish (Eds.), *In Two Minds: Dual Processes and Beyond* (pp. 171-196). Oxford University Press.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107-140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Vasconcelos, M., Monteiro, T., & Kacelnik, A. (2015). Irrational choice and the value of information. *Scientific Reports*, *5*(1), 13874. https://doi.org/10.1038/srep13874

Vazard, J., & Audrin, C. (2022). The noetic feeling of confusion. *Philosophical Psychology*, *35*(5), 757-770. https://doi.org/10.1080/09515089.2021.2016675

Wu, W. (2014). *Attention.* Routledge.

Zipoli Caiani, S. (2014). Extending the notion of affordance. *Phenomenology and the Cognitive Sciences*, *13*(2), 275-293. https://doi.org/10.1007/s11097-013-9295-1

Santiago Arango-Muñoz is Associate Professor of Philosophy at the Universidad de Antioquia (Colombia). His research focuses on philosophy of mind, metacognition, epistemic feelings, and the philosophy of cognitive science.

Address: Instituto de Filosofía, Universidad de Antioquia, Ciudad Universitaria, Block 12, Office 408, Calle 70 No. 52-21, Medellín, Colombia. E-mail**:** santiago.arango5@udea.edu.co — ORCID**:** 0000-0003-3761-3487

Sebastián Mejía-Rendón is a CONICET Doctoral Fellow at the Institute of Humanities (IDH), Universidad Nacional de Córdoba (Argentina). His research focuses on philosophy of mind, philosophy of technology, animal cognition, metacognition, and cognitive artifacts, with particular interest in tool use and metacognitive feelings in non-human animals.

Address: Institute of Humanities (IDH), CONICET – Universidad Nacional de Córdoba, Faculty of Philosophy and Humanities, Córdoba, Argentina. E-mail: joan.mejia@mi.unc.edu.ar — ORCID: 0000-0001-7356-1181

Theoria, 2026, 41(3), 396-433
https://doi.org/10.1387/theoria.27610

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Illusionism as a theory of mental content

(El ilusionismo como teoría del contenido mental)

Marco Facchin*
University of Antwerp
https://orcid.org/0000-0001-5753-9873

Keywords

Mental content
Illusionism
Mind-body problem
Eliminativism
Intentionality
Radical enactivism

ABSTRACT: Analytic philosophers of mind seemingly take the hard problem of consciousness to be the only aspect of the mind-body problem that has not been in principle solved. But phenomenal consciousness is not the only puzzling and mysterious mental phenomena. The capacity of our minds to have contents — to "be about" or "of" objects, and to represent them — is also puzzling and mysterious. And whilst, from the '80s on, it was largely assumed that mental contents could have been easily naturalized, the repeated and systematic failures of naturalistic theories of content invite us to search for an alternative approach. This paper aims to introduce one such alternative approach, which I dub content illusionism. Content illusionism revolves around two main claims. One is an eliminative claim, according to which mental contents do not exist. The other is the illusion claim, according to which there is an illusion of mental content: contents appear to exist and be "carried around" by our mental states. The paper offers a general presentation of content illusionism, explaining its core claims, showing that it is a compelling philosophical view of content, and defending it from some intuitive objections.

Palabras clave

Contenido mental
Ilusionismo
Problema mente-cuerpo
Eliminativismo
Intencionalidad
Enactivismo radical

RESUMEN: Los filósofos analíticos de la mente parecen dar por sentado que el problema difícil de la conciencia es el único aspecto del problema mente-cuerpo que aún no ha sido resuelto. Sin embargo, la conciencia fenoménica no es el único fenómeno mental desconcertante y misterioso. La capacidad de nuestra mente para tener contenidos —para «versar sobre» o «ser acerca de» objetos, y para representarlos— resulta igualmente enigmática. Y si bien desde los años ochenta del siglo XX se dio por hecho que los contenidos mentales podrían naturalizarse sin mayor dificultad, los fracasos repetidos y sistemáticos de las teorías naturalistas del contenido nos invitan a buscar un enfoque alternativo. El presente artículo se propone presentar precisamente uno de esos enfoques alternativos, al que denomino ilusionismo del contenido. Esta posición se articula en torno a dos tesis principales. La primera es de carácter eliminativista y sostiene que los contenidos mentales, en rigor, no existen. La segunda es la tesis de la ilusión, según la cual existe una ilusión de contenido mental: los contenidos parecen existir y ser «transportados» por nuestros estados mentales. El artículo ofrece una presentación general del ilusionismo del contenido, explica sus tesis centrales, muestra que se trata de una posición filosóficamente atractiva sobre el contenido, y la defiende frente a algunas objeciones de corte intuitivo.

Gako-hitzak

Eduki mentala
Ilusionismoa
Gogo-gorputz auzia
Eliminatibismoa
Intentzionalitatea
Enaktibismo erradikala

LABURPENA: Dirudienez, gogoaren filosofo analitikoentzat, kontzientziaren auzi zaila da printzipioz konpondu gabe dagoen gogo-gorputz auziaren alderdi bakarra. Alabaina, kontzientzia fenomenikoa ez da fenomeno mental harrigarri eta misteriotsu bakarra. Gure gogamenak *edukiak izateko* —alegia, «objektuei buruz» edo «objektuez aritzeko», eta haiek irudikatzeko— duen gaitasuna ere harrigarri eta misteriotsua da. Eta 1980ko hamarkadatik aurrera neurri handi batean ontzat eman bazen ere eduki mentalak erraz naturaliza zitezkeela, edukiaren teoria naturalisten porrot errepikatu eta sistematikoek ikuspegi alternatibo bat bilatzera bultzatzen gaituzte. Artikulu honetan, halako ikuspegi alternatibo bat aurkeztu nahi dut, nik *edukiaren ilusionismo* izendatu dudana. Edukiaren ilusionismoa bi tesi nagusiren inguruan egituratzen da. Lehenengoak izaera baztertzailea du, eta, horren arabera, ez dago eduki mental hertsirik. Bigarrena ilusioaren tesia da, eta, horren arabera, eduki mentalaren ilusio bat dago: badirudi edukiak existitzen direla eta gure egoera mentalek "garraiatzen" dituztela. Artikuluan, edukiaren ilusionismoaren aurkezpen orokorra egiten da; haren tesi nagusiak azaltzen dira, edukiari buruzko ikuspuntu filosofiko sinesgarria dela erakusten da, eta objekzio intuitibo batzuen aurrean defendatzen da.

* ***Correspondence to:*** Marco Facchin. Department of Philosophy and Letters, Faculty of Philosophy, Centre for Philosophical Psychology, Universiteit Antwerpen, Building R, Rodestraat 14, Antwerpen 2000, Belgium. – marco.facchin.marco.facchin@gmail.com / marco.facchin@uantwerp.be – https://orcid.org/0000-0001-5753-9873

How to cite: Facchin, Marco (2026). «Illusionism as a theory of mental content»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 41(3), 396-433. (https://doi.org/10.1387/theoria.27610).

Received: 11 June, 2025; Final version: 27 June, 2026.
ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press

This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. A manifesto

Most philosophers take the mind-body problem to have been reduced to the hard problem of consciousness — the problem of explaining why any subjective, *experiential feeling* accompanies our neural activities. The rest of the mind has been figured out — *at least in principle* (Chalmers, 1997).

Yet, phenomenal consciousness is not the only surprising, unique and "weird" feature of the mind. The ability of minds to have mental contents — their capacity to be about, "directed at", or represent things — is quite puzzling too. Think that 2+2=4. How did your mind manage to grasp an abstract mathematical truth, presumably outside time and space? Look at these words. How can your perceptual state "contain" them, making you aware of them and their meaning? Recall any dream of yours: how could such a weird and complex scenario have been shown to your mind's eye?

From the 80's on, a group of philosophers systematically tackled these questions. Their objective was to naturalize mental content — that is, explaining *why* mental states carry the contents they carry by resorting only to naturalistically acceptable, ontologically unsuspicious ingredients (e.g. Millikan, 1984; Fodor, 1987; Dretske, 1988). 40 years later, the problem is still open. The inability of naturalistic approaches to deal with the problem motivates a novel exploration of our conceptual landscape, and the search for an alternative approach to content (see Kriegel, 2011; Egan, 2020).

Here, I sketch *content illusionism* — one such alternative approach. This approach is indebted to Dennett (1991b; 2005) Frankish's (2016; 2023; 2026) and Kammerer (2021) illusionism about phenomenal consciousness, and revolves around the idea that content is an illusion: something that *merely appears* to exist, without *really* existing. If content illusionism is correct, then, our minds are not "about" nor "represent" anything — our mental states come before the mind's eye empty handed, so to speak. Yet, they don't *seem* empty handed. We won't *naturally*, *naively* describe them as such. On the contrary, they appear to "be about" or "represent" many things, sometimes very complicated ones.

Note that, just like its phenomenal counterpart, content illusionism isn't a *single* view, but a *family* of possible views with varying degrees of kinship and similarity (Frankish, 2016; 2023). So, in sketching content illusionism, I am sketching a *motif* (Raja, 2024): a set of usefully vague and open-ended ideas that guide a specific set of epistemic practices. Here, I wish to sketch such a *motif*, introducing it to the literature and showing that it is an attractive position. So, I won't articulate a full-fledged philosophical theory, nor offer a direct argument to the effect that content is an illusion. For such an argument, see (Facchin, 2025).

I shall thus begin by clarifying the terms of the debate, fixing the reference of "mental content" (§2.1); and presenting the core claims of content illusionism (§2.2). I shall then motivate content illusionism, showing that it is a compelling alternative to traditional views of content (§3), and deflect some obvious objections it faces (§4). I will then quickly substantiate the motif, offering a how-possibly explanation of the illusion of content (§5), and then conclude this manifesto by indicating some directions for future research (§6).

2. Presenting content illusionism

2.1. On "content" and its features

Content illusionism claims that mental content — henceforth, just "content" — is an illusion. But what is this content that's supposed to be an illusion? What does "content" mean?[1]

[1] Owing to space limitations, and the programmatic nature of the paper, the discussion of these matters will be highly schematic. See Facchin, 2025) for a more in-depth discussion.

As now customary (e.g. Shea, 2018; Mendelovici, 2018) I fix the reference of "content" by *ostension*. Looking at paradigmatic and uncontested[2] examples of contentful and contentless mental states, it should be possible to identify some natural and salient property states of the first type all have and states of the second type all lack, which is thus what "content" designates (cf. Schwitzgebel, 2016; 2024, ch.8).

As a first example of paradigmatic contentful mental states, consider *perceptual states* such as seeing an apple, hearing an ambulance passing, or smelling the dinner burning in the oven. Notice how, in all these cases, there is some object[3] being perceived — an apple, the passing ambulance, or the dinner turning into charcoal. Other perceptual-like mental states, such as hallucinations, misperceptions, or bouts of imagings share the same feature too, and are indeed paradigmatically contentful.

Consider now *occurrent thoughts* — the kind of sentences we "say aloud" in our heads when we think. Right now, you are "saying aloud" this very sentence — if you are not distracted. So, there is something that you are "innerly saying aloud", something that you are thinking. And the same holds for other explicit acts of thought like judging, deciding, evaluating, considering and inferring.

Consider lastly propositional attitudes — *the* paradigmatic example of contentful mental states (e.g. Fodor, 1987). Believing that p, desiring that q, predicting that r and intending to ϕ all have some sort of object that is believed, desired, predicted or intended introduced by the "that" or "to".

Consider now some paradigmatic *contentless* states. A first paradigmatic case is offered by *non-mental* states, like having a cavity, a broken bone, or a cold. Whilst one might be conscious of these states in virtue of their symptoms, none of these states is considered to carry mental content, on account of being non-mental states.

As a second paradigmatic case, consider now some psychophysical states, such as being hungover, being horny, being thirsty/hungry, being stressed, or being burned out. These states have a rich and complex phenomenology, and are thus mental states in a prominent and clear sense of the term. Yet they lack any object — when you are perceiving, there is something perceived. But when you are hungover, there is nothing *hungovered*, so to speak.

Consider, lastly, moods such as elation, depression, nervousness, or discomfort. These states are clearly mental and are recognized as such by everyone. They have a rich phenomenology, and a complex impact on our behavior. And yet, there does not seem to be anything *elated* in elation, *depressed* in depression or "nervoused" in nervousness; at least, not in the way anything is perceived in perception, or believed in belief.[4]

What sort of visible and salient property is universally present among contentful states and universally absent among contentless states? An *essential link with an object*. All contentful states are partially identified by it. Change the type

---

[2] Almost uncontested - content eliminativists may disagree (e.g. Rosenberg, 2019). But they too would concede that, if content existent, these states would be contentful.

[3] I'm using the term "object" a bit loosely here - the relevant object could be a state of affair, a property, or something else. Still, there is *something* being perceived.

[4] As the reviewers noticed, such a view is not universally held by philosophers, and, popularity-wise, it might even be on the back foot. Crane (1998), for example, called upon the phenomenological tradition (especially Sartre, 1971) to argue that the intentional object of such moods is *the world*. Roughly, just as my (and your) current perceptual state has *this sentence* as its object, or my fear of spiders is about spiders, the depression of the depressed person is "about" the world, or has the world as its object. I must confess I don't really see how depression "is about" the world, or how it has the world as its object in a way that is identical (or even roughly analogous) to the way perceptual states and emotions have objects or "are about" objects. So, I must confess that I cannot but regard that claim as false. I also suspect that the claim is generated by an overextension of a *different* claim coming from the phenomenological tradition, namely that our emotional states (moods included) are partially constitutive of our "openness" to the world; that is, of our *intentional* contact with the world (Heidegger, 1962; see also De Haan, *et al.* 2013 for an updated, "neurophenomenological" rendition of that claim). I find this second claim much more plausible - and indeed I accept the view that (at least some) emotions appear to be contentful. My fear of spiders is indeed *about* spiders, and has spiders as its objects.

of link, and you turn a belief into a desire or a perception into a fantasy. Change the object, and you turn a belief that p into a belief that q, the perception of an orange in a perception of an apple. Nothing of the sort holds for contentless states. Since there is nothing *depressed*, there is nothing whose change turns a depression of a into a depression of *b* — because depressions are not "of" nor "about". And since there is no such object, there can be no linkage with it whose change turns depression into elation, or burnout into hungover.[5]

Note that, strictly speaking, the examples above are examples of contentful and contentless *mental states*, rather than *content*. So, the above fixes the reference of "contentful" and "contentless" mental states. From these, it is easy to fix the reference of "having content" — to have content is to have the linkage with an object contentful states paradigmatically display. But once the reference of "having content" is given, the reference of "content" is too — "content" refers to the relevant objects contentful mental states are essentially linked to.

Is this usage of "content" idiosyncratic? Quite the contrary. Consider:

> "When we introspectively notice intentional states [...] we also notice something we are tempted to describe as what our mental states are "directed at" or what they "say"; *this is their (intentional) content*." (Mendelovici, 2018, p. 8; italics added)
>
> "There is some need for clarification in the literature regarding how we should understand the term 'content' and what a theory of content is about. The term 'content' is commonly used to *refer to the intentional object of a representation* – the thing (or property, abstract entity, etc.) represented." (Ramsey, 2023, p. 2 fn 1; italics added)

Mendelovici and Ramsey have very different philosophical affiliations: the former comes from analytic, "armchair" philosophy of mind, and the latter from (post?)analytic, empirically informed, philosophy of cognitive science. Yet they use "content" identically. Brentano reminds us this usage of "content" is way older than analytic philosophy, and can be traced at least back to the Scholastics[6]:

> "Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, *reference to a content*, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity." (Brentano 2012, p. 88, italics added)

This usage of "content", then, is common, and hides no sleight of hand. Indeed, it is so common that it "leaked" onto the discussion of sub-personal representational posits, to designate what these posits are about (e.g. Neander, 2017; Williams, 2019; Piccinini, 2022).[7] I won't challenge this usage here: after all, even subpersonal representational posits seem to have that kind of essential tie to an object.

Still, this usage of "content" is ambiguous. At times, the world seems to pick up o*rdinary objects*, broadly understood: mind-independent "middle sized dry goods" (cf. Clark, 2009). For example, when I see an apple, think of the colosseum or remember my hotel room in Nuremberg, the content of these states seems to be some ordinary physical object. Yet, other mental states suggest otherwise. When, for example, one hallucinates the proverbial pink elephant, or misperceives a dog turd as a man[8] there is no "middle sized dry good" corresponding to what one misperceived. It is natural, in such cases, to think of content as "mental things" of sort — as *sui generis* objects "in" the mind and "shown" to the

---

5 It is however plausible such changes can be executed by changing the relevant phenomenal features of these states.

6 A reviewer (whom I thank) informed me that it is possible to find even earlier discussions of intentionality, such as in the Nyāya school of philosophy; see (Das, 2021).

7 Presumably, the relevant sub-personal representations also include neural representations, even if they are supposed to be observables rather than posits (Thomason & Piccinini, 2018). For further discussion, see (Facchin, 2023; Drayson, 2025).

8 This actually happened; see Clark (2016, p. 8).

mind's eye; and indeed our folk conception of content does at times conceive contents in this way (cf. Crane & French, 2021).

So, are contents *ordinary objects* or not — or, better, given the eliminativistic view the paper will soon introduce: is "content" *supposed to refer* to ordinary objects or not? I am inclined to give a negative answer — based on the (hopefully uncontested) claim that contents have features *very unlike* the features of ordinary objects. Thus notice, for example, that contents are "satisfiable" (Facchin 2025, p. 206): they are (or "bear"; the literature attests both usages[9]) conditions of satisfaction. And they likely do so essentially, as indeed in many cases contentful states must be either satisfied (e.g. accurate, truthful, or fulfilled) or not. Ordinary objects surely are not satisfiable in this way. Contents are also "apparently worldly" (*ibidem*), in that they present themselves as objects existing over and above the mental state "carrying them" — and yet, in many cases (especially when the content is not satisfied) they seem to be internal or immanent to the state itself. Again, ordinary objects don't seem to possess this feature.

I don't think that this conclusion is (or, should be) controversial. Most philosophers don't take contents to be ordinary objects. This is presumably why philosophers strive to *naturalize* them — ordinary objects stand in no need of naturalization. And notice that one doesn't *need* to buy into some Meinongian theory of non-actual but existent objects to hold the view that contents are not ordinary objects. Taking contents to be *propositions* (e.g. Frege, 1918; Fodor, 1987) or possible worlds (e.g. Wittgenstein, 2001; Lewis, 1973; Kiefer & Howhy, 2018) is enough to take contents *not* to be ordinary objects.

So, "content" designates the objects our contentful mental states (and sub-personal representational states) are essentially linked to, as weird and "non-ordinary" as these objects may be. What does content illusionism say about *them*?

2.2. The core claims of content illusionism

Content illusionism is a (family of) theory (or theories) of content. But "theory of content" is ambiguous between two readings (Ramsey, 2023):

> **Reading (1):** Theories of content are meta-semantic theories, detailing the factors in virtue of which a contentful state carries the content it carries.
>
> **Reading (2):** Theories of content are metaphysical theories, accounting for the ultimate nature of contents.

Reading (1) today prevails: mainstream naturalistic theories of content are theories of content in this sense (e.g. Millikan, 1984; Shea, 2018).[10] And so are their non-naturalistic alternatives (e.g. Bourget & Mendelovici, 2019). Reading (2) is, in contrast, very *ageé*: it calls to mind "Meinong's object theory" (Meinong, 1960) or James's view of contents as physical particulars we bodily relate to (James, 1911).

Shamefully, content illusionism contributes to the confusion concerning the meaning of "theory of content" because it is not a theory of content in any of these two senses. Indeed, it consists of two main claims, the first of which is that *content should be eliminated* — a claim that uneasily sits with both reading (1) and (2).

How should this eliminative claim be understood? Eliminativism comes in two basic variants — entity eliminativism and discourse eliminativism (Sprevak & Irvine, 2020). Entity eliminativism claims that something does not exist — it makes a negative existential claim. Discourse eliminativism makes no negative existential claim (or at least, it's not com-

---

[9] See Zahnoun (2023).

[10] This doesn't, however, mean that such theories are metaphysically agnostic about content. As noticed above, they often take (or end up taking) contents to be propositions or possible worlds.

pelled to make any). Rather, it makes a *linguistic* claim; namely, that we should opt out of certain discursive practices involving the to-be-eliminated word. The eliminative claim advanced by content illusionism should be read as advancing a form of *entity* eliminativism:

> **Eliminative claim:** There are no mental contents.

The eliminative claim entails that the reference of "content" is empty (since there are no contents). And because of this, the eliminative claim *entails* (but is *not* identical to) a weak form of discourse eliminativism, at least for scientific realists.

This is because contents (and the representational states "carrying" them[11]) allegedly play a number of relevant roles in the modern mind sciences. For instance, they are called upon to explain specific patterns of successes and failures of an agent (e.g. Godfrey-Smith, 2006; Shea, 2018) and in identifying common, abstract patterns of response or functioning that are exhibited despite the physical difference of the stimulus or input conditions (e.g. Dennett, 1987; 1991). Now, if one is a scientific realist, one should also discursively eliminate content from these explanations (Ramsey, 2020). The reason is simple: realists take these statements to be ontologically committing. But if the eliminative claim is correct, these commitments are wrongheaded — hence, *they shouldn't be made*. So, the linguistic practices making them should be avoided.

Notice however that the form of discourse eliminativism entailed by the eliminative claim concerns only the *scientific* discourse about the mind.[12] Other relevant forms of content-invoking talk do not bear the relevant ontological commitments, and thus are left untouched. Thus, it does not affect our use of mental contents for *making sense, predicting and rationalizing* the behavior of ourselves and others (e.g. Dennett, 1987; Toon, 2023; Facchin *et al.*, 2025), as these discursive usages of content are entirely possible even without making any ontological commitment to the reality of content. For the same reason, it does not affect usages of content in narrative practices — including self-narration, behavior shaping practices and narrative practices regulating social coordination (Dennett, 1992; Demeter, 2022; Hutto, 2022). Thus, the form of discourse eliminativism entailed by the eliminative claim is remarkably moderate, far removed from Churchlandesque fantasies of all-encompassing linguistic reforms.[13] Indeed, as I shall explain, certain everyday usages of content *cannot* be coherently eliminated by the illusionist, on the pain of contradiction. This is because they are *constitutive* of the illusion of content (see § 5).

The *what*? Well, the illusion of content is what tells apart content illusionism from content eliminativism and fictionalism about mental content. Content eliminativists only say that content does not exist[14] (e.g. Rosenberg, 2019), leaving unexplained why we *think* it exists and behave as if it does. And fictionalism is *primarily* a claim concerning how we should interpret the semantics and/or pragmatics of our content-involving discourse — it's a piece of philosophy of language that, whilst of course sympathetic to anti-realism, neither entails nor is entailed by it (Gan, 2024). *Strictly speaking*, fictionalism has no ontological strings attached.

In contrast, content illusionism *has* ontological strings attached, and not merely negative ones. For, the second main claim of content illusionism is a positive existential claim, the illusion claim:

> **Illusion claim:** There is an illusion of content: whilst there are no mental contents, there seem to be mental contents "given" before our mind's eye.[15]

---

[11] Throughout the paper, I shall assume that representational states *constitutively* involve content. This seems an innocent assumption, given the notion of content at play.

[12] And only on the assumption that some form of scientific realism is correct. Fictionalists won't be troubled by this (see Sprevak, 2013).

[13] Fantasies that, to be fair to the Churchlands, they now have wisely abandoned (e.g. Churchland, 2012).

[14] Or, more often, that *some* contents in *some* mental states or processes do not exist (e.g. Churchland, 1992; Hutto & Myin, 2013).

[15] Or equivalently: content is *seemingly given* to our minds.

What is this illusion, and how should it be characterised? And why is it an illusion, and not just a simple mistake, or an hallucination?

Let's start with the second question. Hallucinations and mistakes are *subjective* and *asystematic*. They are subjective, because they are factually not shared by subjects. My hallucinations, my perceptual mistakes are *uniquely* mine, and it would be very odd if we were *both* hallucinating a pink elephant, or *both* mistaking an apple for a watermelon. Also, hallucinations and mistakes are asystematic. There is no stimulus that reliably causes me to hallucinate or to misperceive. In contrast, illusions are *intersubjectively concordant* and *systematic* — and that's what distinguishes them from simple hallucinations and perceptual errors (Macpherson, 2013; Mandik, 2016). Seeing the two lines of the Müller-Lyer makes *everyone* see two lines of different length. Looking at the complex arrangements invented by Ames makes everyone see rotating windows as oscillating (the so-called "Ames window" illusion), or people changing size when walking into a room (the "Ames room", see Ittelson, 1952).[16] These illusions are readily produced by "consuming" an appropriate stimulus, and are shared by *anyone* consuming it.[17]

Mental contents display the same intersubjective concordance and systematicity — hence they are aptly called an illusion. At least in normal conditions, when gazing at a perceptual scene, we will all see the same thing – i.e., be in a perceptual state with the same content. Similarly, by hearing the premises $\forall(x)\ f(x) \rightarrow g(x)$ and $f(a)$, we will all conclude that $g(a)$; that is, we will all formulate a judgement with the same content. By hearing "Napoleon" we will all think of a French general turned emperor, and so forth. Sure, there are contextual and agential variation: if one does not know modern formalized logic, reading $\forall(x)\ f(x) \rightarrow g(x)$ and $f(a)$ won't let them conclude that $g(a)$, and if we are at the weekly meeting of our *Animal Farm* reading group, hearing "Napoleon" will make us think of a fictional character. But such changes are not random — they are lawfully based on the context, knowledge, and ability of the subject. Even mental contents of states that are *not* tokened as a direct response to any stimulus are largely identical across subjects: for example, we tend to dream anxiety-inducing scenes (Revonsuo, 2000); and subjects affected by schizophrenia tend to hallucinate voices (e.g. Sartorius *et al.*, 1986). This uniformity is quite remarkable: of *all* the things we could dream (or hallucinate) we dream (or hallucinate) only a very limited, and intersubjectively concordant, portion.

How, however, should this illusion be conceived? This is a pivotal question to answer. After all, our intuitive conception of an illusion is that of a state that *misrepresents* something (e.g. Macpherson, 2013) — that is, we conceive illusions of contentful states. Clearly, if this were the relevant sense of illusion, content illusionism would be incoherent.[18]

We can, however, think of illusions in other ways. One such way is *behavioral*.[19] On this view, to be illuded is to be "pulled" to adopt a certain behavioral pattern by a stimulus. When, for example, one is in the grip of the Müller-Lyer illusion, one will behave in certain characteristic ways: they will say that one line is longer, indicate the line with inbound arrowheads as longer, they will show surprise after measuring the two lines, *etc*. And, in a way, the subject *can't but* behave in this way: it is compelled by the stimulus to do so.[20] Similarly, the content illusionist suggests, we can't but behave as if we were constantly shown contents being "carried around" by our mental states. For example, we cannot but make *sincere, but false* reports of mental contents, and we cannot but *sincerely* — that is, without explicit dissimulation — treat others (and ourselves) as if their mental states hosted contents. As I shall sometimes say for the sake of simplicity, we *fit the behavioral stereotype* of a persona believing that mental content is real (cf. Schwitzgebel, 2002; 2013), and

---

16 Both illusions are better experienced in video format: see https://www.youtube.com/watch?v=3t5qQB-FuJg for the Ames room and https://www.youtube.com/watch?v=cVepIZLepVc for the Ames window.

17 Including peoples from very different, "non-western" cultures and non-human animals (see Amir & Firestone, 2025).

18 This is a version of the "cognitive collapse" argument against eliminativism (see Baker, 1987; Sprevak, 2013). This is why the cognitive collapse (and related objections) are not addressed in §4.

19 Another way to think about the illusion will be introduced in §4.

20 Of course, this shouldn't be read *too* rigidly - a person who *knows* about the illusion will typically be able to suppress these behavioral patterns and exhibit different, "not-illusion-oriented" ones. But the pull of "illusory" behavioral patterns remains real - after all, one line *really seems* longer then the other!

thus we believe in a way such that the ascription of such a belief is (given our relevant folk psychological ascriptive practices) *false, but warranted*.[21]

An example to clarify. John wakes up tired and covered in sweat. As he gets to the kitchen, his partner looks at him and asks if everything is ok. John confirms that he just had a terrible night. He says something like:

> "I dreamed that my high school crush was investigating a murder, and I knew they wanted to frame you. So, I was driving you away to hide you, but I realized I wasn't really driving, I was cycling, and when I looked at the road again, I saw I was speeding on a sky-jumping trampoline. So, I tried to brake but I couldn't find the brake on the bike, and so I took this enormous jump and then woke up at 2:45 AM. And I was so upset and anxious I couldn't fall asleep until about 6:00."

What are John and his partner doing? John is presumably explaining why he looks awful, and he does so by means of a strange and incoherent narration. What matters is not if the narration is true or a true report of an experience (the illusionist of course claims that it isn't, but this isn't the focus now). What matters is how John and his partner *treat* it. John is presumably aware that the piece of narration is literally false (nothing of that really happened), and of its poor aesthetic qualities. Yet he is *compelled* to produce that narration — at least, if he's a collaborative speaker.[22] He cannot, for example, recount a different event of his life (it would not *explain* why he looks terrible) or recount a better, more scary story to justify his lack of sleep (e.g. a passage from Stephen King's novel *IT*). Indeed, if his partner noticed that John is doing something like that, they would assume that John is either abandoning the conversation; or trying to make a fool of them. Similarly, John's partner cannot, if they're being a collaborative speaker, remark on the lack of aesthetic qualities or coherence of John's narration. They're both naturally inclined (and compelled by social norms) to treat John as if during the night he was the spectator of the "mental film" he recounted — a mental film so realistic, engaging and anxiety inducing to have disrupted John's sleep pattern. They both fit the *behavioral stereotype* of people believing that John had such and such a dream, and should thus be ascribed the relevant belief (given our folk psychological practices). Most importantly, John and the partner *can't but* fit that stereotype. They can't but *behave as if* John witnessed some sort of contentful "inner movie" while sleeping, justifying the (for the illusionist false) ascription of the corresponding belief.[23]

Aren't John and his partner just pretending, as Toon (2023) recently argued? No. Even if John and his partner behave only as *if* John witnessed an "inner movie" while sleeping, their behavior is not pretense — not even if pretense were *just* a specific, enculturated behavioral pattern, as Toon (2023, p. 72-75) seems to hold. For, even if pretending to ϕ were just a behavior, the behavioral profile of pretending to ϕ would differ, no matter subtly, from *actually* ϕing — indeed, this is what essentially tells apart pretense behavior from non-pretense behavior (see Facchin & Rucinska, 2024). We can thus always contrast pretending to ϕ with really ϕing. But, in the case at hand, we have *no* such contrast. There is no "more serious" version of what John and his partner are doing. There is only a "less serious" version — *pretending* to have had a nightmare. In that case John's behavior will be different — if John only *pretends* to be tired, he won't, for ex-

---

[21] Won't the identification of behavioral stereotypes *presuppose* content? After all, behavioral stereotypes are typically identified in terms of that-clauses (cf. above: "the stereotype of the person *believing that content is real*"). Not really: we can think of behavioral stereotypes as "real patterns" (Dennett, 1991) to which our ascriptive practices "latch onto" - and crucially the existence of ascriptive practices does not *entail* the existence of the contents ascribed (just like the practice of exorcizing demons does not entail the existence of demons, no matter how successful that practice seems). Thanks to Erik Myin for this suggestion. See also Hutto & Satne (2015) for a similar idea defended in more detail.

[22] Why *that* narration rather than any other narration? I don't know. But that's not a problem unique to content illusionism. I know of no theory able to predict and explain why we dream *exactly* what we dream, rather than any other thing we could dream.

[23] What about any phenomenal state John may have undergone in his sleep? Is the content illusionist *forced* to deny their existence? Strictly speaking, the answer is negative. The content illusionist has only to deny that such states carry the content John reported (and any other mental content). This of course goes against *some* ways of conceiving of phenomenal states as essentially contentful (e.g. Mendelovici, 2018), but not against the arguably standard characterization of phenomenal states as non-contentful (Dennett, 1988; Chalmers, 1997).

ample, yawn naturally (though he may *pretend* to yawn, exaggerating his movements) and his movements won't be sluggish in the way the movements of a tired person are. He may even recount a piece of Stephen King's *IT* without triggering any bad reaction in his partner or stepping away from the relevant discourse commitments. So, John's ordinary case can be contrasted with a pretense case, but there is no *non-pretense* case with which to contrast *it*. Thus, it's no pretense.[24]

But how else can we make sense of "behaving as if", if not in terms of pretense? We can do it in terms of (systematic, intersubjective) *mistakes*. Consider this analogy. On a remote island, there is a remote tribe. The tribe's people live mostly on vegetables, because most of the meat they possess serves to stave off the Great Spirit of the Forest — which, if not appeased, would pillage the tribe's garden and cause a famine, or even eat the village people. To appease the Great Spirit, a delegate of the tribe offers it a daily sacrifice. At dusk, she goes near the Sacred Mound near the Forest where the Great Spirit resides. She leaves the corpse of a butchered animal there, and picks up the bones of the one offered the day before. There is, of course, no Great Spirit of the Forest, and the Sacred Mound is just a gigantic ant nest. The ants belong to an undiscovered, nocturnal and mostly carnivorous species. By offering this sacrifice, the tribe is basically ensuring the ants will be occupied throughout the night, and won't eat the vegetables from the village garden (or the village people). Still, villagers know nothing about these ants: it's taboo to get to the sacred mound at any other time, the taboo is strictly observed, and the ants are not yet active at dusk.

What are the tribe's people doing? Notice, the question is not what they *think* they are doing — the question is what they are *really* doing. We can exclude that they are *really* appeasing the Great Spirit, since there is no Great Spirit to appease. But notice that they are not just pretending to appease the Great Spirit. They are *not* pantomiming the prescribed ritual, but actually carrying it out. They are thus behaving as if there was a Great Spirit — thereby fitting the relevant behavioral stereotype - even if there is none, *simply because they are wrong*. Still, their behavior is pragmatically useful (it *really* keeps the carnivorous ants at bay) and able to sustain the belief that *there really* is a Great Spirit: who else would be devouring the sacrificed animals? In the same way, the illusionist suggests, by treating each other as having mental contents we are wrong, but in a useful and epistemically self-sustaining way.

Now, by endorsing the eliminative claim and the illusion claim, illusionism re-arranges our explanatory targets. Rather than explaining content (which does not exist), we should try to explain the *illusion* of content — which is real, and whose origins and grip are at present mysterious. The explanation of it could take a number of different forms, depending on the mechanism called upon to explain the illusion and its grip.[25] It's possible to envisage different forms of content illusionism based on the kind of mechanism invoked to explain the illusion of content. One option could be that of invoking a purely cerebral mechanism — the selective pressure to socially coordinate, for example, may have made us evolve a convenient system to communicate our relevant internal states so as to coordinate with each other. On the opposite side of the spectrum, the mechanism may be entirely "external" and "social", and dependent, for example, on our narrative practices. And between these two extremes, there is a vast sea of possibilities I cannot explore here. In §5, I will sketch a minimal mechanism generating the illusion which calls upon both neural and social (really sociocultural) resources. I do believe that such a "hybrid" view is the correct one. However, in line with the programmatic objectives of the paper, I won't defend it here. Indeed, in line with these objectives I wish to highlight that the "hybrid" view I will later sketch is not the only possible one, and that the space of possible mechanisms generating the illusion is quite wide.

Notice that the above means that, *strictly speaking*, content illusionism is not a theory of content (in any of the two senses seen above), but a theory of the *illusion of content.* Still, theories of content (in any of the two senses) and content illusionism have an important *explanandum* in common: our content-invoking introspective reports. For the illusionist, such reports are a constitutive part of the illusion, and should be explained by the relevant mechanism responsible

[24] I make this argument at length in a paper currently under review.

[25] If this sounds preposterous, keep in mind that philosophers as diverse as Clark (2015), Lee (2019) and Egan (2020) make an analogous claim about content *itself*. In comparison, the claim that a mechanism can explain the *illusion* of content seems fairly mild.

for it — indeed, such a mechanism would provide an error *theory* accounting for the production of these (false) reports. For the realist, such reports are genuine, truthful reports (at least in most cases), and their production should be explained by an appropriate introspective mechanism. Moreover, illusionism and theories of content (in any of the strict senses seen above) are clearly mutually exclusive, in the sense that the truth of content illusionism implies the falsity of any theory of content and *vice versa*. As such, it makes practical sense to continue to talk of content illusionism as an *alternative* to realism.

This concludes my general presentation of content illusionism. This is, admittedly, a very impressionistic, general description of content illusionism. It doesn't substantiate — or even suggest — any particular illusionistic doctrine. This is by design: recall that here I wish only to present the general *motif* of content illusionism, rather than any specific illusionist view. But why should one bother with illusionism?

3. Motivating content illusionism

Here, I wish to give some very general reasons as to why content illusionism is an attractive position, which we should prefer to the alternatives. But to do so, I must first clarify what the relevant alternatives are. To do so, I follow (Frankish, 2016), and take the relevant alternative to be only *realism*, either conservative or revolutionary.[26]

Conservative realism, like illusionism, adopts an austere ontology, wherein the only admissible entities and properties are the ones accepted by our *normal* (in the sense of Kuhn 1970) natural sciences. Unlike illusionism, however, the approach is committed to the existence of content — it is a form of realism, after all. Theories of content that identify contents with regular, mundane objects — such as James's (1911) account wherein contents are concrete particulars — are one straightforward example of conservative realism. Yet, conservative realism is nowadays more represented by a specific, "reductive" reading of naturalistic metasemantic theories, which tries to *reduce* or even *identify* content with the relevant naturalistic features mentioned in the relevant meta-semantic theories (see e.g., Lee, 2019; 2021 for a modern articulation of this view).[27]

Revolutionary realism retains the realistic commitments of conservative realism, but it doesn't try to squeeze content in the mold of normal natural sciences. On the contrary, it argues that content cannot be so squeezed. To account for content, one must either operate *outside* of the framework offered by natural sciences; or revolutionize that framework introducing some novel ontological ingredients. Phenomenologists who doubt that the mind can be naturalized are a valid representative of the first tendency (Husserl, 1970; Gallagher, 2017), whereas semantic primitivists taking semantic features as a basic ingredient of reality (Burge 2010) are a good example of the second tendency.

Now, why should one prefer content illusionism to these forms of realism?

---

[26] What about all non-realistic views of content? My answer is that, when properly analyzed, these are either not alternatives to illusionism, or *extremely* unattractive ones. As seen above, fictionalism in its various incarnations is a theory about the semantics and/or pragmatics of our content-involving talk. It's not a theory of how that talk is produced and why we believe in it (i.e., behave as if mental contents are real). So, illusionism and fictionalism are not alternatives (though they can still clash on the semantics and/or pragmatics of the content-involving talk). Similar considerations also hold for *instrumentalism*. Eliminativism about content is a truncated and unconvincing form of illusionism: it lacks the illusion claim, and so it misses an obvious *explanandum*, namely the illusion of content. Since I cannot see any other non-realistic alternative, I conclude that realism, in its various forms, is the only palatable alternative to illusionism.

[27] Notice that in this case the conservative realist would be *weakly* illusionist, in that they would claim that content exists, but are really different from what they appear (they just are certain relevant naturalistic ingredients called upon by theories of content). For the relationship between reductionism and "weak" forms of illusionism, see Frankish (2016).

3.1. Motivation #1: Parsimony

A first reason to prefer content illusionism is that (ceteris paribus[28]) it's more *ontologically parsimonious* than the alternatives — it denies the existence of contents, whereas the alternatives are committed to it. In the case of conservative realism, the parsimony is only *quantitative*: both views agree on the *kinds* of things admissible in our ontology, but disagree on their *number*: according to the illusionist, their number does not include contents. According to the conservative realist, it does. In the case of revolutionary realism, the parsimony is also *qualitative*: illusionist and revolutionary realist disagree also on the *kinds* of things our ontology admits. For the revolutionary realist is committed not only to the kinds that figure in our normal sciences, but also to some other kind pertaining to our revolutionary sciences. That's why it's *revolutionary* realism, after all.

What about *mechanistic* (and, relatedly, *explanatorily*) simplicity? One may worry that the mechanism generating the illusion of content *must* be pretty complicated, generating pretty complex explanations. This would compare them unfavorably to the mechanisms — and the explanations — at the realist's avail. After all, the realist could call upon a simple detection mechanism, which would enable a straightforward explanation of why we *behave as if* our minds were populated by contents (namely, we do so *because we detect them*). I'll deal with this issue in §5, showing that the mechanisms responsible for the illusion of content can be remarkably simple, and that a *satisfactory* realistic explanation of our reports and behavior must be *at least as complex* as the illusionist's one.

Now, if *parsimony* is a reason to prefer a theory over the alternatives, we have a reason to prefer illusionism over realism. Yet, parsimony offers such a reason only *everything else being equal*, and the realist would claim that everything else is *not* equal. I agree — I think it's even better for illusionism; which brings me to the second reason to prefer it.

3.2. Motivation #2: Better fit with the mind sciences

Representations and contents were the central explanatory tools of "classical" cognitive science.[29] But, pretty unsurprisingly, the landscape of the mind sciences has changed from the '50s — and at present, it's far from clear that representations and content play any role within the sciences of the mind (cf. Ramsey, 2007; Favela & Machery, 2024). Sure, the term "representation" is used, thereby invoking contents. Yet, empirical surveys suggest that the term "representation" is currently used to refer to any relevant internal state (and "content" anything that caused the tokening of the state), regardless of whether the state satisfies any definition of representation. Indeed, often the usage of "representation" outstrips the definitions of the term that practitioners of the mind science accept (Villaroya, 2017; Favela & Machery, 2023).

Examples of this over-liberal usage of representations are not hard to find. For example, Kaplan & Zimmer (2020) speak of brain-wide representations of current behavior — in spite of the fact that representations are typically taken to be causally decouplable from what they represent (Haugeland, 1991; Baker *et al.*, 2022), unlike brains from current behaviors.[30] Brette (2019) talks of cell assemblies *representing* models, where he clearly means *implementing* models which (in turn) represent external objects (see also Jakobson, 2013 for a sample of other such cases).[31] In the "predictive processing" literature it is commonplace to treat brains, and even entire organisms, as models of their niches (e.g., Friston,

---

[28] That is, keeping constant other, independent ontological commitments (e.g. to mathematical entities).

[29] Notice, however, that this is something even very "classical" cognitivists deny (Stich, 1983; Chomsky, 1995).

[30] Aren't brain-in-a-vat scenarios proof that brains *can* in fact be causally decoupled from current behaviors? No, they aren't. At best, they prove that the brain can be *very wrong* about how reality is. As many philosophers have already argued at length (Hurley, 1998, Ch.8; Thompson & Cosmelli, 2011; Hohwy, 2017), even an envatted brain must be immersed in the causal regularities provided by an active (i.e., behaving) body. Indeed, the vat is precisely providing *that coupling in* the absence of an actual behaving body.

[31] The difference matters: the expression "MOSAIC" *represents* a pretty famous model of motor control (see Haruno *et al.*, 2001). But "MOSAIC" does not compute a thing. For that, you need the model *implemented*, not *represented*.

2011) — where "modelling" boils down just to a non-trivial form of matching (e.g., Conant & Ashby, 1970; Clark, 2017). In their *Principles of Neural Sciences*, Kandell *et al.* (2012, p. 370) talk of the job of neural sciences being that of discovering how the brain represents mental states, but they clearly mean something like *constitute or realize* — and examples could be easily multiplied.

Now, perhaps this unruly usage of "representation" plays some relevant role in current cognitive science (e.g., it may play some communicative function, see Favela & Machery, 2024). But, as philosophers of the mind sciences have repeatedly argued, it does currently play no *referential* role — at least, if the term "representation" is supposed to refer to *contentful* representations. For, in many — indeed, *most* — sectors of current cognitive scientific theorizing, the word "representation" (and its cognates) fails to refer to any *genuinely representational* posit or observable.[32] This point has been convincingly made for components of cognitive mechanism in general (Kohar, 2023), various neurological structures (Anderson & Champion, 2022; Facchin, 2023; Krakauer & Ramsey, 2026), different kinds of — currently extremely popular — "predictive" models of the mind (Orlandi, 2014; 2016; 2018; Facchin, 2021a, b) and, most importantly, artificial neural networks — from relatively shallow classifiers to deep auto-encoders (Ramsey, 2007; Davies Barton *et al.*, 2024). This covers the majority of cognitive models and theories.

This doesn't (of course) mean that the word "representation" fails to refer *altogether* — it is used to refer to many different *non-representational* posits or observables. It means, however, that the word is *misused* so that its usage is no longer diagnostic of any genuine representational commitment. Quite literally *any* other word could do the same job — even ridiculous ones such as "demon". Yet, in such a case, "demon" would fail to refer to demons, and so presumably no one would rely on such a usage of "demon" to argue that the cognitive sciences are ontologically committed to demons, or in favor of a *demonic theory of mind*. Indeed, no one invoked the *pandemonium model* (Selfridge, 1958) to defend these views, even if it *explicitly* used the term "demon" and quantified over it!

What about *information-carrying states*? Surely at times the word "representation" refers to those, and such states have some sort of affinity with representations. Aren't they enough to justify the claim that cognitive science is still committed to representations and content? No, they aren't. Even leaving aside the — quite substantial — problems concerning whether cognitive science is *really* committed physical states "carrying" information (see Hutto & Myin, 2013; Ramsey, 2020; Zahnoun, 2025; Sganzerla *et al.*, 2025) and whether such a commitment is actually beneficial or warranted (Nizami, 2019; Brette, 2019, 2026), cognitive science remains committed only to a Shannon-inspired[33], *correlational* notion of information; and such a notion of information has no essential tie to representations and contents. At the most basic level, such a notion of information captures only *correlations* between states, such that observing that one state is the case makes the occurrence of the correlated state more likely. Formalizing a bit, a state S carries information about a state Z just in case $p(Z|S) > p(Z)$ (see Shea 2018, p. 76, see also Dretske, 1998; Hutto & Myin, 2013; Neander, 2017; Krakauer & Ramsey, 2026). Of course, such a purely correlational notion of information is not semantic, representational, and contentful in any way.[34] The fact that the street is wet raises the probability that it has just rained, but streets are in no sense contentful representations. The fact that the tide is high raises the probability that the moon is in a certain position in respect to the body of water, but bodies of water do not represent the moon. Thus, the fact that modern cognitive science is committed to information-bearing states neither entails nor supports the view that it is committed to representations.

---

[32] Notice that such claim holds for *most* sectors of current cognitive science. As a reviewer correctly noticed, it may be that in some sectors of cognitive science (e.g. formal semantics) the notion of content cannot be eliminated. It is worth noticing, however, that even in this case there are theories that eliminate meanings and contents in the relevant sense. Distributional semantics (Harris 1954) could be an example.

[33] I say "Shannon-inspired" because Shannon's original formalism has aspects that are often "glossed over" by neuroscientific practice and related philosophical discussion. For example, little attention is given to channel conditions (for more on this point, see Martínez, 2025).

[34] This is also why it makes a good fit for naturalistic theories of content - that is, theories that try to ground content in natural (non semantic, non mental) factors.

All in all, it appears that the current mind sciences are largely doing fine without representations. Hence, we should not be committed to them — which favors illusionism over realism. But the fit between illusionism and the current mind sciences goes even deeper.

3.3. Motivation #3: Better explanations of behavior

The fact that the mind sciences have "silently" turned non-representational suggests that eliminating content won't hinder their explanatory power. Indeed, as I shall now illustrate, the opposite is true: *explicitly* eliminating content and representation puts us *more* in contact with how the mind works, and afford us *better* explanations of behavior.

For example, we intuitively tend to think of visual perception as a sort of "mental photo" of our surroundings. Iconic, depiction-like conceptions of perception are still alive in some sectors of (classical) cognitive psychology.[35] But we now have enough evidence to be sure that perception *cannot* consist in the creation of such a mental photo — and thus that research projects aiming to explain perception in these terms are misguided. For, our perceptual system is nothing like a camera for mental photos — it's rather a system made to retrieve the relevant environmental information just in time to guide our online action (see Churchland *et al.*, 1994; O'Regan, 1992; 2011; Noe, 2004).[36]

Wouldn't recent work on mental imagery (e.g., Nanay, 2023) rescue the "mental photo" view of perception, as a reviewer suggests? I don't think so. Indeed, consider *the predictive processing* account of perception, which I take to be a prominent neuroscientific account of the continuity between mental imagery and perception (cf. Clark, 2013; 2014; 2016; see also Van Leeuwen, 2025; Nanay, 2025 for further discussion). Without getting into too much detail on the account (I did so in, Facchin, 2022) the idea is that the machinery of perception does not passively wait for perceptual inputs, but rather tries to *guess* what the perceptual input is. Crucially, such guesses take the form of the *endogenous generation of the incoming input*. When it comes to perception, this means that the guessed and the actual input can be easily compared, generating an error signal that can be used to "revise" the guesses and keep it in contact with reality. The endogenously generated input, however, can also be used "offline", and kept free from the control of the error signal, generating various kinds of imaging phenomena (Clark, 2012; 2016). Perception, then, is just a form of reality-controlled imagery, and consists in the production of a reality-tested "mental image" of the actual visual scene — or so, at least, predictive processing suggests.

Notice that predictive processing is taken to be a strongly representationalist theory, indeed it may be "as representational as cognitive scientific theories get" (Gładziejewski, 2016, p. 562), to use the words of a prime defender of the theory's representational reading. Appealing to it, then, should put significant pressure on my own view.

Yet, the opposite is true. Not only does predictive processing *support* — and indeed it *mandates* — an active account of vision (see Friston *et al.*, 2012), wherein exploratory movements (such as, for example, saccadic eye movements) are constantly performed to continuously provide the relevant error signal and keep perception anchored to reality, but the theory also *presupposes* mastery of sensorimotor contingency (Seth, 2014; Facchin, 2021a; 2022). The reason is trivial: to economically generate the right inputs, the brain must be able at least to anticipate how our movements will impact them. And such a mastery need not have any representational basis, it can simply be realized by direct neuronal connec-

---

[35] See for example Block (2023).

[36] Isn't "retrieving the relevant information" something like operating contentful structures? No, at least not given the sense talk of information has for philosophers and cognitive scientists. Information is (as recalled above) understood as just a correlation between types of states - but surely correlations are not in and by itself contentful. Indeed, this is precisely what makes informational theories of content (e.g. Dretske, 1988; Usher, 2001; Heemskerk, 2024) *naturalistic* theories. The correlational nature of information means that talk of "retrieving information" is metaphorical (Hutto & Myin, 2014; Ramsey, 2020). To speak literally, we should - as Orlandi (2014; 2016; 2018) correctly does - talk about the visual system re-organizing itself in response to environmental perturbations. But that would make the paper a bit too cumbersome to read.

tions allowing motor neurons to influence the neurons endogenously generating the incoming input.[37] Additionally, as it has been argued time and time again (Orlandi, 2014; 2016; 2018; Facchin, 2021a,b, 2022; Downey, 2016; 2018; Ramstead *et al.*, 2020, Kirchhoff & Robertson, 2019; Robertson & Kirchhoff, 2019; Van Es & Myin, 2020) the machinery of predictive processing is, upon closer analysis, markedly representation-free. Indeed, predictive processing has been one important wake up call for philosophers, who noticed — and criticized[38] — its abuse of representational language (e.g., Friston, 2011).

So, looking at predictive processing — to repeat, what is commonly taken to be a *strong* representationalist theory — and the way it depicts the role of mental imagery in perception does not support a "mental photo" view of perception. On the contrary, it ends up supporting an active view of perception based on our sensorimotor mastery. Indeed, in the predictive processing view, it is precisely because perception is *active* (and thus actively generating the relevant feedback signal) that perception remains controlled and "in touch" with reality.

The same holds for other mental states that seem to heavily revolve around the manipulation of content. For example, insight, imagination and problem solving are not the result of the mingling of mental contents, but rather a product of an agent's intelligent and controlled engagements with a number of epistemically supportive environmental props (Gedenryd, 1998; Van Leuween *et al.*, 1999; Van Leuween, 2011; Vallée-Tourangeau *et al.*, 2016; Carvalho & Nolfi, 2016; Pezzulo & Nolfi, 2019; Bocanegra *et al.*, 2019).[39] Getting away from content, then, bolsters our theories in the mind sciences, affording us a better grip on our very mental life.[40]

But what about *folk psychological* explanations? Surely these are often content-based explanations. Won't eliminating mental content deprive us of them? No, it will not. As said above, content illusionism entails only a very limited form of discourse eliminativism, limited to our *scientific* explanations.[41] Of course, if one were to think (a la Fodor, 1987) that folk psychology offers a literal, representational causal-explanatory theory of how our mind works one would have to eliminate contents from folk psychology too. But I reckon that few, even amongst realists, hold such a view of folk psychology today. After all, we do know that such a view makes hefty, almost certainly false, assumptions about our cognitive architectures (Churchland, 1992; Ramsey *et al.*, 2013; Schwitzgebel, 2026), and that we can use folk psychology to causally explain our behavior even without making such hefty assumptions (e.g., Schwitzgebel, 2013; Langland-Hassan, 2022). So, I take it that folk psychology remains basically untouched by the elimination of content — which happens painlessly, without any significant loss of explanatory power.

3.4. Motivation #4: Getting a new, easier explanatory target

Talking of explanatory power, it's worth noting that endorsing content illusionism allows one to trade away a metaphysically hard explanatory target (mental content) for a metaphysically easy one (the illusion of content).

Of course, the mechanism explaining the illusion can be very complicated, and will likely mention various features of our nervous systems as well as the way they interact among them and with our environment (including, quite possibly, our *social* environment). Still, there is nothing in principle *metaphysically* complex in explaining our behaviors recurring

---

37 See Bovet (2007) for a simple model, Leinweber *et al.* (2017) for analogous neurological findings, and Facchin (2021a) for a philosophically oriented discussion of them.

38 For such criticism, see for example Van Es (2021), Baltieri *et al.* (2020). Notice also that such a criticism was shared by philosophers who defend a representationalist position - such as Clark (2017).

39 It's worth noticing that such a reliance on external props to solve complex cognitive tasks seems a quite basic feature of minds - it can be seen even in fairly "lowly" animal minds, like the minds of wasps and arachnids, which presumably don't trade in any content whatsoever (see Facchin & Leonetti 2024, for examples).

40 Anderson (2014) makes the same point concerning the explanation of our brain and its organization.

41 Assuming that some form of scientific realism is correct. See Ramsey (2020) for an argument to this effect tailored to representations.

to neural and environmental features — cyberneticists do so routinely (Braitenberg, 1986; Brooks, 1999; Nolfi, 2002; Carvalho & Nolfi, 2016). The same cannot be said for conservative and revolutionary realism. Revolutionary realism acknowledges that accounting for content from a naturalistic perspective is *metaphysically hard*, calling for a revolutionary revision of our naturalistic worldview. Conservative realism disagrees, and tries to account for contents in terms of familiar ingredients taken from our normal natural sciences. Such attempts, however, have thus far only produced failures, and the whole program seems to be stalling, if not degenerating (for this quite grim verdict, see Kriegel, 2011; Egan, 2020; Lycan, 2022). This strongly suggests that the problem of content is metaphysically hard — even if perhaps not as hard as the hard problem of consciousness (cf. Hutto & Myin, 2013; Mendelovici & Bourget, 2023). Thus, abandoning realisms in favor of content illusionism is an *advantageous* trade, one that we should be willing to make. Hence, we should prefer content illusionism over both forms of realism.

3.5. Motivation #5: Avoiding the problems plaguing realism

Relatedly, content illusionism avoids certain prominent pitfalls and problems that plague realistic approaches. For example, conservative realism is plagued by the *determinacy* problem: no matter how we try to refine our naturalistic metasemantic theories, such theories fail to ascribe determinate contents to the relevant mental states.[42] Another problem is that these theories overgeneralize: the relevant naturalistic "ingredients" that, according to these theories, ground contents are almost everywhere — and so too should be content, if these theories were correct. But this result seems wrong, at least insofar it ascribes mental contents to presumably mindless systems like smoke detectors and rifles (the examples are from Ramsey, 2007; Orlandi, 2014).[43] As for revolutionary realism, it typically insists on the role of *introspection* in the study and characterization of mental contents. But this leaves them open to a potent methodological objection — introspection is an extremely unreliable source of information about our mental life (e.g., Schwitzgebel, 2008). Indeed, even "rigorously trained" introspectors end up *disagreeing* on the specific structures of their mental contents (as attested by the failure of the introspectionist program in early scientific psychology). Moreover, in many (arguably most) cases our introspective reports are not the result of *introspection*, but rather confabulation (see Chater 2018, for a number of examples). And notice that we are unable to discern between introspection and confabulation — indeed, it's often shocking to discover that many of our introspective reports are groundless, and invented "on the spot". So revolutionary realism seems methodologically flawed, and thus an unreliable guide to our mental reality.

Content illusionism, of course, avoids these problems. By denying the existence of content, it avoids the problems of indeterminacy and overgeneralization that plague conservative realism. And, since it claims that there are no contents, it has an easy time to account for why introspective reports appear to be systematically confabulated on the spot — they are because there is nothing "inside" to report![44]

3.6. Motivation #6: Content illusionism accounts for the features of content better than any form of realism

A last reason to prefer content illusionism is that it accounts for certain salient features of content better than the realistic alternatives.[45] What are these features? I have already mentioned two — systematicity and intersubjective

---

[42] This problems changes in its articulation depending on the theory of content under scrutiny, see Artiga & Sebastián (2018) for informational accounts, Cummins (1989) for causal accounts, Fodor (1990) for teleological theories, Segundo-Ortin & Hutto (2021) for accounts based on abstract similarities, and Egan (2020) for a review.

[43] Importantly, even some revolutionary realistic theories of content suffer from the same problem (Nes, 2008; Raimondi, 2021).

[44] Well, what about the (seemingly) non-confabulatory cases of introspection? I have an account of these too. The important point to notice is that they are not reports of independently existing mental contents, but ways in which an agent makes sense of themself and his situation (Facchin et al., 2025).

[45] I make this point in full detail in Facchin (2025).

concordance (§2). Another important feature is that illusions are persistent in the face of opposing evidence. No matter how many times we measure the two lines of the Müller-Lyer illusion, they will always look different (Fodor, 1984).

Now, I have argued that content is systematic and intersubjectively concordant above, and I won't rehearse the point here. And, to be persuaded that content is persistent, is sufficient to notice that content eliminativists like Rosenberg or Hutto & Myin never — despite their sincere conviction of the inexistence of content — reported mental content vanishing. To the contrary, current content eliminativists clearly voice the persistence of content, by wondering how it is possible, if the mind really is a non-representational system, that something nevertheless seems to stand in for environmentally absent objects and features (Downey, 2024). So, content quite trivially has these features. But these features are only weak indicators of illusion. In fact, whilst they discriminate illusions from errors and hallucinations (which are unsystematic, intersubjectively discordant and, in the case of errors, non-persistent), they do not discriminate between illusions and reality. After all, the appearance of real objects, properties and events is subjectively concordant, systematic and persistent too.

Illusions, however, have other features too, which real objects typically lack. One is that they are *bound to a single mode of access*, and trying to access them in other ways, or to sensorily explore or manipulate them, typically reveals their illusory nature. Spectators of illusions, has been noted, are extremely passive (Dennett, 1991b; Hohwy, 2013), as exploring illusions "spoils the trick". Many illusions discovered by Ames, for example, force the spectator to look with just one eye through a peeping hole, minimizing the spectator's ability to explore the scene (Ittelson, 1952), and the appearance of diversity in the Müller-Lyer illusion can be dispelled by connecting the edges of segments via parallel lines. Illusions also show something that fails to fit in *the causal order of the world*. By looking at the Müller-Lyer lines, one could have the impression that the arrowheads caused the lines to take different lengths — something we know can't be the case. Or, by looking in the Ames room, one has the impression that moving causes people to grow or shrink in size, something that we know is not (and cannot) be the case.

Notice, importantly, that these features are *highly diagnostic* of illusions. Natural objects are typically accessible from multiple channels and freely explorable. Moreover, they are tightly knit in the causal order of the universe. Compare: if I were to show you my Ferrari, but force you to see it only from a *single* point of view and insisting that you cannot see it from any other angle, or perceptually explore it, you would likely think I really have no Ferrari, but just a Ferrari-shaped cardboard façade with which I'm trying to fool you. And this impression would likely be reinforced if my Ferrari fails to "fit" in the causal order of the world as a Ferrari would (e.g., if it's impalpable, or cannot be damaged).

Content, however, displays these two features. For one thing, the only *direct access* we have to our mental contents is by means of introspection. Mental states, after all, are *private*. Of course, we can *indirectly* access them and their contents (for example, through inference from behavioral data[46]) — but the same holds for illusions: one can indirectly access the Ames room illusion by reading about it. But to access it directly – that is, for the illusion to be *given* – one can only look through the peeping-hole. And the same goes for content — for it to be given, it must be introspected. Indeed, attempts to access contents from a different perspective — for example, by using sophisticated techniques of feature visualization, which should make transparent to us the contents represented by neurocomputational models — do not uncover things that seem even remotely mental contents. Consider, for example, the features visualized in **figure 1**:

[46] More recently, "mind-reading" techniques have allowed to ascribe contents by inference from *neural* data - but such technologies remain inferential through and through, requiring sophisticated probabilistic and information theoretic assumption in order to work (e.g. Kriegeskorte & Douglas, 2019, pp. 169-170).

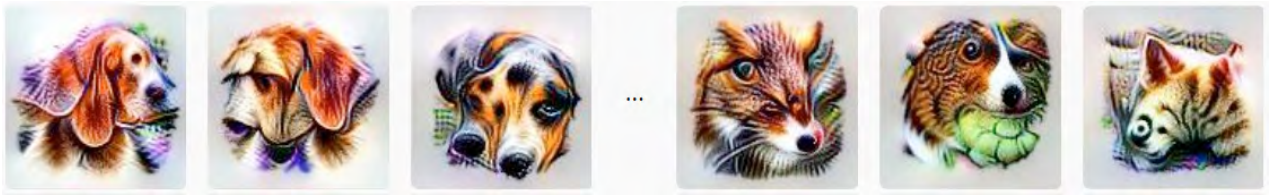


Figure 1

Several "mental contents" of a neurocomputational model, respectively labeled as "Tiger Cat" (right) and "Labrador Retriever" (left). A visual inspection of such things seems sufficient to confirm that states presenting these things would not in fact be presenting any tiger cat or Labrador retriever to any subject. (Reprinted from Olah et al., 2018, reused under CC-BY)

The features visualized are *labelled* as "beagles" and "tigers". But they are *nothing like* the contents of our mental states representing tigers and beagles. If when you try to think of a Beagle you were to end up thinking *these things*, you'd likely see a neurologist. And this is no isolated case — generally the features "visualized" by such techniques don't even remotely resemble our mental contents.[47]

Content also exhibits apparent causality.[48] We "see" contents causing our behaviors all the time: my desire for another beer makes me walk towards the bar, and my belief that drinking and driving is unsafe makes me leave my car keys home. But we also know that content itself can hardly be the cause of my movements and behaviors. All the causal work seems to be made by its "vehicle" — the physical (neural) state carrying content around. It is that state, the particular firing of some of my neurons, that makes me order another beer, or that makes me leave the car keys on the table, back home. Subtract these physical states from the vignette, and you will subtract my behavior. Subtract the contents, and I will move just the same. You can, for example, imagine that I have recently died (just before hitting the bar), and that a mad scientist is using some electrodes in my brain to make me move, Galvani-style.

Notice that the causal inertia of content is an upshot of the fact that (most) philosophers endorse some form of externalism about content: content, they hold, is grounded in factors that are *external* to the subject and thus extrinsic to the subject's state. This leaves the door open to atom-by-atom duplicates of a subject that are nevertheless embedded in an environment wherein the relevant external factors do not obtain and are thus contentless. Being atom-by-atom duplicates of a normal subject, these duplicates *will duplicate any causal relation the normal subject has with the environment*: the subject and the duplicate will behave identically, react in identical ways to identical stimuli, and even produce identical reports.[49] Yet, the duplicate *simply has no content*, strongly suggesting that content is *not at all causally involved* in the production of our behavior. Davidson's (2016) swampman scenario offers the most famous example of such duplicates — and the one that most often lets content realists concede the point (e.g., Millikan, 1993; Dretske, 1988; Shea, 2018). But similar thought experiments could be replicated for *other* theories of content, leading to a similar conclusion.[50]

Realists, whether conservative or revolutionary, have a hard time accounting for these features of content. Conservative realists take contents to be regular parts of the natural world. But such parts are typically accessible from a number of different perspectives, and causally efficacious. Why shouldn't content be? Conservative realists, as far as I can see, have

---

47 For another example, see https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

48 In the efficient sense of causality. Content may still be causal in some other sense.

49 Such a duplicate would thus be, in a way, an intentional zombie - the analog of a philosophical zombie in regards to mental content (rather than phenomenal consciousness).

50 If at all needed: many theories of content nowadays are some sort of variation on teleosemantic themes (e.g. Neander, 2017; Shea, 2018; Heemskerk, 2024), and are thus vulnerable to the Swampman counterexample.

never offered a successful answer to this question. Revolutionary realists could, in principle, sidestep these problems by positing novel, special objects (contents) that are accessible only through introspection and that are causally insulated from the rest of the world. But I take that most revolutionary realists would like not to. After all, semantic primitivists such as Burge (2010) give contents a role in the mind sciences, and thus in the causal explanation of our behavior. Presumably, such a role would contrast with content lacking any causal powers. Similarly, phenomenologists seem to take our most basic experiences to be grounded in our bodily interactions with the world (e.g., Husserl, 1970; Heidegger, 1926/1962; Merleau-Ponty, 1963), which suggests that our mental contents are causally influenced by bodily goings on — a view that, again, clashes with the causal insulation of content. In contrast, content illusionists have a straightforward time in accounting for them. After all, the illusionist claims that contents *are illusions*. The fact that continents have the prototypical feature of illusions, then, should be quite expected.

3.7. Summing up

To recap, I have offered six general reasons to prefer content illusionism over realism. Content illusionism (1) is more parsimonious than realism, it (2) better fits with the mind sciences, (3) affords better explanations of our behavior, (4) allows us to trade away a metaphysically hard explanatory target for an easier one, (5) avoids the puzzles and pitfalls of realism, and (6) better accounts for certain relevant features of content. Taken individually, no such reason is powerful enough to motivate the adoption of content illusionism. Nor are they intended to. Even taking them together, I don't want to suggest that these six reasons offer a *knockdown argument* against realism — they clearly don't. Still, together, these six considerations indicate that content illusionism has significant attraction over realism, and should thus be preferred to it.

4. Defending content illusionism

I shall now try to defend content illusionism from some of the most obvious objections to it. I will here consider three objections.

4.1. Objection #1: Content illusionism denies the data

The first objection contends that content illusionism *denies the data*. According to the objector, mental contents are simply and undeniably given to our awareness. Any denial of this is extremely implausible, and methodologically unwise. Surely our first-personal, introspective knowledge of our mind is an important — if not *primary* — source of data concerning our mind.[51]

The illusionist's reply is threefold: (i) it is false that content illusionism denies the data; (ii) even if contents were given, the objection would fail; (iii) no methodologically pernicious move is being made. Let me expand.

(i) Content illusionism does not deny any *data*, if "data" means what it usually means: the result of third-personal, public and repeatable observation or measurement of *something*. In the case at hand, the data are introspective reports concerning mental contents, and seemingly content-involving behaviors (see Dennett, 1991b), whose existence the illusionist does not deny.[52] To the contrary, they are central to the illusionist — for they are a constitutive part of the il-

[51] This objection is inspired by an analogous one leveled against illusionism about phenomenal consciousness (e.g. Searle, 1992, 1995; Strawson, 2004).

[52] What if by "data" one means something else – like a private apprehension of something mental (cf. Searle 1995)? Well, these are not data in the common, *unproblematic* sense of the term. On the contrary, these "data" are highly contentious entities whose existence is *explicitly*

lusion of content, which is the illusionist explanatory target. Of course, the illusionist denies *the truthfulness* of these reports, as it entails the existence of content (which the illusionist negates). But to deny this is not to deny the data. Compare: no one would claim that the humanist Lorenzo Valla denied the data when he exposed Constantine's Donation as a forgery. He only accounted for the data in a different way.

(ii) Allow — *only* for the sake of discussion — essentially private data to be relevant in the discussion. What does it mean that these data are *given*? More precisely, should "being given" be read as a success (factive) term? This is a yes/no question that admits only two answers — none of which puts the desired pressure on illusionism.

If the answer is no, then "being given" works roughly like "being believed": something can be given, or be believed, without being real. But then the fact that contents are given (in this sense) poses no challenge to illusionism. Indeed, the illusionist *could* reasonably say that the illusion of content consists in contents seemingly being given to awareness without these contents being real!

If the answer is yes, then "being given" works roughly like "being known": something can be given, or be known, only if that thing is real. That's why from the *knowledge* of p we can infer that p is the case. So, knowing p *presupposes* that p. Hence, were content given in this factive sense, its existence would be presupposed. But to presuppose the existence of content just is to presuppose the falsity of content illusionism. Hence, were "being given" factive, the objection would be circular.[53]

(iii) It is also false that our "first-person" knowledge of our mind is an important, or even *primary*, source of data in the study of the mind. Psychological enterprises relying on such knowledge historically *failed*, and the study of the mind has progressed mainly through *third-personal*, indirect measures of mental activity. This holds even for the subjective phenomenon *par excellence*, namely phenomenal consciousness (see Seth, 2021).

So, content illusionism does not deny the data – it only accounts for them in a different way. At this point, it is easy to imagine that the realist would still argue that a problem persists: the realist can immediately (and seemingly unproblematically) call for *an abductive inference to the best explanation* to infer the existence of content — what else could explain the production of the relevant report? Surely the illusionist must concoct some sort of complex, highly *ad-hoc* account to explain it. And such a complex explanation can't be better than the simpler realist one.[54] But it can, as I will argue in §5.

4.2. Objection #2: The "no appearance/reality distinction" objection

The second objection argues that content illusionism is wrong because, just like phenomenal consciousness, content leaves no room for the appearance/reality distinction. If it seems to me that I am thinking of a green apple, then I'm really thinking of a green apple. Of course, I may be wrong — the apple I was thinking about turned out to be the tiniest of watermelons. But even in this case, my thought would be wrong only because its content was *apple* rather than *watermelon*. So, if it seems to one that one is thinking about x, one is really, *infallibly and eo ipso* thinking about x. There simply is no room for error.[55]

I must confess I'm really baffled by this line of reasoning. It's quite easy to imagine a case of error in regards to content.[56] Imagine, for example, a super-neurosurgeon, able to mess with a person's introspective mechanism in the follow-

---

denied by illusionism. The objection, then, would be a re-statement of the eliminative claim (§2.2), and so it would not be an *objection* at all – at least, not until an independent argument for the existence of these "data" is provided.

53 Identical considerations, it seems to me, hold if "givenness" is substituted with "acquaintance" (e.g. Balog, 2023) - or any other "direct, unmediated" epistemic relation.

54 Thanks to a reviewer for pressing this point.

55 For an analog point concerning phenomenal consciousness, see for example Kripke (1980).

56 And phenomenal consciousness: see Dennett's (1988) "alternative neurosurgery" scenario, from which I'm loosely drawing here.

ing way: whenever the person is in a state whose content is related to apples, the mechanism introspects a state whose content is related to oranges, and vice versa. So, when a person who underwent this surgery will think about *apfelstrudel*, she will token a state whose content is a particular apple pie typical of the alpine region — but this person will introspect (and report) that they are thinking about an orange pie. And when this person sees a glass of orange juice, the person will introspect seeing a glass of *apple* juice — and so forth. Of course, such fine-grained interventions are (currently) not real — but they obviously are physically possible. It's thus quite easy to imagine a case where a person introspects one content when they actually have another. There is a space for the appearance/reality distinction.

The realist may reply that the illusionist is not entitled to use such a scenario to defend their position, as the scenario seems to presuppose the existence of certain contents, including the contents of introspective states. The point is well taken, but not fatal to the illusionist, for at least two reasons. First, when articulating the point, the illusionist can simply grant the existence of content for the sake of discussion – and then proceed to show that the realist is not entitled to the objection, given that they must allow for an appearance/reality distinction given their own theoretical commitments. Secondly, it's relatively easy to modify the scenario sketched above in a way that does not presuppose the existence of any content. Indeed, it is sufficient to change what the introspective mechanism monitors (and does) to make the scenario compatible with content illusionism. Thus, imagine that rather than detecting contentful mental states, the introspective mechanism is sensible to the — surely enormously complex — *dispositions to behave* towards apples and oranges.[57] Imagine further that the introspective mechanism does not produce contentful states, but "reroutes" the agent behavior; that is, when the introspective mechanism tracks the activation of apple-related behavioral dispositions (including the disposition to report apple-related contents) it deactivates those and activates orange-related dispositions (and vice versa). Thus, the mechanism sketched above can easily be made compatible with content illusionism, and so there's nothing *problematic* in the illusionist appealing to it.

4.3. Objection #3: Content illusionism entails semantic nihilism

The third objection argues that content illusionism cannot be *true*, because, if the world were like content illusionism describes there would be no content, and thus nothing that can be true or false. After all, it seems that the meaning — and thus the truth — of what I am writing hinges upon our (shared) interpretation of these signs; that is, it seems to hinge on (presumably contentful) mental acts of interpretation. Eliminate those, and eliminate the meaning from the writing — and thus, the potential truth or falsity — of what I am writing.

The content illusionist can react to this challenge by noticing that it hinges on a common, but fairly substantial, presupposition — namely, that mental content is (directly or indirectly) the source of all content, meaning, and every semantic evaluable property (e.g., Grice, 1989; Aizawa & Adams, 2005; Mendelovici, 2026).

Yet, this assumption should not be granted — at least, not without argument. On the one hand, the assumption presupposes a specific, typically conservative realist (see references above) view of mental content as the ultimate ground for all content. But, as noticed above, conservative realism is not in good health, and so it cannot offer us a reason to espouse this assumption. Nor does the assumption seem particularly "natural" or "obvious" to make — the question of whether certain (non-mental) states have intrinsic, mind-independent content is fairly old and admits a variety of solutions (e.g., Plato, 1997). And, as weird as it may sound, some terms almost universally seem to be "natural names" for certain things, as attested by the kiki/bouba effect (see Cwiek *et al.*, 2022). But even without resurrecting an (admittedly quite cryptic) Cratylean view of "natural significance", the content illusionist can allow for semantic properties

[57] Wouldn't these dispositions be so multifaceted and unruly to make necessary for the introspection mechanism to transduce them into some sort of inner representational medium (for this point, see Clark & Toribio, 1994)? Not at all. Ordinary, non-representational devices can unproblematically yield the same output in response to a multifaceted and unruly input class. See (Ramsey, 2007, pp. 219-220; Facchin, 2026) for examples.

not grounded in mental content in at least two ways. Illusionism, in and by itself, is neutral in respect to both (it doesn't compel or favor one over the other).

First, the illusionist can join a broadly *pragmatist* tradition which takes semantic properties to originate from specific social, cultural practices (e.g., Wittgenstein, 2009; Haugeland, 1990; Brandom, 1994; Hutto & Satne, 2015). The illusionist can simply help themselves with one such account, and dodge the objection. And notice that, even on the quite dubious assumption that such an account is an account of the origins of contents in the relevant sense discussed here (i.e. content as "something" carried around by contentful states, see Myin & van den Herik, 2021), the account itself does not confer content to our mental states. After all, our cultural practices do not typically revolve around the neuronal states that should carry the relevant contents, and so would not assign contents to them.

The fact that the content illusionist has such a move to their avail has one important consequence I wish to highlight here: it allows the illusionist to also block a different intuitive realistic objection. The objection, roughly, contends that content illusionism cannot be coherently formulated, for illusions are *misrepresentations* (e.g., Macpherson, 2013): they are representations *whose content gets things wrong*. If this were the case, then, the illusion claim (i.e. mental content is an illusion) would presuppose exactly what the eliminative claim (i.e. there is no mental content) denies. Hence content illusionism would be internally inconsistent.

I already dealt with this objection in §2.2 when I elucidated the illusion of content in behavioral terms: to be under the spell of the illusion of content is to behave so as to satisfy the stereotype of the believer that one's mental contents are thus-and-so. Some may find these terms objectionable — for example, because they hold that the identification of such a behavioral stereotype still presupposes content. Such a burden (if at all present) can now be discharged by characterizing the illusion of content in terms of public representations (see Toon, 2023 for a similar idea). To be under the illusion of having some visual content, for example, is *to behave as if one has an "inner picture" of the external world*. To be under the illusion of having a mnemonic content is to behave *as if one was able to retrieve knowledge from a sort of "inner notebook"*. To be under the illusion of having a contentful imaginative state is *like painting a mental painting*. Dreams are like *inner movies*, and to think *is like talking to oneself*. Such metaphors offer the familiar basis through which we understand our own contentful mental states, and can aptly characterize the illusion of content in a way that *does not* presuppose mental content.[58]

Going back to semantic nihilims, there is a second move the illusionist can try to identify semantic, or at least proto-semantic, notions. After all, even in the absence of content, there can still be *functional success* and/or failure conditions that can be treated as at least proto-semantic under a broadly correspondence notion of semantics. A world inhabited only by paramecia would (most likely) be a world with no mental content, but the organelles inside the paramecia would still have functions, which they can still fail or succeed to perform. For example, ribosomes would still have the function of producing proteins by "reading" RNA sequences, and RNA bases would still have the function of "copying" DNA sequences. And these forms of functioning seem to presuppose a form of structural matching which can be usefully treated as a semantic or proto-semantic, natural form of "picturing" (see Sachs, 2019) — even if we are in no way tempted to think that RNA sequences somehow carry a "content" (a sort of special, *sui generis* object somehow representing or replicating the protein to be produced).

I have here defended content illusionism by three of the most obvious, most direct objections against it. At this point, content illusionism may appear as a genuine contender when it comes to dealing with mental content and its philosophical problems. Or, maybe, it would appear as a genuine contender if I had provided a positive story about illusionism — concerning, for example, how the illusion is generated. But, thus far, I haven't.

Time, then, to fill in this gap.

---

[58] Won't this characterization of the illusion clash with the behavioral one? I don't think so. To fit the behavioral stereotype is, in many cases, to behave as if one is guided by a number of mental analogs of external representations.

5. The illusion of content: Sketching an explanation

The aim of this section is to sketch an explanation to the illusion of content (an "error theory", as some philosophers might prefer to call it). As is common in the illusionist literature (see Dennett, 1991b, Frankish, 2026), the explanation will take the form of a mechanism.[59] The functioning of such a mechanism will, as I shall show, cause us to adopt the relevant behavioral stereotype of a person believing that they have mental contents thus-and-so.

Owing to space limitations and the programmatic character of the paper, I will be light on empirical details. For this reason, I will also try to sketch the *simplest possible* mechanism generating the illusion. As said above (§ 2.2) it is possible to call upon other, different and more complex mechanisms. For now, however, this simple one must do.

Notice that whilst here I aim at providing the simple possible mechanism, I don't think my explanation is a *how-possibly explanation*. Whilst surely simplified, the mechanism I am sketching is made up of two components that *are actually present* in our cognitive architecture. Let me introduce them.

5.1. Component #1: An internal controller

Some devices exhibit a behavior which is entirely controlled by environmental variables. So called "passive walkers", for example, are simple robots whose walking behavior is guided by the slope they are placed upon (McGeer, 1990).

Most animals are not devices of this kind. Whilst they remain open to environmental influences, they also have an inner *controller* enabling them to flexibly interact (and partially decouple from) environmental variables: the brain. More in detail, what such a controller must minimally do is to *keep track* of the organism's bodily variables (i.e. the organism's bodily state — think of it as the set of receptors and actuators "wired" to the brain), and keep them within precise bounds of viability. And notice that this "keep track" can simply take the form of detecting — and correcting for — deviations from a "default" state (Aikins, 1996; Bruineberg *et al.*, 2018). For example, if the appropriate receptors notice a deviation from the desired bodily temperature of 36.7°, the brain can correct for it inducing rhythmic muscle spasm (trembling), which will, in some circumstances, raise the temperature. Or, in the case of a more sophisticated control system, the correction can happen by operating on one's central heating system. The predictive processing theory of perception mentioned above (§3) offers a modern — and importantly *representation-free*[60] — rendition of this idea (see Clark, 2013, 2016; Facchin, 2021a,b; Seth, 2014b; Williams, 2018a,b; Bruineberg *et al.*, 2018). But the idea that the brain is such a control system predates it (e.g., Powers 1973; Ashby, 2013; Clark, 1997; Anderson, 2014) and seems to be certain beyond any reasonable doubt.[61] What is relevant to my purpose here is that a controller of this kind must exist (else animals would have a hard time keeping themselves alive) and that it must constantly keep track of the organism's physiological states: its bodily variables and the state of receptors and actuators.[62]

5.2. Component #2: Agent-detectable manipulative social signalling

Most animals must coordinate with other animals (of same or different species) to some end. This need for coordination gives rise to "signaling games" (Skyrms, 2010). Descriptively, the point is very familiar: green vervets emit specific

---

59 Really at the level of generality I will operate, it is a mechanistic sketch (Piccinini & Craver, 2011).

60 See references given in §3. The "modeling" talk that abounds in the predictive processing literature uses a sense of model that comes from the "good regulator theorem" (Conant & Ashby, 1970) and which is markedly non-representational. See Thobani (2024).

61 Of course this is not to claim that the brain performs only that role as a controller - a claim that runs into a vast number of psychological (Sun & Firestone, 2020) and neurological (Facchin & Viola, 2025) counterexamples.

62 Such architectures are also importantly entangled with illusionism about phenomenal consciousness, see (Dennett, 2013; Clark *et al.*, 2019; Frankish, 2026).

calls to "invite" other vervets to run away from specific predators (Seyfarath *et al.*, 1980); bees "dance" to signal to other bees the location of flowers to "harvest" (Von Frisch, 1993), *homo sapiens* emits sound waves to coordinate with each other in a range of situations, and so forth.[63]

Signalling games are at least at times open to *manipulation*: a participant of the game may emit a signal when the external circumstances *do not* call for its production, thereby forcing other participants to act inappropriately — which typically results in some (perhaps short-term) advantage for the signaler. A smart vervet may emit a predator call when no such predator is in sight, to force other vervets to run and stop a losing fight.[64] Humans lie. Some fireflies mimic the mating signals emitted by other species of fireflies — thereby luring their prey (Lloyd, 1965).[65]

Crucially, in at least some cases, such manipulations are *agent-detectable*: they have physiological markers the agent emitting them (or, its control system) can keep track of, and act upon. This makes good sense. Consider the vervet case above: the vervet emitting the call does so to "force" the fight to end, to regain energy and perhaps "steal" the good over which the vervets were fighting as other vervets run to hide. If that vervet were to run too, they would be very poorly served by their attempted manipulation! In the human case, such markers are many, and range from physiological changes in skin conductance to particular patterns of neural activity (Godert *et al.*, 2001, Meijer *et al.*, 2014; Mandal, 2014).[66]

5.3. The mechanism at work

To see how these two components interact to generate the illusion, consider the following — no doubt rather simplistic — scenario. Let S_1 and S_2 be two humans engaged in a common activity. We can imagine them being prehistoric humans hunting for a bison — this will help us keep their mind and signaling games simple, and keep S_1 and S_2 as culturally unsophisticated as a human can be. Thus, imagine that, with some reliability, S_1 spots bisons, signals their presence to S_2, and then the two proceed to hunt their prey.

Let me now vary their routine. Imagine that, by some accident (*which* accident is not very important), S_1 enters in the physical states of action readiness that follow the spotting of a bison *when no bison is actually around* (and so no bison can have been spotted). S_1's brain is therefore preparing and "launching" all the relevant bison-related (motoric and non-motoric) responses: S_1 is thus emitting the relevant signal (e.g., turning towards S_2 saying "Bison!") and starting the hunting routine (which includes a number of bodily movements and physiological changes, such as the release of adrenalin). Yet no bison is there. Thus S_1's internal state and behavior will fail to *evolve* in the usual way, and S_1 (or its brain) will register the fact. Roughly put, there is no bison-related sensory stimulus incoming, and S_1's brain will arrest the relevant motor routine and physiological changes just launched.

S_1 would, at this point, be quite perplexed. S_1 sees now that there is no bison out there: his brain is not tracking the presence of bison-related states at the sensory periphery, if you wish. Nor has his brain tracked any bison-moving-away related sensorimotor state — so S_1 knows the bison has not run away. So, S_1's behavior fits the stereotype of the believer that there is no bison out there (and indeed, S_1 is not trying to track the bison, to run away from it, to surprise it in an

---

[63] Aren't such signals contentful? Perhaps. But these signals would at best be external (non-mental, public) representations, which the illusionist can allow for (§4). Do signaling games presuppose mental contents - for example, communicative intentions, or representations of communicative goals? Arguably, no. Bacteria take part in - sometimes very complex - signaling games (Diggle *et al.*, 2007; Hasrem & Van Impe, 2022), and presumably bacteria do not mentally represent a thing.

[64] Anecdotically attested in Cheney & Seyfarth (1990, pp. 213-216).

[65] Again, fireflies do so presumably without having to internally represent anything, so again manipulative signaling does not presuppose content.

[66] Can elements outside the nervous system of an agent be components of the mechanisms responsible for the agent's mentality? The short answer is: yes (Kaplan, 2012). Components of mechanisms need not be spatially contiguous, or part of the same anatomically defined system.

ambush, etc.). But S_1 will also fit the stereotype of the believer *that they saw a bison* — their brain didn't register *any* of the markers of manipulative signaling. Thus, if asked, S_1 would defend the claim that they really saw a bison, that the signal they omitted was no joke, that they themself was ready to attack the bison, and so forth. More simply put: S_1's behavior is such that we are licensed to say that S_1 believes that a bison was seen, but there is no bison out there.

The obvious conclusion — that is, the behavioral stereotype S_1 is forced towards by the behaviors above — is that *the seen bison must have been "in here"*, in S_1's mind. Thus, S_1 will say that they saw a bison that nevertheless wasn't out there (but really *seemed* out there), that S_2 wasn't able to see such bison, and that when they looked at the spot again the bison simply wasn't there anymore. It's easy to see how a little bit of cultural evolution could lead S_1 to say he had a sort of "private mental picture" of the bison — and this is roughly our folk conception of perceptual content (Fischer *et al.*, 2026).[67] But even without such cultural scaffolds[68], it is clear that S_1 is now pushed towards the belief that there are certain mental "objects" that are not "out there" but rather inhabit our minds. That is, S_1 is pushed towards the behavioral stereotype of a believer in such objects: S_1 will talk about them, say that they are present — perhaps even given — to their mind, produce genuine introspective reports about those, and so forth.

I could add bells and whistles to this mechanism sketch. For example, adopting a "predictive processing" view and its idea of precision may explain why we are *particularly certain* that mental contents exist (cf. Clark *et al.*, 2019). Calling upon a specific socio-material embedding may explain why the illusion has, for us, the specific features it has (why, for example, we take perception to be some sort of picture, cf. Draisma, 2000; Toon, 2023). But these are, in an important sense, bells and whistles. The core of the mechanism consists in the non-manipulative production of signals that falsely express and report what's going on in our mind.[69]

5.4. Illusionism as the best explanation

In §3.1 I argued that illusionism is quantitatively and qualitatively more parsimonious than realism. I also asserted (without much of an argument) that it is also more mechanistically (and thus, explanatorily) parsimonious. Moreover, in §4.1 I mentioned that the realist could uphold their belief in content on the abductive grounds provided by an inference to the best explanation. And while above I anticipated something in the way of an answer, my points were left largely unsubstantiated. Effectively, I took an *argumentative loan* for the sake of exposition. It is now time to start paying that loan back.

Let me then start with the shape of the realist's alternative. In a nutshell, the realist's alternative is a simple *detection mechanism*. The mechanism has some part that immediately detects the relevant content, and an output channel whereby the detector mechanism can influence the production of the subject's reports concerning their mental contents. Such a mechanism surely seems simpler than the one I sketched above, and so allows us also to offer a better — simpler and more direct — explanation of our reports.

Isn't such a simple detection mechanism *obviously* mechanically simpler than the mechanism I sketched above? How can the illusionist argue that this mechanism is more *mechanistically complicated* than the one I sketched? I cannot deny that the simple detection mechanism looks simpler than the one I sketched. Once we try to think about the mechanism's *details*, however, the apparent simplicity rapidly disappears.

---

[67] Indeed, the picture is so popular that it's also shared by the non-folk. For example, Chris Frith is a neuroscientist having a sophisticated and complex theory of our minds, and yet still he subscribes to that picture of content (Frith, 2007, p.135).

[68] Their mention is suggestive of an "extended mind" (Clark, 2008) picture of the illusion, where the illusion is at least in part constituted by our external representational practices and our reflection on them (cf. Dennett, 2000). This is a fascinating theme I must pass under silence for the sake of brevity.

[69] This is also what John was doing in the example above (§2.2).

The first, and most obvious problem we run into is that the realist *must* commit to all the components of my story too. These components are so simple, undemanding — and I dare to say *obviously real* — that to deny their existence seems if not impossible, at least a super-herculean task. To deny that brains track and monitor bodily states, or that there are manipulative signals in signaling games one must have a pretty revolutionary theory of brains and communicative behaviors. And to motivate the adoption of such a theory *is on them*.

Noticing this is already sufficient to "cash in" the previous point concerning mechanistic simplicity. The realist is prima facie committed to the existence of all the components I have pointed out above, to which they add a commitment to the simple detection mechanism I outlined above.[70] The realist, then, puts *more mechanisms* in our head, giving us a *more mechanistically complicated* picture of our cognitive architecture.

And, once we think about *how* the simple detection mechanism gets "wired" to the ordinary circuitry of the brain, the complexity becomes quite dazzling. For, the simple detection mechanism is something like a "Cartesian theatre" (Dennett, 1991b): a mechanism *isolated from* the ordinary mechanism of the brain, to which all contents converge and are shown for our benefit. Contents must thus be forwarded to the detection mechanism (or the detection mechanism must be able to move around the brain), complexifying the (already quite complex) connection architecture of our brain — to the point that such a centralized detection mechanism may be *biologically* impossible (Mesulam, 2008). And notice that making such a mechanism a "virtual" mechanism, or part of the brain "software", does not solve the problem. According to all physical theories of computation (Piccinini, 2025) there must be an *isomorphism* between computational states and the substrate implementing them. Thus, were such a "virtual mechanism" implemented in the brain, *something in the brain* should go through the relevant state changes of the virtual mechanism. There would still be a piece of the brain wherein all content must be forwarded. Thus, the realist actually *complexifies* the overall picture of mental mechanisms, adding a component. In contrast, the illusionist does not call upon ad hoc mechanisms, and explains the illusion with "the regular" mechanisms of the mind and brain (cf. Dennett, 1991b, 2025; Frankish, 2026).[71]

But the above is only the first problem (or, to be more precise, set of problems) the realist must face. A second problem (or set thereof) concerns *content detection*. The realist must in fact offer an account of it, and yet it is very hard to see how such an account could go. Normally, detection happens in virtue of the fact that the detected state causally influences the detector. Some examples to clarify: litmus papers detect the Ph. of a substance because the substance Ph. causes them to change color. (Bulb) thermometers detect temperature because the temperature causes the fluid in the thermometer to expand or contract, going "up" and "down". Optic smoke detectors detect smoke (and thus, fires), because the smoke enters the inner chamber of the detector, causing the beam of light inside it to scatter (which causes the detector's photosensitive cell to ring the alarm).

As argued above, however, content seems to lack the causal powers needed to causally influence anything. This means that the realist either lets go of the idea that contents are detected — which would of course force us to reject the simple model sketched above — or comes up with some *non-causal* story about detection. I don't see what form such a model may take, and more generally this seems a significant complication on the simple story the detection mechanism allegedly affords.

Thirdly and lastly, the simple detection mechanism *can't be the whole story about our introspective reports*. It is well known that some introspective reports are confabulations, and thus they report contents that have not (indeed, cannot) have been detected. Some of the most famous cases include neurotypical subjects making up on the spot some "in-

---

[70] This inflation of commitments is of course also made more acute by the realist's commitment to the existence of content.

[71] These problems are also compounded by the fact that this mechanism also must interface and communicate with all other neural and cognitive mechanisms, in a way that can generate all sorts of conflicts between them. Consider again the "bison" example sketched above. Why hasn't the simple introspective mechanism vetoed the utterance of "bison", given that no corresponding visual content was introspectable (notice that, by construction of the case, S1 enters in the states that follow the perception of a bison, so no "visual" bison content can be introspected)? These kinds of conflicts are a problem for the realist to solve - and their solution is guaranteed to further complicate the realist's story. Dennett's (1991b) discussion of Orwellian and Stalinesque models offers a vivid illustration of (some) such problems.

trospected preferences" to justify their choices (Nisbett & Wilson, 1977), and the case of split brain patients confabulating a reasons for their actions when the stimulus motivating them was given to the right cerebral hemisphere, which is unable to produce verbal reports (Gazzaniga, 1995).[72]

Indeed, introspective reports are likely produced by many different mechanisms (Schwitzgebel, 2012), generating a multiplicity of reports and a variety of introspective practices. Some reports are confabulations, invented "on the spot" to justify our behaviors (Chater, 2018). Other reports conform to our socially shared narrative practices (Hutto, 2012), allowing us to partake in a social community.[73] Other reports are free acts of self-exploration, whereby we strive to interpret and make sense of how we physiologically feel (Facchin *et al.*, 2025). Other reports are optimistically biased to make us self-confident and keep us going, no matter how objectively bleak the situation is (e.g., Sharot, 2011). Reports may simply express how we feel and what is going on inside our minds (e.g., Wittgenstein, 2008), or the upshot of our decisions (Moran, 2001), or even be an attempt to mold us and our behavior into a form we find more congenial to us (Hutto, 2016; 2022). Our introspection, and the rich mental life it purportedly shows, is manifestly *not* just the effect of a mechanism that detects what's currently going on in our mind.

Sure, to acknowledge such a complexity — and such a variety of introspective practices — *does* force me to complexify the mechanism I sketched above too. But notice how such a complexification is not really damaging for the illusionist.

On the one hand, the basic story I sketched about the illusion can be replicated. Regardless of the introspective practice and the specific mechanism generating them, *if* the participant produces a genuine (non-manipulative, in the sense seen above) report, then the participant will end up *fitting the behavioral stereotype* of the believer that their mental content is thus-and-so. And, if there is no such content (as I have argued here), then the participant is under the illusion. The illusionist, then, can simply (roughly) repeat the same explanatory story for each and every introspective practice. The realist cannot. They need to posit different, additional mechanisms. As a result, their explanation will also be way less unified and homogeneous.

On the other hand, whilst such a plurality of mechanisms and introspective practices is surely compatible with realism (cf. Schwitzgebel, 2012), it is not really *predicted* by most realist accounts. Realists tend to have a monistic view of introspection, where introspection is the output of a single mechanism. This is indeed the basis of the various "-isms" competing to offer accounts of introspection. For example: expressivism explains introspection in terms of the expression of mental states (e.g., Finkelstein, 2008), whereas inferentialists explain introspection calling upon inferential mechanisms, perhaps operating on a theory of mind or folk psychology (e.g., Gopnik, 1993). Illusionists, in contrast, have historically expected such a plurality of mechanisms. Dennett's (2005, pp. 72-76) *tuned deck* story colorfully conveys the point. The story's protagonist is the conjurer Ralph Hull, apparently able to play a "tuned deck" trick no other conjurer was able to replicate, or even understand. Later, Hull revealed the trick: there simply was no "tuned deck" trick! Hull had at his avail a *plurality* of (already well known) different moves yielding the same output (e.g., guessing the card a member of the public chose, or whatever). When Hull announced he was going to perform the "tuned deck" trick, he simply performed one of those tricks — cycling through the moves made between each iteration of the "trick", which made the tuned deck trick impossible to track: when another conjurer were looking for a move A, Hull made a move B; when a conjurer were controlling for a move B, Hull performed a move C — and so forth.

By analogy, illusionists expect the illusion to be the product of multiple different mechanisms — which make the illusion hard to disciple, as there is no single mechanism always responsible for it we can always call upon to explain content away.

---

[72] See also Chater (2018) for several other cases. Also notice that such confabulations are not things that subjects willingly make up. They indeed reflect the subject's "honest" view of reality (see Schneider, 2003).

[73] Mesquita (2022 pp. 200-203) offers a striking example with the case of Ramla, a Muslim girl that, in virtue of his Muslim upbringing (and the emphasis on mutual social obligations such upbringing placed on her) reported feeling ashamed in many instances wherein a westerner would have rather felt sad.

The above, I take, makes a case for illusionism being *less* mechanistically complex than realism. The very same considerations help the illusionist respond to a different realist's counterargument; namely that the existence of contents can be abductively inferred by the *best explanation* of their apparent givenness.

The first step in this direction is that of noticing that the *existence* of contents isn't in and by itself sufficient to explain their *apparent givenness*.[74] Hence, appealing only to the existence of content does not explain (let alone best explains) their apparent givenness. To see why this is the case, imagine an agent having only non-introspectively detectable contentful states (for example, an agent having only "neural", sub-personal representational states). Were such an agent to exist, content would (by construction of the case) exist. And yet, the mere existence of content would not explain why content appears to be given to the agent. Indeed, in the scenario described, these constants wouldn't be given (apparently or not) to the agent *at all* — for the agent has no way to access them (again, by construction of the case).

So, in and by itself, the existence of contents *does not* explain their apparent givenness. The existence of content can of course be *part* of the explanation of why content seems given only if it is joined by something else; namely a mechanism that allows the agent to access their own content. And the detection mechanism I sketched above is the simplest mechanism at the realists' avail. And, as I argued above comparing it to its illusionistic alternative, that mechanism does not offer a *better* explanation of the apparent givenness of content.

The realist could of course reply that the simple detection mechanism I have put forth and examined is *too* simple, and that some more complex and realistic mechanism would indeed provide a better explanation. Indeed, the space of possible introspective mechanisms is quite large (see Kammerer & Frankish, 2023) — and whilst space limitations forbid its exploration here, I must confess I have no straightforward argument to *deny* that there is no *simpler* mechanism able to satisfy the realist.

Perhaps, then, it is convenient to put the point as a *challenge* to the realist: the realist has to provide a mechanism that is (a) simpler than the one I sketched, and (b) able to overcome the problems I have highlighted. I have a hard time imagining how the realist could do that — and it is on the realist to prove that I'm wrong.

6. Conclusions

I have introduced content illusionism to the literature. I have motivated it, defended against some obvious objections, and sketched a mechanism responsible for the illusion. Of course, this is just the first step. In an important sense, the real work to develop and defend content illusionism begins now — after the more programmatic points have been made.

In particular, it is necessary to:

(i) provide a fuller, more detailed mechanistic sketch of the mechanisms responsible for the illusion of content.
(ii) elaborate a richer and clearer picture of what the illusion of content amounts to and its origins, both developmentally (are we born with it or we acquire it?) and phylogenetically (presumably, it is not coextensive with life and it has "popped up" on *some* branch of the tree of life).
(iii) gauge the epistemological implications of such a view. If there are no contents, how should we understand belief and knowledge? How should we understand rationality and inference?

[74] Of course, were the realist able to demonstrate that contents exist, they would prove that illusionism is wrong. But, in the objection I'm considering, the realist is inferring (abductively) to the existence of content based on the fact that it best explains their apparent givenness. And I'm contesting that the existence of content alone offers such an explanation.

(iv) explore the relations between content illusionism and illusionism about phenomenal consciousness. It is intuitive to think that the two views must be coherent and mutually supportive — and yet illusionists about phenomenal consciousness heavily rely on representations in their explanatory practices (e.g., Kammerer, 2021).

There is, then, plenty of work for illusionists to do. I hope the present paper has persuaded you this is *valuable* work, and enticed you in developing your view of content in an illusionist direction (*especially* if you are a radical enactivist).

Acknowledgements

An early version of this paper (in late '22, early '23) was read and commented upon by (in random order) Bruno Cortesi, Valentina Martinis, François Kammerer, Angela Mendelovici, Thomas van Es, Zuzanna Rucinska, Erik Myin, Farid Zahnoun, Giulia di Rienzo, David Schoute and Wouter van Hooydonk. Thank you for your comments and questions. The paper in this version has been presented to the *Expaning Mind* seminars in Wollongong. I thank the whole crowd of the Gong, but a special thanks goes to Dan Hutto and Erik Myin for their engagement during the seminar.

This work has been financed with the generous FWO (*Fonds Wetenschappelijk Onderzoek*) grant "Towards a globally non-representational theory of the mind" (grant number 1202824N)

REFERENCES

Aikins, K. (1996). Of sensory systems and the aboutness of mental states. *Journal of Philosophy, 93*(7), 337-372.

Aizawa, K., & Adams, F. (2005). Defending non-derived content. *Philosophical Psychology, 18*(6), 661-669.

Amir, D., & Firestone, C. (2025). Is visual perception WEIRD? The Müller-Lyer illusion and the cultural byproduct hypothesis. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000549

Anderson, M. L. (2014). *After Phrenology*. The MIT Press.

Anderson, M. L., & Champion, H. (2022). Some dilemmas for an account of neural representation: A reply to Poldrack. *Synthese, 200*(2), 169.

Artiga, M., & Sebastián, M. Á. (2020). Informational theories of content and mental representation. *Review of Philosophy and Psychology, 11*(3), 613-627.

Ashby, W. (2013). *Design for a brain: The origin of adaptive behaviour*. Springer Science & Business Media.

Baker, L. R. (1987). *Saving belief: A critique of physicalism*. Princeton University Press.

Baker, B., *et al.* (2022). Three aspects of representation in neuroscience. *Trends in Cognitive Sciences, 22*(11), 942–958

Balog, K. (2023). The abolition of phenomena: A voyage among the zombies. *Klēsis Revue Philosophique* 55. https://www.revue-klesis.org/pdf/klesis-55-07-katalinbalog-the-abolition-of-phenomena-a-voyage-among-thezombies.pdf

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020, July). Predictions in the eye of the beholder: An active inference account of Watt governors. In J. Bongard, J. Lovato, L. Hebert-Defruésene, R. Dasari & L. Soros, *Artificial Life Conference Proceedings, 32* (pp. 121-129). The MIT Press. https://direct.mit.edu/isal/proceedings-pdf/isal2020/32/121/1908564/isal_a_00288.pdf

Block, N. (2023). *The border between seeing and thinking*. Oxford University Press.

Bocanegra, B. R., Poletiek, F. H., Ftitache, B., & Clark, A. (2019). Intelligent problem-solvers externalize cognitive operations. *Nature Human Behaviour*, 3(2), 136-142.

Bourget, D., & Mendelovici, A. (2019). Phenomenal intentionality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/phenomenal-intentionality/

Bovet, S. (2007). *Robots with Self-Developing Brains*. (Doctoral dissertation). University of Zurich.

Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. The MIT Press.

Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.

Brentano, F. (2012). *Psychology from an empirical standpoint*. Routledge.

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences, 42*, e215. https://doi.org/10.1017/S0140525X19000049

Brette, R. (2026). *The Brain, in theory*. Princeton University Press.

Brooks, R. A. (1999). *Cambrian intelligence: The early history of the new* AI. The MIT Press.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. S*ynthese, 195*(6), 2417-2444.

Carvalho, J. T., & Nolfi, S. (2016). Cognitive offloading does not prevent but rather promotes cognitive development. *PloS One, 11*(8), e0160679. https://doi.org/10.1371/journal.pone.0160679

Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory*. Oxford University Press

Chater, N. (2018). *Mind is flat: The remarkable shallowness of the improvising brain*. Yale University Press.

Cheney, D. L, and Seyfarth, R. M. (1990). *How Monkeys See the World*. University of Chicago Press

Chomsky, N. (1995). Language and nature. *Mind, 104*(413), 1-61.

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

Churchland, P. M. (1992). *A neurocomputational perspective*. The MIT Press.

Churchland, P. M. (2012). *Plato's camera*. The MIT Press.

Churchland, P. S., *et al.* (1994). A critique of pure vision. In C. Koch & J. Davies. *Large scale theories of the brain* (23-60). The MIT Press.

Clark, A: (1997). *Being there*. The MIT Press.

Clark, A. (2008). *Supersizing the mind*. Oxford University Press.

Clark, A. (2009). Spreading the joy? Why the machinery of consciousness is (probably) still in the head. *Mind, 118*(472), 963-993.

Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind, 121*(483), 753-771.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences, 36*(3), 181-204.

Clark, A. (2014). Perceiving as predicting. In D Stokes, M. Matthen & S. Biggs (Eds.) *Perception and its modalities* (pp. 23-43). Oxford University Press.

Clark, A. (2015). Embodied prediction. In T. Metzinger & W. J. Windt. (Eds.). *Open Mind, 7* (pp.1-21). MIND Group. https://doi.org/10.15502/9783958570115

Clark, A. (2016). *Surfing Uncertainty*. Oxford University Press.

Clark, A. (2017). How to knit your own Markov Blanket: Resisting the second law with metamorphic minds. In T. Metzinger, & W. Wiese (Eds.), *Philosophy and Predictive Processing, 3* (pp. 1-19). MIND Group. https://doi.org/10.15502/9783958573031

Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies, 26*(9-10), 19-33.

Clark, A., and J. Toribio. 1994. Doing without representing? *Synthese 101*(3), 401–431

Conant, R. C., & Ashby W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systemic Science, 1*(2), 89-97

Crane, T. (1998). Intentionality as the mark of the mental. *Royal Institute of Philosophy Supplements, 43*, 229-251.

Crane, T., & French, C. (2021). The problem of perception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/perception-problem/

Cummins, R. (1989). *Meaning and mental representation*. The MIT Press.

Ćwiek, A., *et al.* (2022). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B, 377*(1841), 20200390.

Das, N. (2021). Vācaspati on aboutness and decomposition 1. In C. Meyns (Ed.). *Information and the History of Philosophy* (pp. 79-106). Routledge.

Davidson, D. (2016). Knowing one's own mind. In A. Pessin & S. Goldberg (eds.). *The Twin Earth Chronicles* (pp. 323-341). Routledge.

Davies-Barton, T., Raja, V., Baggs, E., & Anderson, M. L. (2026). Debt-free intelligence: Ecological information in minds and machines. *Philosophical Psychology, 39*(2), 239-265.

De Haan, S., Rietveld, E., Stokhof, M., & Denys, D. (2013). The phenomenology of deep brain stimulation-induced changes in OCD: an enactive affordance-based model. *Frontiers in human neuroscience, 7*, 653. https://doi.org/10.3389/fnhum.2013.00653

Demeter, T. (2022). A mental fictionalism worthy of its name. In T. Demeter, T. Parent, & A. Toon (eds.). *Mental Fictionalism: Philosophical Explorations* (pp. 86-103). Routledge.

Dennett, D. C. (1987). *The Intentional Stance*. The MIT Press.

Dennett, D. C. (1988). Quining qualia. In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science* (pp. 42–77). Oxford University Press.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy, 88*(1), 27-51.

Dennett, D. C. (1991b). *Consciousness Explained*. Little Brown

Dennet, D. C. (1992). The self as a center of narrative gravity. In F. Kessel, P. Cole, D. Johnson, M. Hakel (Eds.) *Self and Consciousness: Multiple Perspectives*. Erlbaum.

Dennett, D. C. (2000). Making Tools for Thinking. In D. Spreber (Ed.) *Metarepresentation: A Multidisciplinary Perspective* (17-29). Oxford University Press.

Dennet, D. C. (2005). *Sweet Dreams*. The MIT Press.

Dennett, D. C. (2013). Expecting ourselves to expect: The Bayesian brain as a projector. *Behavioral & Brain Sciences, 36*(3), 209-210.

Diggle, S. P., Gardner, A., West, S. A., & Griffin, A. S. (2007). Evolutionary theory of bacterial quorum sensing: when is a signal not a signal? *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1483), 1241-1249.

Downey, A. (2016). Radical sensorimotor enactivism & predictive processing. In T. Metzinger, & W. Wiese (Eds.). *Philosophy and Predictive Processing*, 10 (pp.1-26). MIND Group. https://doi.org/10.15502/9783958573123

Downey, A. (2018). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese, 195*(12), 5115-5139.

Downey, A. (2024). Theatre Hunger: An Underestimated 'Scaling Up' Problem. *Erkenntnis, 90*(6), 2503-2521

Draaisma, D. (2000). *Metaphors of Memory: A History of Ideas about the Mind*. Cambridge University Press.

Drayson, Z. (2025). Representations are (still) theoretical posits. THEORIA. *An International Journal for Theory, History and Foundations of Science, 40*(3), 240–255. https://doi.org/10.1387/theoria.27209

Egan, F. (2020). A deflationary account of mental representations. In J. Smortchkova, K. Dolega, & T. Schlicht (Eds.), *What are Mental Representations?* (pp. 26–54). Oxford University Press.

Facchin, M. (2021a). Predictive processing and anti-representationalism. *Synthese, 199*(3), 11609-11642

Facchin, M. (2021b). Are generative models structural representations? *Minds and Machines, 31*(2), 277-303.

Facchin, M. (2022). *The metaphysics of Predictive Processing: A non-representational account*. Ph.D. (Doctoral dissertation). IUSS Pavia.

Facchin, M. (2023). Neural representations unobserved — or: A dilemma for the cognitive neuroscience revolution. *Synthese, 203*(1), 7.

Facchin, M. (2025). Is mental content an illusion? *Synthese, 206*(5), 1-29.

Facchin, M. (2026). Defusing the Representation-Hungry Challenge. *Review of Philosophy and Psychology*, 1-28. https://doi.org/10.1007/s13164-026-00822-2.

Facchin, M., & Leonetti, G. (2024). Extended animal cognition. *Synthese, 203*(5), 138.

Facchin, M., & Rucińska, Z. (2025). Public charades, or how the enactivist can tell apart pretense from non-pretense. *Erkenntnis, 90*(5), 2095-2117.

Facchin, M., Rucińska, Z., & Fondelli, T. (2025). Radically embodied introspection. *Topoi, 45*, 373-389.

Facchin, M., & Viola, M. (2025). Structure and function in the predictive brain. *Biology & Philosophy, 40*(6), 1-26.

Favela, L. H., & Machery, E. (2023). Investigating the concept of representation in the neural and psychological sciences. *Frontiers in Psychology, 14*, 1165622. https://doi.org/10.3389/fpsyg.2023.1165622

Favela, L. H., & Machery, E. (2025). The concept of representation in the brain sciences: The current status and ways forward. *Mind & Language, 40*(2), 215-225.

Finkelstein, D. H. (2008). *Expression and the Inner*. Harvard University Press.

Fischer, E., Allen, K., & Engelhardt, P. E. (2026). Scientific or naïve? Perceptions of direct and indirect realism, and why they matter. *Philosophy and Phenomenological Research, 112*(1), 102-129.

Fodor, J. (1984). Observation reconsidered. *Philosophy of science, 51*(1), 23-43.

Fodor, J. (1987). *Psychosemantics*. The MIT Press.

Fodor, J. (1990). *A theory of content and other essays*. The MIT Press.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies, 23*(11-12), 11-39.

Frankish, K. (2023). What is Illusionism? *Klēsis – Revue de philosophie, 55*. https://www.revue-klesis.org/pdf/klesis-55-04-keith-frankish-what-is-illusionism.pdf

Frege, G. 1918 [1997]. Thought. In M. Beaney (ed.), *The Frege Reader* (pp. 325-45). Blackwell.

Frankish, K. (2026). The Emancipation of the drone: A Dennettian thought experiment. *Philosophical Psychology, 39*(3), 910–940.

Friston, K. (2011). Embodied inference: Or "I think therefore I am, if I am what I think." In W. Tschacher & C. Bergomi (Eds). *Implications of Embodiment: Cognition and Communication*, (pp. 89–125). Imprint Academic.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in psychology, 3*, 151. https://doi.org/10.3389/fpsyg.2012.00151

Frith, C. (2007). *Making Up the Mind*. Oxford University Press.

Gallagher, S. (2017). *Enactivist interventions.* Oxford University Press.

Gan, N. (2024). What Kind of Non-Realism is Fictionalism? *Ergo: An Open Access Journal of Philosophy, 15*. https://doi.org/10.3998/ergo.5720

Gazzaniga, M. S. (1995). Principles of human brain organization derived from split-brain studies. *Neuron, 14*(2), 217–228.

Gedenryd, H. (1998). *How designers work: Making sense of authentic cognitive activities*. (Doctoral dissertation). Lund University.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*(2), 559-582.

Gödert, H. W., Rill, H. G., & Vossel, G. (2001). Psychophysiological differentiation of deception: The effects of electrodermal lability and mode of responding on skin conductance and heart rate. *International Journal of Psychophysiology, 40*(1), 61-75.

Godfrey-Smith, P (2006). Mental representations, naturalism and teleosemantics. In G. MacDonald & D. Papineau (eds.) *Teleosemantics* (pp.42-68). Oxford University Press.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain sciences, 16*(1), 1-14.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press.

Hall, M., & Robinson, D. (2021). Acoustic signalling in Orthoptera. *Advances in Insect Physiology 61*, 1-99.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2-3), 146-162.

Haruno, M., Wolpert, D. M., & Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural computation, 13*(10), 2201-2220.

Hashem, I., & Van Impe, J. F. (2022). Dishonest signaling in microbial conflicts. *Frontiers in microbiology, 13*, 812763. https://doi.org/10.3389/fmicb.2022.812763

Haugeland, J. (1990). The intentionality all-stars. *Philosophical perspectives, 4*, 383-427.

Haugeland, J. (1991). *Representational genera.* In *Having Thought* (pp.171-207). Harvard University Press.

Heemskerk, J. (2024). *How informational teleosemantics works: Towards a realist theory of content* (Doctoral dissertation). University of Warwick.

Heidegger, M. (1926/1962). *Being and Time*. Harper & Row.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger, & W. Wiese (Eds.), *Philosophy and Predictive Processing, 2* (pp. 1–15). MIND Group. https://doi.org/10.15502/9783958573048

Hurley, S. (1998). *Consciousness in Action*. Cambridge University Press.

Husserl, E. (1970). *The crisis of European sciences*. Northwestern University Press.

Hutto, D. D. (2012). *Folk psychological narratives: The sociocultural basis of understanding reasons*. The MIT Press.

Hutto, D. D. (2016). Narrative self-shaping: A modest proposal. *Phenomenology and the cognitive sciences, 15*(1), 21-41.

Hutto, D. (2022). Getting real about pretense: A radical enactivist proposal. *Phenomenology and the Cognitive Sciences, 21*(5), 1157-1175.

Hutto, D. D., & Myin, E. (2013). *Radicalizing Enactivism*. The MIT Press.

Hutto, D. D., & Satne, G. (2015). The natural origins of content. *Philosophia, 43*, 521-536.

Ittelson, W. H. (1952). *The Ames demonstrations in perception; a guide to their construction and use*. Princeton University Press.

Jacobson, A. J. (2013). *Keeping the world in mind: Mental representations and the sciences of the mind*. Palgrave MacMillan

James, W. (1911). The tigers in India. In *The meaning of Truth* (pp. 43-50). Longman Green & Co.

Kammerer, F. (2021). The illusion of conscious experience. *Synthese, 198*(1), 845-866.

Kammerer, F., & Frankish, K. (2023). What forms could introspective systems take? A research programme. *Journal of Consciousness Studies, 30*(9-10), 13-48.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., & Mack, S. (Eds.). (2012). *Principles of neural science* (Vol. 4, pp. 1227-1246). McGraw-hill Medical.

Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. *Biology & Philosophy, 27*(4), 545-570.

Kaplan, H. S., & Zimmer, M. (2020). Brain-wide representations of ongoing behavior: A universal principle? *Current opinion in neurobiology, 64*, 60-69

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese, 195*(6), 2387-2415.

Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical explorations, 21*(2), 264-281.

Kohár, M. (2023). *Neural machines: A defense of non-representationalism in cognitive neuroscience*. Springer.

Krakauer, J. W., & Ramsey, W. (2026). Mental representation without neural representation: Understanding the evidence. *Philosophy and the Mind Sciences, 7*(1). https://doi.org/10.33735/phimisci.2026.12204

Kriegel, U. (2011). *The sources of intentionality*. Oxford University Press

Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current opinion in neurobiology, 55*, 167-179.

Kripke, S. (1980). *Naming and Necessity*. Harvard University Press

Langland-Hassan, P. (2022). Why pretense poses a problem for 4E cognition (and how to move forward). *Phenomenology and the Cognitive Sciences, 21*(5), 1003–1021

Lee, J. (2019). Structural representation and the two problems of content. *Mind & Language, 34*(5), 606-626.

Lee, J. (2021). Rise of the swamp creatures: Reflections on a mechanistic approach to content. *Philosophical Psychology, 34*(6), 805-828.

Leinweber, M., *et al.* (2017). A sensorimotor circuit in the mouse cortex for visual flow prediction. *Neuron, 95*(6), 1420–1432

Lewis, D. (1973). *Counterfactuals*. John Wiley & Sons.

Lloyd, J. E. (1965). Aggressive mimicry in Photuris: Firefly femmes fatales. *Science, 149*(3684), 653-654.

Lycan, W. G. (2022). A Rylean Mental Fictionalism. In T. Demeter, T. Parent, & A. Toon (eds.). *Mental Fictionalism: Philosophical Explorations* (pp. 70–85). Routledge

Macpherson F. (2013). The philosophy and psychology of hallucination: An introduction. In F. Macpherson, D. Platchias (eds.). *HALLUCINATION - philosophy and psychology*. (pp. 1-39). Oxford University Press.

Mandal, F. B. (2014). Nonverbal communication in humans. *Journal of Human Behavior in the Social Environment, 24*(4), 417–421.

Mandik, P. (2016). Meta-illusionism and qualia quietism. *Journal of Consciousness Studies, 23*(11-12), 140-148

Martínez, M. (2025). The information-processing perspective on representation. *Philosophy and the Mind Sciences, 6*. https://doi.org/10.33735/phimisci.2025.11321

McGeer, T. (1990). Passive dynamic walking. Int. J. Robotics Res., 9(2), 62-82.

Meijer, E. H., Selle, N. K., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology, 51*(9), 879-904.

Meinong, A. (1960), The theory of objects, in R. Chisholm (ed.). *Realism and the Background of Phenomenology*. Free Press.

Mendelovici, A. (2018). *The Phenomenal Basis of Intentionality*. Oxford University Press.

Mendelovici, A. (2026). Belief as self-verifying fictions. In J. Lewis-Jong & E. Schwitzgebel (eds.) *The Nature of Belief* (pp. 40-66). Oxford University Press.

Mendelovici, A., & Bourget, D. (2023). Facing up to the problem of intentionality. *Philosophical Perspectives, 37*(1), 228-247.

Merleau-Ponty, M. (1963). *The structure of behavior*. Beacon Press.

Mesquita, B. (2022). *Between Us: How Cultures Create Emotions*. WW Norton & Company.

Mesulam, M. (2008). Representation, inference, and transcendent encoding in neurocognitive networks of the human brain. *Annals of neurology, 64*(4), 367-378.

Millikan, R. G. (1984). *Language, Thought and other biological categories*. The MIT Press.

Millikan, R. G. (1993). *White queen psychology and other essays for Alice*. The MIT Press.

Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton University Press.

Myin, E., & Van den Herik, J. C. (2021). A twofold tale of one mind: revisiting REC's multi-storey story. *Synthese, 198*(12), 12175-12193.

Nanay, B. (2023). *Mental imagery: Philosophy, psychology, neuroscience*. Oxford University Press.

Nanay, B. (2025). Mental imagery, predictive processing, aphantasia, and the interaction between philosophy and cognitive science: Responses to Amy Kind, Christian Scholz, and Neil Van Leeuwen. *Mind & Language, 40*(3), 333-340.

Neander, K. (2017). *A Mark of the Mental*. The MIT Press.

Nes, A. (2008). Are only mental phenomena intentional? *Analysis 68*, 205–15

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Nizami, L. (2019). Information theory is abused in neuroscience. *Cybernetics & Human Knowing, 26*(4), 47-97.

Noe, A. (2004). *Action in Perception*. The MIT Press.

Nolfi, S. (2002). Power and the limits of reactive agents. *Neurocomputing, 42*(1-4), 119-145.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). *The building blocks of interpretability. Distill, 3*(3), e10. https://doi.org/10.23915/distill.00010

Orlandi, N. (2014). *The Innocent Eye*. Oxford University Press.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics, 44*(2), 327-352.

Orlandi, N. (2018). Predictive perceptual systems. *Synthese, 195*(6), 2367-2386.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology/Revue canadienne de psychologie, 46*(3), 461-488.

O'Regan, J. K. (2011). *Why red doesn't sound like a bell: Understanding the feel of consciousness*. Oxford University Press.

Pezzulo, G., & Nolfi, S. (2019). Making the environment an informative place: A conceptual analysis of epistemic policies and sensorimotor coordination. *Entropy, 21*(4), 350. https://doi.org/10.3390/e21040350

Piccinini, G. (2022). Situated neural representations: Solving the problems of content. *Frontiers in Neurorobotics, 16*, 846979. https://doi.org/10.3389/fnbot.2022.846979

Piccinini, G. (2025). Computation in physical systems. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2025 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2025/entries/computation-physicalsystems/

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183*(3), 283-311.

Plato (1997). *Cratylus*. Hackett.

Powers, W. T. (1973). *Behavior: The control of perception*. Aldine Publishing.co.

Raimondi, A. (2021). Crane and the mark of the mental. *Analysis, 81*(4), 683-693.

Raja, V. (2024). The motifs of radical embodied neuroscience. *European Journal of Neuroscience*, 60(5), 4738-4755.

Ramsey W. (2007). *Representation Reconsidered*. Harvard University Press.

Ramsey, W., Stich, S. P., & Garon, J. (2013). Connectionism, elimiantivism, and the Future of Folk Psychology. In W. Ramsey, D. Rumehalrt & S. Stich (Eds.) *Philosophy and connectionist theory* (pp. 199-228). Routledge.

Ramsey, W. (2020). Defending representational realism. In J. Smortchkova, K. Dolega, T. Schlicht (eds). *What are mental representations?* (pp. 54-78). Oxford University Press.

Ramsey, W. (2025). The hard problem of content is neither. *Review of philosophy and psychology, 16*(1), 153-174.

Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. Adaptive behavior, 28(4), 225-239.

Revonsuo, A. (2000). The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and brain sciences, 23*(6), 877-901.

Robertson, I., & Kirchoff, M. D. (2019). Anticipatory action: Active inference in embodied cognitive activity. *Journal of Consciousness Studies, 27*(3-4), 38-68.

Rosenberg, A. (2019). *How history gets things wrong: The neuroscience of our addiction to stories*. The MIT Press.

Sartorius, N., Jablensky, A., Korten, A., Ernberg, G., Anker, M., Cooper, J. E., *et al.* (1986). Early manifestations and first contact incidence of schizophrenia in different cultures. *Psychological Medicine, 16*, 909-928

Sartre, J. P. (1971). *Sketch for a Theory of the Emotions*. Routledge.

Schnider, A. (2003). Spontaneous confabulation and the adaptation of thought to ongoing reality. *Nature Reviews Neuroscience, 4*(8), 662-671.

Schwitzgebel, E. (2002) A phenomenal, dispositional account of belief. *Noûs, 36*, 249-275.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review, 117*(2), 245-273.

Schwitzgebel, E. (2012). Introspection, what?, In D. Smithies & D. Stoljar (Eds.), *Introspection and Consciousness* (pp. 29-48). Oxford University Press

Schwitzgebel E. (2013), A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box. In N. Nottelman (ed.), *New Essays on Belief: Constitution, Content, and Structure* (pp. 75–99). Palgrave Macmillan.

Schwitzgebel, E. (2016). Phenomenal consciousness, defined and defended as innocently as I can manage. *Journal of Consciousness Studies, 23*(11-12), 224-235.

Schwitzgebel, E. (2024). *The weirdness of the world*. Princeton University Press.

Schwitzgebel, E. (2026). Dispositionalism, yay! Representationalism, Boo!. Belief as self-verifying fictions. In J. Lewis-Jong & E. Schwitzgebel (eds.) *The Nature of Belief* (pp. 286-308). Oxford University Press.

Searle, J. (1992). *The rediscovery of the mind*. The MIT Press.

Searle, J. (1995). John Searle Replies. In Dennett, D. & Searle, J. 'The Mystery of Consciousness: an exchange'. *The New York Review of Books*. https://www.nybooks.com/articles/1995/12/21/the-mystery-of-consciousness-an-exchange/

Segundo-Ortin, M., & Hutto, D. D. (2021). Similarity-based cognition: Radical enactivism meets cognitive neuroscience. *Synthese, 198*(Suppl 1), 5-23.

Selfridge, O. G. (1958). Pandemonium: A paradigm for learning in mechanisation of thought processes. Proceedings of a Symposium Held at the National Physical Laboratory (pp. 513–526). *HMSO*. https://gwern.net/doc/ai/nn/1959-selfridge.pdf

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive neuroscience, 5*(2), 97-118.

Seth, A. K. (2014b). The cybernetic Bayesian brain: From interoceptive inference to embodied selfhood. In T. Metzinger & J. M. Windt (Eds.), *Open MIND, 35* (pp. 1–24). *MIND* Group. https://doi.org/10.15502/9783958570108

Seth, A. (2021). *Being you: A new science of consciousness*. Penguin Publishing.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Vervet monkey alarm calls: Semantic communication in a free-ranging primate. *Animal Behaviour, 2*8(4), 1070-1094.

Sharot, T. (2011). *The Optimism Bias*. Penguin Publishing.

Shea, N. (2018). *Representation in cognitive science.* Oxford University Press.

Sganzerla, L., Hutto, D. D., & Kirchhoff, M. D. (2025). Beyond the extended mind: New arguments for extensive enactivism. *Synthese, 205*(3), 118.

Skyrms, B. (2010). Signals: *Evolution, learning, and information*. Oxford University Press.

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist, 96*(4), 539-560.

Sprevak, M., & Irvine, L. (2020). Eliminativism about consciousness. In U. Kriegel (ed). *The Oxford handbook of the philosophy of consciousness* (pp. 348-370). Oxford University Press.

Strawson, G. (2004). *Mental Reality (II ed)*. Oxford University Press

Stich, S. P. (1983). *From folk psychology to cognitive science: The case against belief*. The MIT Press.

Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in cognitive sciences, 24*(5), 346-348.

Thobani, I. (2024). A triviality worry for the internal model principle. *Synthese, 204*(1), 36.

Thompson, E., & Cosmelli, D. (2011). Brain in a vat or body in a world? Brainbound versus enactive views of experience. *Philosophical topics, 39*(1) 163-180.

Thomson, E., & Piccinini, G. (2018). Neural representations observed. *Minds and Machines, 28*(1), 191-235.

Toon, A. (2023). *Mind as Metaphor*. Oxford University Press.

Usher, M. (2001). A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind and Language, 16*(3), 311-334.

Vallée-Tourangeau, F., *et al.* (2016). Insight with hands and things. *Acta psychologica, 170*, 195-205.

Van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior, 29*(3), 315-329.

Van Es, T., & Myin, E. (2020). Predictive processing and representation: How less can be more. In D. Mendoça, M. Curdao & S. Grouveia (Eds.). *The philosophy and science of predictive processing* (p. 7-25). Bloomsbury.

Van Leeuwen, N. (2025). Three questions on imagery and perception: A comment on Nanay's Mental imagery. *Mind & Language, 40*(3), 325-332.

Von Frisch, K. (1993). *The dance language and orientation of bee*s. Harvard University Press.

Vilarroya, O. (2017). Neural representation. A survey-based analysis of the notion. *Frontiers in psychology, 8*, 1458. https://doi.org/10.3389/fpsyg.2017.01458

Williams, D. (2018a). Predictive minds and small-scale models: Kenneth Craik's contribution to cognitive science. *Philosophical Explorations, 21*(2), 245-263.

Williams, D. (2018b). Predictive processing and the representation wars. *Minds and Machines, 28*(1), 141-172.

Williams, J. R. (2019). *The Metaphysics of Representations*. Oxford University Press.

Wittgenstein, L. (2001). *Tractatus logico philosophicus*. Routledge.

Wittgenstein, L. (2009). *Philosophical investigations*. Wiley & Sons.

Zahnoun, F. (2023). Some inaccuracies about accuracy conditions. *Phenomenology and the Cognitive Sciences, 22*(2), 461-477.

Zahnoun, F. (2025). Information in the brain: From metaphor to truth. *Theoria, 91*(5), e70021. https://doi.org/10.1111/theo.70021

Marco Facchin is a postdoctoral fellow at the University of Antwerp (FWO grant number 1202824N). He works in the philosophy of mind and cognitive science, with a focus on mental content, extended cognition, and "radically embodied" views of cognition.

Address: Department of Philosophy and Letters, Faculty of Philosophy, Centre for Philosophical Psychology, Universiteit Antwerpen, Building R, Rodestraat 14, Antwerpen 2000, Belgium. E-mail: Marco.facchin.marco.facchin@gmail.com / Marco.facchin@uantwerp.be – ORCID: 0000-0001-5753-9873

Theoria, 2026, 41(3), 434-444
https://doi.org/10.1387/theoria.28294

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

The opacity of knowledge: An externalist's case for widespread KK-failure

(La opacidad del conocimiento: un argumento externista a favor del fracaso generalizado del principio KK)

Murali Ramachandran*
Thapar University
https://orcid.org/0000-0001-5611-518X

Keywords

Epistemic externalism
KK
Epistemic closure
Basic knowledge
Easy knowledge
Bootstrapping
Epistemic entitlement

ABSTRACT: An epistemic externalist should allow that whatever proposition about the world, *P*, one knows to be true, there will be some other fact one is not in a position to know, but which is required for the belief in *P* to count as knowledge. We contend that ordinary knowers are cognizant of such 'unknown knowledge-makers'. On this basis, we present a novel case —we distinguish it from rivals— for *widespread* KK-failure: that, quite generally, if one knows *P*, one will not be in a position to know that one knows it.

Palabras clave

Externismo epistémico
KK
Clausura epistémica
Conocimiento básico
Conocimiento fácil
Bootstrapping
Derecho epistémico

RESUMEN: Un externista epistémico debería admitir que, para cualquier proposición sobre el mundo, *P*, que se sepa verdadera, habrá algún otro hecho que uno no está en condiciones de conocer, pero que resulta necesario para que la creencia en *P* constituya conocimiento. Sostenemos que los sujetos cognoscentes ordinarios tienen noticia de tales 'hacedores de conocimiento desconocidos'. Sobre esta base, presentamos un argumento novedoso —que distinguimos de sus rivales— a favor del fracaso generalizado del principio KK: a saber, que, en términos muy generales, si alguien sabe que *P*, no estará en condiciones de saber que lo sabe.

Gako-hitzak

Esternalismo epistemikoa
KK
Itxiera epistemikoa
Oinarrizko ezagutza
Ezagutza erraza
Bootstrapping
Eskubide epistemikoa

LABURPENA: Esternalista epistemiko batek onartu beharko luke munduari buruzko edozein proposizio (*P*) egiazkoa dela jakinda ere beste egitateren bat egongo dela, norberak ezagutzerik ez daukana baina *P* proposizioarekiko sinesmena ezagutzatzat jotzeko ezinbestekoa dena. Baieztatzen dugu ezagutzaile arruntak «ezagutza-sortzaile ezezagun» horien kontziente direla. Hori oinarri hartuta, argudio berritzaile bat aurkezten dugu —beste argudio batzuetatik bereiziaz batera— KK printzipioaren porrot orokortuaren alde: alegia, eskuarki, batek *P* ezagutzen badu, bat horrek ez du jakiterik izango *P* ezagutzen duela.

* ***Correspondence to:*** Murali Ramachandran. Thapar School of Liberal Arts and Science, Thapar University, Bhadson Road, Patiala, Punjab 147004, India. – murali.ramachandran@thapar.edu – https://orcid.org/0000-0001-5611-518X

How to cite: Ramachandran, Murali (2026). «The opacity of knowledge: An externalist's case for widespread KK-failure»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 41(3), 434-444. (https://doi.org/10.1387/theoria.28294).

Received: 18 February, 2026; Final version: 06 September, 2026.
ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press

This work is licensed under a *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International* License

1. Preliminaries

The key thought behind *epistemic externalism* is that knowledge must meet some kind of 'reliability' condition, but also that one may not always be in a position to know that this condition is met:

> Knowledge has a *'reliability'* requirement of sorts, e.g. that one's knowing a proposition *P* to be true requires one's belief in *P* to be *reliably-acquired* (*or safe from error, indefeasibly justified, non-accidentally true,* etc.) [**REL**]
>
> One may know a proposition *P* to be true and yet *not* be in a position to know whether one's belief in *P is* reliably-acquired (or safe from error, indefeasibly justified, non-accidentally true, etc.)—Following Cohen (2002), we'll call such knowledge *basic knowledge*[1] [**BK**]

We wish to draw attention to an overlooked argument showing that these conditions lead quite straightforwardly to widespread violations of the KK-principle:

> If one knows a fact *P*, then one knows —or one is at least in a position to know— *that one knows P*. [$K(P) \rightarrow K\,K\,(P)$] [**KK**]

This view is certainly not new; but, we contend, earlier proponents held it on the basis of faulty reasoning. Consider e.g. the following passages from Williams (1991) and Dretske (2004) ('modest contextualism' in the Dretske quote amounts to epistemic externalism as we have characterised it):

> I agree that no item of knowledge guarantees full knowledge of its own reliability conditions: indeed, *this is just another way of stating* the externalist point that knowing does not guarantee knowing that one knows. (Williams, 1991, p. 348, our emphasis)
>
> It is interesting to note, however, that modest contextualism (and, hence, externalism) provides an illuminating explanation of why KK fails. It fails because factual knowledge, according to modest contextualism, depends for its existence on circumstances of which the knower may be entirely ignorant. So the knower can know that *P* without knowing (as required by KK) that he knows that. (Dretske, 2004, p. 176)

What Williams and Dretske are suggesting is that [KK] will fail when, and *because,* the knower is ignorant of pertinent factors —the satisfaction of certain reliability (or safety) conditions— *in virtue of which* she has that knowledge. The underlying thought is presumably this: such a knower, being ignorant of these pertinent factors, *ipso facto* cannot *verify* that these obtain; and, hence, she cannot know *that she knows.*

However, externalism does not demand as much. Consider e.g. a crude causal theory of knowledge: one knows *P* iff one believes *P* and this belief is caused by the fact that *P*.[2] What this account requires for someone to know that they know *P* is only that they believe that they know *P* and this belief is caused by their knowing *P*; they are *not* required to know or be in a position to establish that this causal relation obtains; moreover, they are *not* required to know or be in a position to establish that their belief in *P* itself is caused by the fact that *P*. More generally, Williams's and Dretske's reasoning fall short because the reliability (safety, indefeasibility, etc.) of one's *belief that one knows P* does not require that one is aware of what makes one's *belief that P* reliable (safe, etc.).

---

[1] Actually, Cohen (2002) introduces the term 'basic knowledge' for knowledge one purportedly has without being in a position to know that *its source is reliable* (Cohen, 2002, p. 310). But his central concern—the so-called *problem of easy knowledge* —arises on our broader conception of basic knowledge too. We make some brief remarks about that problem in § 3.4.

[2] Obviously, a feasible causal theory would need to rule out 'deviant' causal chains—but our crude account serves for the point we wish to make.

Of course, Williamson (2000, § 5.1) has famously argued against [KK], making appeal to epistemic *closure*[3]—-i.e. for our purposes:

> If one knows that P_1, ..., and that P_n, and at the same time also knows that these propositions jointly entail Q, then one knows, or is at least in a position to know, that Q [**EC**]

But his argument does not purport to *identify* any proposition which is supposedly known but not known to be known; rather, it aims to establish merely that such a proposition must exist.[4] To this extent, his refutation has little impact on everyday verdicts on second-order knowledge: we can allow that knowers *on the whole* know that they know. By contrast, if the rationale we present in this paper is accepted —which also assumes [EC]— pretty much *whatever* we know about the world, *we do not know that we know it!*

We shall lead up to the key strategy by way of considering two sorts of cases an externalist might categorize as cases of 'unacknowledged knowledge', as it were, and, so, count as examples of KK-failure. (But, just to be clear, these are not arguments we ourselves would wish to defend—rather, they shall serve to introduce, and highlight features of, the case we wish to draw attention to.)

2. *Unacknowledged knowledge*

2.1. *The cautious skeptic*

Suppose a cautious skeptic goes around asserting the following sorts of things: she *believes* she is sitting at her desk but doesn't *know* it; she *believes* the room is illuminated, but doesn't *know* it; she *believes* she is writing down her ideas, but doesn't *know* it; etc.. Suppose her perceptual capacities and logical acumen are fine. Her perceptual beliefs are true, and they change accordingly with changes in her environment.

Externalists will surely hold that she *does* know these things despite her denials—because she is appropriately and adequately sensitive to what is going on around her; her beliefs are reliably acquired and safe from error. However, they may allow that she does not *know that she knows* these facts, but only because she does not believe —and, given her misconceptions about knowledge, is not even *in position to believe*— that she knows these facts, and knowledge requires belief.

So, we have KK-failure here, but, arguably, the skeptic herself is at fault.

2.2. *Closure versus KK*

What is less widely appreciated is that the (many) alleged counterexamples to closure —of the sort famously advanced by Dretske (1970, 1971) and Nozick (1981)— may be recast as examples of 'unacknowledged knowledge' instead, thereby providing a case for KK-failure in the same scenarios.

We'll use Dretske's (1971) famous zebra example to illustrate the strategy. Suppose Meena is visiting the zoo, and sees a zebra in an enclosure. Dretske allows, *contra* [EC], that she may know the premises of the following (valid) argument but not be in a position to know its conclusion:

---

[3] The following quotation shows that Dretske himself is not appealing to closure—which he in any case famously rejects (Dretske, 1970, 1971): "Although modest contextualism rejects KK, it does not, taken by itself, deny closure" (Dretske, 2004, p. 176).

[4] To give a bit more detail —a full elaboration is unnecessary here— the argument has Mr Magoo seeing a tree in the distance; what Williamson's [KK]-*reductio* supposedly establishes is that for *some n*, Mr Magoo knows $[\neg T_n]$ *that the tree is not n inches tall,* but he does not know *that he knows* $[\neg T_n]$.

ZEBRA

(1) [Zeb] The animal in the pen is a zebra.

(2) If [Zeb], then [¬CPM] it is not a cleverly painted mule.

(C) Hence, [¬CPM].

According to Dretske, Meena's sighting the zebra suffices for her to know [Zeb], but not for her to know [¬CPM], that what she is looking at is not a cleverly painted mule. This accords with ordinary folk's intuitions: intuitively, to know *that,* she would have to examine the animal up close. And Meena herself would most likely deny that she knows [¬CPM].

Many find the intuitive cost of denying closure too high, however.[5] For example, if one takes ZEBRA to be an [EC]-counterexample, one must surely allow that Meena also knows [¬M] *that the animal in the pen is <u>not a mule</u>* despite not knowing [¬CPM] (that it is not a *cleverly painted <u>mule</u>*). We prefer to pay the costs associated with *endorsing* closure (and make partial payment in § 3.4.)

The pertinent point for present purposes is that the putative closure-counterexamples can be re-modeled as KK-counterexamples. To begin with, closure-endorsers will insist, on the basis of closure, that if Meena *does* know the ZEBRA-premises, then she is simply mistaken in thinking she does not know the conclusion. But, they can then explain her reticence in claiming knowledge of [¬CPM] by putting it down to her *not knowing that she knows* it: she does not know *that she knows* because she does not *believe* that she knows (and knowledge requires belief).[6]

And once it is granted that Meena knows, but does not know that she knows, [¬CPM] in the zoo scenario, our closure-endorsers can then argue that she does not know that she knows [Zeb] either—by consideration of the following argument:

K-ZEBRA

(1) Meena knows [Zeb].

(2) Meena knows (*If* [Zeb], *then* [¬CPM].)

(C) Hence, Meena knows [¬CPM].

Allowing that Meena herself recognizes the validity of [EC], she will recognize that K-ZEBRA is (trivially) valid. And we can surely allow that she knows the second premise—it is not unreasonable to assume that simple *a priori* truths can be known and be known to be known. So, since, as our closure- endorser has it, Meena is not in a position to know that she knows [¬CPM], i.e. (C), her knowledge of [EC] dictates that she does not know the first premise—i.e. we must deny that she knows that she knows [Zeb].[7]

---

5 Contemporary closure-deniers, e.g. Alspector-Kelly (2019) and Kelp (2021), avoid some pitfalls of the original Dretske (1971) and Nozick (1981) accounts—including e.g. Kripke's (2011) notorious Red Barn example. But equally counterintuitive results are inevitable—such as the one we mention here.

6 Skeptics and closure-deniers accommodate people's reluctance to affirm that they know the likes of [¬CPM] by maintaining that they indeed *do not know* it. (Epistemic contextualists take the same line while also attempting to preserve closure). To our mind, the reluctance (or awkwardness) people feel is just as plausibly accommodated by taking it to signify that they *do not believe that they know.* Williamson (2005, p. 227 ff.) likewise argues for KK-violations in such cases by appealing to the knowers' *unpreparedness to use the proposition for practical reasoning* in the pertinent context. We think it simpler to appeal directly to the fact that the subjects do not believe *they know.*

7 We thank a referee for forcing clarification of our reasoning here.

Finally, since putative closure-counterexamples such as ZEBRA can be devised on demand for just about any known proposition, the table-turning anti-KK strategy just outlined yields a case for ubiquitous KK-failure instead.

But this case against [KK] has two, not insignificant, drawbacks. First, it rests on ordinary folk having *misplaced* qualms about what they know: e.g. it is precisely, and only, because Meena mistakenly believes she does not know [¬CPM] that [KK] is violated in the ZEBRA scenario. One might well balk at a case against [KK] which pivots solely on how diffident or cautious subjects are. The strategy would appear to have the consequence that a dithering subject fluctuates between not knowing one minute, and knowing the next, that she knows. The second shortcoming is that one may rightly question whether demonstrating that an individual knows, but does not know that they know, a certain fact is in itself sufficient to refute [KK]: for, [KK] —our version of it at any rate (p. 2)— is violated only if the individual is *not even in a position to* know that fact. Our ditherer, one might think, was always *in a position* to know that she knew.

So, as we proclaimed at the outset, we are not minded to press the foregoing arguments against [KK]. However, the anti-KK strategy we are about to outline is more compelling; and it has the consequence that an *entirely faultless* subject may know she is seeing a zebra, for example, without being in a position to *know that she knows* she is. But, let us give fair warning: it will not suit all closure-endorsers.

3 The case from basic-knowledge bootstrapping

3.1 Strategy

The following case against [KK] rests on a specific type of basic-knowledge situation (call it a *BK-bootstrapping* situation):

— An individual, X, knows some empirical fact, *P* (A)

— *Q* is a proposition which, given the way X acquired her belief in *P*, has to be true in order for X's belief in *P* to count as knowledge (e.g. reliably-acquired, safe from error, indefeasible, non-accidentally true, etc.). Thus:

(X knows *P*) *only if Q* (B)

— X knows (B)—i.e. knows that: (she *knows P*) *only if Q* is true (C)

(For any such *Q*, we'll call [¬*Q*] a *knowledge-defeating possibility* for X's belief *P*).

— But, X is not in a position to know *Q* (D)

(A), (B) and (D) are what our initial characterization of epistemic externalism, i.e. [REL] and [BK], allows: X's knowledge of *P* is basic knowledge; (C) is what makes this a special case of basic knowledge—*special, but commonplace* we contend. For, we think that most people would acknowledge certain possibilities as *knowledge-defeating* possibilities for (just about) any belief they had about the world. (So, our view is that even ordinary folk are really externalists, advocates of basic knowledge, at heart—they just need the right prodding!)

Assuming [EC] (epistemic closure), it is easily demonstrated that [KK] is violated in any BK-bootstrapping situation: X does not know that she knows the pertinent *P*:

The BK-bootstrapping KK-reductio

1. X knows that she knows *P* from (A) and [KK], for reductio
2. X knows that [X knows *P only if Q* is true] from (C)
3. X knows *Q* from (1), (2) and [EC]

4. X doesn't know Q from (D)
5. $\perp$ (3) and (4) contradict
6. So, [KK] is refuted from reductio of (1)

So, [KK] is violated in any situation meeting conditions (A)-(D). We contend that such situations are ubiquitous.

3.2. Examples

Example 1: LIAR

Suppose Meena meets someone at a party who introduces himself as Dilip and, in the course of the ensuing conversation, reveals he is a physicist. She believes him. There is no deception here. We can allow, then, that Meena knows [P] *that Dilip is a physicist.* Further, given her grounds for [P], we can also assume Meena recognizes that she does not *know* [P] if [L] *Dilip is in fact a convincing compulsive liar.*[8] In our jargon, [L] is a knowledge-defeating possibility for Meena's belief [P]—and she knows it. But, she surely does not know [$\neg L$] on the basis of their brief conversation.

So, consider the following valid argument:

LIAR

(1) Meena knows [P] that Dilip is a physicist.
(2) (Meena knows [P]) only if [$\neg L$] Dilip is not a convincing compulsive liar.
(C) Hence, [$\neg L$] Dilip is not a convincing compulsive liar.

The case against [KK] is straightforward: we have allowed that Meena knows LIAR(2); if she also knew LIAR(1), closure ([EC]) would dictate that she is in a position to know what is entailed by them, namely, [¬L]—which she doesn't. And Meena cannot come to know [$\neg L$] by way of deducing it from the premises; for, in the given circumstances, the reasoning exemplifies a form of *bootstrapping:* in affirming LIAR(1), Meena has *already* taken the falsity of [L] *as granted,* as it were, albeit tacitly. So, Meena does not know LIAR(1), period. Hence, Meena knows, but does not know that she knows, [P]. As it happens, Dilip is a wonderful, honest person; but, she simply is not yet in a position to know that.

Before proceeding to the second example, let us spell out the *formal* differences between ZEBRA, K-ZEBRA and LIAR just to highlight what is distinctive about our bootstrapping case against KK, even though all three assume [EC], epistemic closure.[9]

— ZEBRA is simply an instance of *modus ponens:* [$A, A \rightarrow B \models B$]; the externalist case against KK we floated in this case granted that Meena knows the premises, and, hence, by [EC], that she knows the conclusion, B; it accommodates her own awkwardness about affirming that she knows B by proposing that she does not *believe* that she knows B, and, hence, since knowledge is factive, that she does not *know that she knows B*.

— *K*-ZEBRA basically *piggybacks* on the above stance; it is of the form [$K(A), K(A \rightarrow B) \models K(B)$]. This is an *instance* of [EC]; so, *assuming Meena herself knows* [EC], she knows that *K*-ZEBRA is trivially valid. Meena's knowledge of [EC] therefore dictates that if she knows *K*-ZEBRA's premises, she is at least in a position to know that she knows B, which

[8] We emphasize 'know ' to highlight that [P] itself is not at odds with [L]; it is Meena's *knowing* [P] which is. The same point applies to the arguments we'll be considering later.

[9] Again, we thank a referee for pressing the need for such clarification.

our externalist maintains is *not* the case. So, our externalist argues, one must deny that Meena knows both premises—the premises in this case being that she knows A and that she knows $[A \rightarrow B]$. Next step: since $[A \rightarrow B]$ in this specific case is an *a priori* truth, it is granted that she knows this, and also that she knows that she knows this. We are thereby compelled to deny that she knows the first premise—i.e. to deny that she *knows that she knows A*.

— LIAR is of the same logical form as ZEBRA (i.e. *modus ponens*), just with 'A' replaced by '$K(A)$': $[K(A), K(A) \rightarrow B \models B]$. But, unlike ZEBRA, this is a bootstrapping argument for Meena. So, our externlaists takes there to be no grounds for maintaining that Meena knows the conclusion, B, to be true. Hence, as with K-ZEBRA, we are led to deny that she knows both premises; and, again, since the second premise appears *a priori* true, we can grant that she knows it to be true. In which case, she does not know the first premise, i,e, she does not *know that she knows A*.

Let us return to our other examples of BK-bootstrapping.

Example 2: HOLIDAY

Here is different sort of example illustrating that the reliability (or safety) of a belief depends on far more than its *provenance.* Suppose Meena has just booked her flight for a holiday in the Bahamas a month hence. Presumably, she *may* then know [B] *that she will be in the Bahamas the following month*. But, as she herself will grant, she knows [B] only if [$\neg A$] *she does not come close to having a serious accident days before her departure*. However, she surely is not in a position to know the latter. So, we have another counterexample to [KK]: given that Meena knows that *her knowing* [B] is incompatible with [A], if she *knew that she knew* [B] as well, [EC] would dictate that she *is* in a position to knew [$\neg A$]—which she isn't; so Meena does not know that she knows [B].[10]

Example 3: MEETING

Our third candidate KK-counterexample is a variation of ZEBRA (§ 2.2). It suggests that beliefs have to be *indefeasibly justified,* not just reliably-acquired or safe from error, in order to qualify as knowledge. Consider the earlier zebra-scenario: Meena sees a zebra at the zoo and thereby comes to know [Zeb] *that the animal in the pen is zebra.* But, presumably, her belief would not have qualified as knowledge if e.g. [Meet] *there had been a recent meeting of the zoo's governors where the possibility of deploying mules cleverly painted to look like zebras was seriously considered*—(there is no need to delve into possible explanations for such a meeting). We allow that Meena knows that *her knowing* [Zeb] is incompatible with [Meet], but that she is in no position to know that the latter is false. As before, given that Meena knows [Zeb], and knows that her knowing this is incompatible with [Meet], she cannot also know that she knows [Zeb], for by closure this would dictate that she is in a position to know [¬Meet], which she isn't.

The diverse variety of illustrative examples we have just considered would seem to suggest that BK-bootstrapping arguments could be devised for just about any known facts about the world. If our case against [KK] is sound, KK-failure is *rife;* knowledge *in general is opaque:* there are hardly any known facts about the world which are known to be known.

3.3. A disparity in need of explanation

However, an obvious anomaly needs to be addressed: why should someone who maintains that Meena *knows* [¬CPM] (that the animal in the pen is not a cleverly painted mule) in the zoo scenario —*despite* her own protestations to the contrary one might add— *deny* that she knows [¬Meet] (that the zoo's governors have not met recently to consider the deployment of

[10] As a referee points out, some may question whether one can have knowledge of future events at all; so here is a variation on HOLIDAY involving knowledge of the *past.* Decades after Meena's trip to the Bahamas, her son, Raju, comes across her booking among her belongings; we can surely allow that Raju knows [B] *that Meena visited the Bahamas;* but Raju realizes that he *knows* [B] only if [$\neg C$] *Meena did not come close to canceling the trip in order to attend a wedding,* which he presently is not in a position to know. In this BK-bootstrapping situation, Raju knows, but does not know that he knows, [B].

cleverly painted mules) in the same setting? Our stance is apt to seem even more puzzling given that closure-endorsers must maintain that Meena *knows* —because it is entailed by [Zeb]— [¬BIV]: *that she is not a deceived brain in a vat in a zebra-less world.* So, why the disparity? Why not take Meena to know [¬Meet], even if she does not believe that she does?

Closure-endorsers will indeed have diverging responses here. We shall briefly demarcate what we take to be a feasible stance that underwrites our treating the likes of [¬CPM] and [¬BIV] differently from the conclusions of our BK-bootstrapping arguments in § 3.2.

4. Epistemic priority, transfer of warrant and collateral knowledge

We shall distinguish three kinds of closure-endorsers by their take on putative counterexamples such as ZEBRA.

Knowledge by deduction

(A) First, there are those who allow that Meena can *come to know (acquire knowledge that)* [¬CPM] holds by way of *competently deducing it* from [Zeb], which (by hypothesis) she knows on the basis of her sighting.

We would put Williamson (2000, p. 117) and Hawthorne (2013, p. 29) in this camp, since they appear to maintain that one can, for any valid inference, *acquire* knowledge of the conclusion by way of *competent deduction.* Other possible (A)-listers are Beebee (2001) and Markie (2005), who take the arguments from [Zeb] to [¬CPM] and [¬BIV] to *beg the question* against those who doubt or question the conclusions, but, apparently, nevertheless allow the 'inferential-route', as it were, to knowledge of those conclusions.

(A)-type closure-endorsers may resist our case against [KK], insisting that Meena can indeed come to know the conclusions of the arguments considered in § 3.2 by way of deduction—so long as she does know the pertinent premises. But to insist this just begs the question against our contention that these are *bootstrapping* arguments, which, intuitively, can give no succor to the conclusions in the specified circumstances.

Epistemic entitlement

(B) The second type of closure-endorser *denies* that Meena can acquire knowledge of [¬CPM] by way of deduction from the premises of ZEBRA, on the grounds that the *warrant* for the first premise *'fails to transfer'* to the conclusion; but, rather than *rejecting* closure on these grounds as Dretske (1970, 1971) does, they maintain that Meena *'already' knows* [¬CMP] to be true.

Davies (2004) and Wright (2004) are prominent (B)-listers; they attempt to preserve closure by maintaining that Meena is *'epistemically entitled'* to accept [¬CPM], that the animal she sighted is not a cleverly painted mule. In their view (albeit for different reasons), Meena's knowledge of [Zeb] *presupposes* (depends on, and is partly explained by, the fact) that she knows [¬CPM]—the latter item of knowledge is *epistemically prior,* so to speak.

But, to be blunt, we consider the appeal to 'epistemic entitlement' —and we include the invocation (of knowledge) of so-called *hinge* propositions here— to be wishful thinking. The fact that we would have to rule out knowledge *tout court* in a certain area of discourse, a 'region of inquiry' as it were, unless a proposition *p* is assumed to be true, does not, in our view, simply *entitle* us —as Wright (2004) appears to maintain— to take *p* to be true, let alone to take ourselves to *know* that is.

Collateral knowledge

Our misgivings about these approaches lead us to the third closure-preserving camp—though we fear we may be the only occupants as yet:

(C) This camp sides with (A)-listers in taking Meena's knowledge of [¬CPM] to be explained by —to stem ('flow') from— her knowledge of [Zeb] rather than, as (B)-listers have it, the other way around. But it sides with (B)-listers in denying that Meena can *acquire* knowledge of [¬CPM] by way of *deduction* from ZEBRA's premises. Rather, it takes her knowledge of [¬CPM] to emerge *in sync* with, at the same time as, her coming to know [Zeb]; but, this is not a case of *inferential*-knowledge.[11]

Crudely, we think of Meena's knowledge of [¬CPM] —and let us not forget [¬BIV] (that she is not a deceived brain in vat in a zebra-less world)— as *collateral knowledge:* she attains this knowledge *by virtue* of coming to know [Zeb]. To give a more mundane example: Meena's knowing [Zeb], given the concepts she possesses, dictates her knowing a host of other facts—e.g. that the animal in the pen is not a mule; that it is not a kangaroo; not a giraffe; etc. These latter items of knowledge are not, in our view, known by way of inference from [Zeb]; for, Meena's belief that the animal is zebra *comes with* the (albeit tacit) belief that it is not a mule, not a kangaroo, not a giraffe, and so on. She likewise acquires more esoteric knowledge of the likes of [¬CPM] and [¬BIV] in our view, since her understanding of what it is to be a zebra, given other concepts she possesses, makes it rationally impossible for her not (to be disposed) to believe these other propositions if she believes [Zeb].[12]

Of course, some story will have to be told about what makes her knowledge of [Zeb] different, epistemically prior or more basic; we hazard that it is distinguished by the fact that her belief in [Zeb], unlike her belief in [¬CPM] or [¬BIV], meets some sort of 'sensitivity' condition—but it would take us too far afield to detail or defend a specific account here.[13]

By contrast, coming back to the disparity 'in need of explanation', [¬Meet] is *not* a consequence of [Zeb]—as we have stressed, it is Meena's *knowing* [Zeb] which entails [¬Meet]. And, as we have argued, since her knowing [Zeb] *(a priori) depends* on the truth of [¬Meet], she cannot come to know that conclusion by way of deducing it from the premise that she knows [Zeb]—for, this would be a case of epistemic bootstrapping. So, we deny that Meena knows [¬Meet] because, in our eyes, there are *no satisfactory explanation* of how she attains this putative knowledge. More generally, we distinguish between the (obvious) consequences of 'worldly' propositions one knows, such as [¬CPM] and [¬BIV] —these comprise collateral knowledge one has— and the consequences of one's *knowing* such worldly propositions, which *ipso facto* are not known if they conform to our bootstrapping template in § 3.1 (p. 7).

It is incumbent on any externalist who still maintains that Meena really is in a position to know the conclusions of the bootstrapping arguments in § 3.2 —e.g. some steadfast A- or B- lister— to explicate the sorts of 'reliability' (safety, defeasibility) conditions they take to underpin Meena's knowledge of [Zeb], say, but which she is not in a position to know. Unless some such condition is advanced, we invite clarification on what makes them externalists at all—for, presumably, the existence of basic knowledge will not be key.

Conclusion

We can summarise our final position by comparison with the remarks from Williams (1991) we quoted earlier (p. 2). We wholly agree with him that "*no* item of knowledge guarantees *full* knowledge of its own reliability conditions" (Williams, 1991, p. 348, our emphases). But we would go further: not only are there reliability conditions which the

---

[11] We note a divergence from Steup (2019), who defends an analogous view: (roughly) that one attains knowledge of the *reliability* of the source of one's knowledge of *P as one comes to know P;* but we shy away from this view, since it is one's *knowing P,* not *P* itself, which entails the reliability the source. Nevertheless, he may count as a (C)-lister by our characterisation.

[12] It should be clear from these remarks that we definitely are not denying that one may *extend* one's knowledge by way of competent deduction in other sorts of cases.

[13] But, to give a heads up, we have in mind something along the lines of accounts proposed by Roush (2005, 2012); Ramachandran (2015).

knower is not in a position to know, we reckon the knower herself will be aware of some such condition. For example, we took it for granted that Meena herself would realize that her knowing [*P*], that Dilip is a physicist, in the LIAR scenario *depends on* [¬*L*], Dilip's not being a convincing complulsive liar.

Secondly, we agree with Williams's view that the ramification of there being reliability conditions which the knower is not a position to know, given externalism, is that "knowing *does not guarantee* knowing that one knows" (Williams, 1991, p. 348, our emphasis); but, again, we would go further: *if* our BK-argument is sound, the existence of such conditions guarantees that one never (at any rate, *hardly ever*) knows that one knows—as far as 'wordly' knowledge is concerned, at least.

We have presented a (neglected) case for widespread KK-failure from the perspective of a certain (neglected) externalist position. We submit that there is, at any rate, a case to answer. We close with the following question: *does it actually matter if we generally don't know that we know?* We are not sure what of significance is lost if the present case is accepted.

REFERENCES

Alspector-Kelly, M. (2019). *Against Knowledge Closure.* Cambridge University Press.

Beebee, H. (2001). Transfer of warrant, begging the question, and semantic externalism. *Philosophical Quarterly,* 51(204), 356–74.

Cohen, S. (2002). Basic knowledge and the problem of easy knowledge. *Philosophy and Phenomenological Research,* 65(2), 309–329.

Davies, M. (2004). Epistemic entitlement, warrant transmission and easy knowledge. *Aristotelian Society Supplementary Volume,* 78(1), 213–245.

Dretske, F. (1970). Epistemic operators. *Journal of Philosophy,* 67(24), 1007–1023.

Dretske, F. (1971). Conclusive reasons. *Australasian Journal of Philosophy,* 49(1), 1–22.

Dretske, F. (2004). Externalism and modest contextualism. *Erkenntnis,* 61(2-3), 173–186.

Hawthorne, J. (2013). The case for closure. In Steup, M. and Turri, J. (Eds.), *Contemporary Debates in Epistemology.* Wiley-Blackwell, 26-43.

Kelp, C. (2021). *Inquiry, Knowledge, and Understanding.* Oxford University Press.

Kripke, S. A. (2011). Nozick on knowledge. In Kripke, S. A. (Ed.), *Philosophical Troubles: Collected Papers,* Volume 1. Oxford University Press, 162-224.

Markie, P. J. (2005). Easy knowledge. *Philosophy and Phenomenological Research,* 70(2), 406–416.

Nozick, R. (1981). *Philosophical Explanations.* Harvard University Press.

Ramachandran, M. (2015). Knowing by way of tracking and epistemic closure. *Analysis,* 75(2), 217–223.

Roush, S. (2005). *Tracking Truth: Knowledge, Evidence, and Science.* Oxford University Press.

Roush, S. (2012). Sensitivity and closure. In Becker, K. and Black, T. (Eds.), *The Sensitivity Principle in Epistemology.* Cambridge University Press, 242-268.

Steup, M. (2019). Easy knowledge, circularity, and the puzzle of reliability knowledge. *Episteme,* 16(4), 453–473.

Williams, M. (1991). *Unnatural Doubts: Epistemological Realism and the Basis of Scepticism.* Blackwell.

Williamson, T. (2000). *Knowledge and its Limits.* Oxford University Press.

Williamson, T. (2005). Contextualism, subject-sensitive invariantism, and knowledge of knowledge. *Philosophical Quarterly,* 55(219), 213–235.

Wright, C. (2004). Warrant for nothing (and foundations for free)? *Aristotelian Society Supplementary Volume,* 78(1), 167–212.

Murali Ramachandran is Associate Professor of Philosophy at TSLAS, Thapar University. He specializes in epistemology, philosophical logic and metaphysics, and has published on counterpart theory, the theory of descriptions, counterfactual accounts of causation, paradoxes, and Williamson's anti-luminosity arguments.

Address: Thapar School of Liberal Arts and Science, Thapar University, Bhadson Road, Patiala, Punjab 147004, India. E-mail: murali.ramachandran@thapar.edu – ORCID ID: https://orcid.org/0000-0001-5611-518X

Theoria, 2026, 41(3), 445-460
https://doi.org/10.1387/theoria.27988

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

A CASE OF ENTITY REALISTS USING ABDUCTIVE CONFIRMATION

(Un caso de realistas de entidades que recurren a la confirmación abductiva)

Ken Aizawa*
Rutgers University
https://orcid.org/0000-0002-7143-4901

Keywords

Abduction
Action potential
Confirmation
Entity realism
Hacking
Hodgkin
Huxley

ABSTRACT: The paper has three objectives. First, it argues that Hodgkin, Huxley, and Katz were entity realists who confirmed the existence of specific entities by abduction. Second, it outlines a theory of this abductive reasoning. Third, it shows how the case is a counterexample to Hacking's contention that experimentalists do not accept abductive evidence. The paper is not a "mere counterexample" paper, as it outlines a positive account of how to understand Hodgkin, Huxley, and Katz's abductive reasoning and their entity realism. This single case invites the question of how many more such cases are to be found.

Palabras clave

Abducción
Potencial de acción
Confirmación
Realismo de entidades
Hacking
Hodgkin
Huxley

RESUMEN: Este artículo tiene tres objetivos. Primero, argumenta que Hodgkin, Huxley y Katz fueron realistas de entidades que confirmaron la existencia de entidades específicas mediante la abducción. Segundo, esboza una teoría de este razonamiento abductivo. Tercero, muestra cómo este caso constituye un contraejemplo a la afirmación de Hacking de que los experimentalistas no aceptan la evidencia abductiva. El artículo no es un mero contraejemplo, ya que presenta una explicación positiva de cómo comprender el razonamiento abductivo de Hodgkin, Huxley y Katz y su realismo de entidades. Este caso particular invita a preguntarse cuántos casos similares existen.

Gako-hitzak

Abdukzioa
Ekintza-potentziala
Berrespena
Entitate-errealismoa
Hacking
Hodgkin
Huxley

LABURPENA: Artikulu honek hiru helburu ditu. Lehenik eta behin, argudiatzen du Hodgkin, Huxley eta Katz entitate-errealistak zirela eta entitate jakin batzuen existentzia abdukzio bidez berretsi zutela. Bigarrenik, abdukzio bidezko arrazoibide horren teoria bat laburbiltzen du. Hirugarrenik, erakusten du kasu hau kontradibide bat dela Hacking-en harako baieztapen hartarako, alegia, esperimentalistek ez dutela abdukzio bidezko ebidentziarik onartzen. Artikulua ez da «kontradibide soil» bat; izan ere, Hodgkin, Huxley eta Katzen abdukzio bidezko arrazoibidea eta haien entitate-errealismoa ulertzeko moduari buruzko azalpen positibo bat ematen du. Kasu zehatz honek galdera egitera garamatza antzeko beste zenbat kasu aurki litezkeen.

* ***Correspondence to:*** Ken Aizawa. Department of Philosophy, Rutgers University, Newark. 404 Conklin Hall, 175 University Avenue, Newark, NJ 07102. USA – ken.aizawa@gmail.com – https://orcid.org/0000-0002-7143-4901

How to cite: Aizawa, Ken (2026). «A case of entity realists using abductive confirmation»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 41(3), 445-460. (https://doi.org/10.1387/theoria.27988).

Received: 29 October, 2025; Final version: 14 September, 2026.
ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press

This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Around the middle of the twentieth century, Alan Hodgkin, Andrew Huxley, and Bernard Katz came to believe that action potentials are associated with specific entities, namely, changes in axonal membrane permeability and ion fluxes through that membrane. They were realists about these entities. Through their experimental work, they supported these beliefs even though, given the technology of their day, they could not observe the changes in the membrane or the movements of sodium ions. Instead, they supported these beliefs using abductive reasons. They believed in changes in membrane permeability and ion fluxes, because those unobserved changes and fluxes explained the changes in the electrical potential across the axonal membrane. In brief, Hodgkin, Huxley, and Katz were entity realists who used abductive confirmation.

The primary objective of this paper is to draw attention to this case. Many recent discussions of entity realism do not consider such cases.[1] Further, the existence of one case raises the question of whether there are others. A secondary objective is to outline more explicit theories of Hodgkin, Huxley, and Katz's thinking. These physiologists obviously do not explicitly claim that they are entity realists who have confirmed the existence of unobserved entities using abduction. Historians and philosophers of science must, therefore, draw attention to features of Hodgkin, Huxley, and Katz's experiments, their reasoning, and their commentary on their work, to show that they were implicitly entity realists who used abductive confirmation. A tertiary objective is to draw attention to the implications of this case for Ian Hacking's views regarding entity realism and confirmation. Hacking suggests that scientists do not confirm the existence of unobservable entities using abductive reasons. Instead, they confirm them by manipulating them to reason about other entities.

This paper is organized according to the foregoing three objectives. Section 1 will briefly review some of Hodgkin, Huxley, and Katz's experimental work from 1939 to 1949, making the case that they believed in specific unobserved entities for abductive reasons. Section 2 will outline a theory of abduction and entity realism that characterizes what is implicit in Hodgkin, Huxley, and Katz's work. This means that Hodgkin, Huxley, and Katz's work will serve as more than a "mere counterexample" to theses by Hacking. Their work motivates a theory of scientific ontology and methodology. Finally, section 3 will relate the Hodgkin-Huxley-Katz case to Hacking's views on entity realism and confirmation.[2]

1. Hodgkin, Huxley, and Katz on the Action Potential

The work discussed here is Hodgkin, Huxley, and Katz's research on the biophysical basis of the neuronal action potential published between 1939 and 1949. Section 1.1 will review an experiment by Hodgkin and Huxley, which they took to support the conclusion that there is an "overshoot" in the action potential, thereby challenging Julius Bernstein's "membrane hypothesis". Section 1.2 will review an experiment by Hodgkin and Katz, which they took to support the conclusion that sodium influx is responsible for the initial inward current of the action potential. These two experiments form part of the backdrop to the series of five papers by Hodgkin, Huxley, and Katz published in 1952. These latter papers led to Hodgkin and Huxley sharing the 1963 Nobel Prize in Medicine or Physiology with John Eccles.

1.1. Hodgkin and Huxley 1939-1945

In 1939, the accepted view of the neuronal action potential was due to Bernstein. On his view, at rest, the interior of the axon is more negative than the exterior. There is, Bernstein proposed, an imbalance of ions across the membrane.

---

1 See Eronen (2019), Khalili (2023, 2024), Khalili & Kazemi (2025), Nanay (2019, 2022), and Patton (2025).

2 Aizawa & Headley (2022, 2025) broaches the thesis that Hodgkin and Huxley used abductive reasoning. Aizawa (2025) adds to this a more elaborate theory of how Hodgkin and Huxley connected abduction and confirmation. Aizawa (forthcoming) shows how the theory of abduction applies to Jacobus van 't Hoff's theory that carbon atoms have a tetrahedral structure. The present work indicates how this earlier work bears on entity realism and Hacking's views on the confirmation of unobserved entities, a topic not discussed in earlier work.

With a sufficiently large depolarization of the axon, however, he further proposed that there is a temporary breakdown in the axonal membrane's resistance to the flow of ions. As a result, ions flow through the membrane down their concentration gradients and approach an equilibrium. In other words, one of the physical consequences of this proposal was that careful measurements of the action potential would yield a curve like that shown in the left half of Figure 1, wherein the action potential would "peak" near zero.

In 1939, Hodgkin and Huxley managed to insert microelectrodes into the recently discovered squid giant axons and obtain more precise measurements of the action potential.[3] These measurements revealed a dramatic "overshoot" of the action potential, as shown in the right half of Figure 1. Note that another common way of describing this result was to say that the action potential exceeded the resting potential. Still another way of describing the result was to note that the potential reverses from the resting state, in which the axon is more negative on the inside than on the outside, to an active state in which the axon is more positive on the inside than on the outside. It is important to note these alternative descriptions since Hodgkin, Huxley, and Katz used them in quotes provided below. Although Hodgkin and Huxley did not comment upon it in 1939, they appreciated that this experimental result challenged the membrane hypothesis.[4]

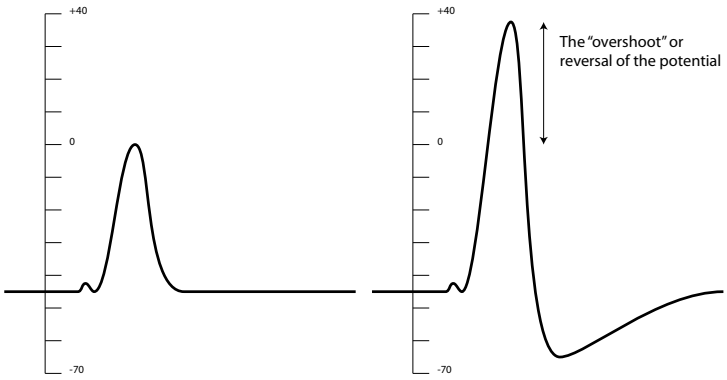


Figure 1

Action potential expected on Bernstein's hypothesis (left)
and action potential as measured in Hodgkin & Huxley, 1939 (right)

In 1945, following the interruption of World War II, Hodgkin and Huxley returned to measuring the action potential, further confirming its overshoot. In the introduction to their paper, Hodgkin and Huxley comment,

> A method of inserting microelectrodes into squid axons was developed, and direct measurements of action and resting potentials were obtained in the summer of 1939 (Hodgkin & Huxley, 1939). At about the same time, Curtis & Cole (1940) evolved a rather different technique for performing the same experiment, and in a subsequent paper (Curtis & Cole, 1942) concluded that the action potential exceeds the resting potential by a large amount. The present paper provides an independent confirmation of this conclusion and extends it to crustacean nerve fibres. (Hodgkin & Huxley, 1945, p. 176)

Note Hodgkin and Huxley's claim to be providing confirmation. In performing further experiments, Hodgkin and Huxley concluded that "The difference between action and resting potentials is too large to be explained by a liquid

[3] See Hodgkin & Huxley (1939).
[4] See Hodgkin (1994).

junction potential between the axoplasm and the microelectrode. It must be due to a genuine reversal of potential at the surface membrane of an active section of nerve" (Hodgkin & Huxley, 1945, p. 194).[5] Note here the invocation of explanation in connection with disconfirmation of the hypothesis that the difference is due to a liquid junction potential.

1.2. Hodgkin and Katz, 1946-1949

In the fall of 1946, Katz told Hodgkin about a forthcoming paper of his reporting that crab axons were inexcitable in sodium-free media. This inspired Hodgkin and Katz to examine the role of extracellular sodium in squid giant axons during the summer of 1947. The results of this work were published in (Hodgkin & Katz, 1949) and presented the first formulation of the Hodgkin-Huxley-Katz theory of the action potential.

The theory in (Hodgkin & Katz, 1949), in outline, proposed that an action potential has two phases. See Figure 2. In the first phase, lasting about a millisecond, the stimulation of the axon quickly increases the sodium permeability of the membrane, allowing excess extracellular sodium to enter the axon. This drives the initial "upward" shift and overshoot/reversal of the action potential. In the second phase, the sodium permeability decreases and the potassium permeability increases. With the increase in potassium permeability, potassium flows down its concentration gradient, exiting the cell. This drives the later "downward" shift of the action potential. As Hodgkin and Katz observed, "A simple consequence of the hypothesis is that the magnitude of the action potential should be greatly influenced by the concentration of sodium in the external fluid" (Hodgkin & Katz, 1949, p. 38). This is what they proposed to investigate. The following discussion will, therefore, largely set aside the role of potassium efflux and the change in potassium membrane permeability.

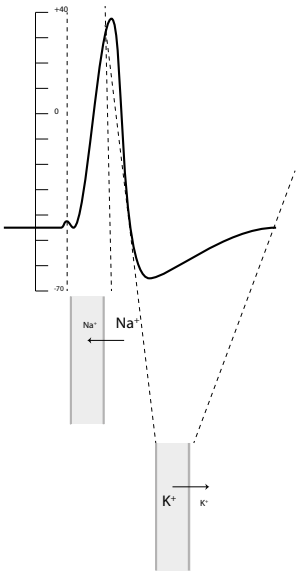


Figure 2

The 1949 Hodgkin-Huxley-Katz theory of the basis of the action potential.
NB, the scale here follows the convention used in (Hodgkin & Katz, 1949), rather than contemporary conventions

[5] A liquid junction potential is a potential difference that arises at the interface of two solutions of different ionic compositions. In the case at hand, there was one solution in the axoplasm and another in the microelectrode. Although Hodgkin and Huxley did not know the exact constituents of the axoplasm and hence did not know the precise value of the junction potential, they claimed that their estimates were smaller than could be used to explain the measured reversal potential.

(Hodgkin & Katz, 1949) reports on several experiments; however, for present purposes, it will suffice to review their second one. In this experiment, Hodgkin and Katz depolarized an axon in three different conditions: 1) the standard seawater solution (containing sodium), 2) a reduced sodium solution (33% sea water, 67% isotonic dextrose), and then again 3) the standard seawater solution (containing sodium). For ease of reference, call this the "three-curve experiment". This experiment is representative of the other experiments in (Hodgkin & Katz, 1949) insofar as it examines action potentials in different media over time.[6]

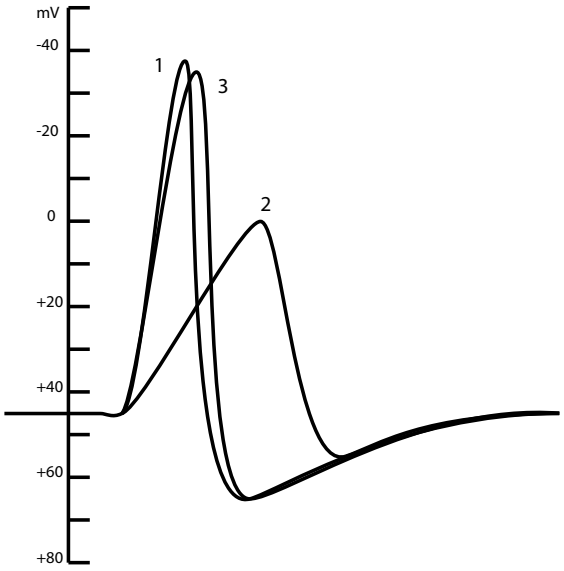


Figure 3

Action of sodium-deficient solutions on the resting and action potential. 1, response in seawater; 2, after 16 min. in 33% seawater, 67% isotonic dextrose; 3, 13 min. after reapplication of sea water. (Redrawn from (Hodgkin & Katz, 1949, p. 48, Figure 4.1a)).

For present purposes, the principal observations from this experiment are 1) that, in the reduced-sodium medium, the action potential has a lower peak (i.e., the peak is at 0 mV rather than near -40mV) and 2) that the time to reach this lower peak is delayed (i.e., the peak is shifted to the right). To appreciate Hodgkin and Katz's thinking about their experimental work, consider some of the commentary they offered on it. In the second paragraph of their paper, they write, "According to the membrane theory excitation leads to a loss of the normal selectively permeable character of the membrane, with the result that the resting potential falls towards zero during activity. This aspect of the theory is at variance with modern observations and must be rejected" (Hodgkin & Katz, 1949, p. 38). This is the conclusion based on the Hodgkin-Huxley work from 1939-1945, namely, that Bernstein's membrane theory is incorrect. It is not based on the three-curve experiment described in the preceding paragraph. Next, they comment: "However, a large reversal of membrane potential can be obtained if it is assumed that the active membrane does not lose its selective permeability, but reverses the resting conditions by becoming highly and specifically permeable to sodium. ... The essential point in the hypothesis is that the permeability to sodium must rise to a value which is much higher than that to potassium and chloride" (Hodgkin & Katz, 1949, p. 38). The "reversal of membrane potential" refers to the overshoot reported in (Hodgkin & Huxley, 1939), shown in the right half of Figure 1. The comment points to the existence of a change in

[6] Because of space limitations, it is left to the reader to verify that the three-curve experiment is representative of the other experiments in Hodgkin & Katz (1949), that the three-curve experiment is not cherry-picked.

membrane permeability. This is reinforced by the following: "The hypothesis ... presupposes the existence of a special mechanism which allows sodium ions to traverse the active membrane" (Hodgkin & Katz, 1949, p. 38). Finally, in the summary of their paper, Hodgkin and Katz write,

> The reversal of membrane potential during the action potential can be explained if it is assumed that the permeability conditions of the membrane in the active state are the reverse of those in the resting state. The resting membrane is taken to be more permeable to potassium than sodium, and the active membrane more permeable to sodium than to potassium. A reversed membrane potential can arise in a system of this kind if the concentration of sodium in the external solution is greater than that in the axoplasm.
>
> This hypothesis is supported by the following observations made with a microelectrode in squid giant axons:
>
> 1. The action potential is abolished by sodium-free solutions, but returns to its former value when sea water is replaced.
>
> 2. Dilution of sea water with isotonic dextrose produces a slight increase in resting potential, but a large and reversible decrease in the height of the action potential. (Hodgkin & Katz, 1949, p. 73)

Note, in the first sentence, that Hodgkin and Huxley take themselves to have an explanation of the reversal potential. In this text, 1. refers to Hodgkin and Huxley's first experiment (which has not been discussed here), whereas 2. refers to the three-curve experiment. (The present discussion does not address the slight increase in resting potential.)

This commentary indicates that, based on the experiments reported in 1939 and 1945, Hodgkin and Huxley drew three principal conclusions. First, Bernstein's theory of the action potential was not correct. Second, there exists an action potential (which was never seriously in doubt at that time) with an overshoot/reversal in potential (which they take to have been established by 1945). Third, there exists a change in membrane permeability to sodium and the movements of sodium ions into the cell during the initial phase of the action potential.

2. Hodgkin, Huxley, and Katz's Abduction and Entity Realism

The preceding discussion of Hodgkin, Huxley, and Katz's experimental work was meant to present the science in terms that are relatively neutral with respect to the philosophical theories to be presented in this section. This section presents theories of abduction and entity realism that are meant to accurately describe Hodgkin, Huxley, and Katz's experimental work. Section 2.1 will introduce the theory of abductive confirmation and disconfirmation whereas Section 2.2 will introduce the theory of entity realism. As noted in the introduction, this combination of scientific views has not been discussed in the literature on entity realism.

2.1. Hodgkin, Huxley, and Katz's Abduction

To start, consider some general features of how the term "abduction" is understood here. The proposal is that abduction is inferring some hypothesis H because H explains some evidence E. This interpretation does not embrace some of the more common interpretations of "abduction".

First, one should distinguish between what is here meant by "abduction" and what Gilbert Harman meant by "inference to the best explanation" (IBE).[7] For Harman, inference to the best explanation is *warranted* abduction, and ab-

[7] There are, of course, other versions of IBE, such as those in Lipton (2004), Bird (2022), and Davey (forthcoming), to name but a few. Present space limitations prevent an examination of them.

ductive inferences are warranted when one is able to reject all the rival hypotheses.[8] Moreover, for Harman, there is an open-ended list of considerations that one might use to reject rival hypotheses. These considerations include simplicity, plausibility, breadth, and so forth. To put matters in another way, Harman's IBE is meant to answer two questions. Question 1: What reason is there to believe H? Answer 1: H explains E.[9] Question 2: Why think that H is the correct explanation, rather than H'? Answer 2: There could be many reasons, such as H is simpler than H', H has a broader scope than H', and so forth. By contrast, and consistent with Harman's view, what is here meant by "abduction" is merely part of this story. "Abduction", as used here, just implicates H's explaining E. To put this matter another way, abduction, as understood here, does not compare rival hypotheses in the way that IBE does.[10]

Second, "abduction" is not meant to be the same as "Peircean abduction"; it is not meant to be used as Peirce is commonly taken to have used it.[11] According to a common reading of Peirce, scientists use abduction to introduce (pursuit-worthy) new hypotheses that are subsequently confirmed using hypothetico-deductive methods.[12] The proposal in this paper is that Hodgkin, Huxley, and Katz use abduction to confirm hypotheses in the sense that abduction gives a reason to believe some hypothesis, even when those reasons may not amount to *compelling reasons*. They are reasons that, to critics, may not justify belief in hypothesis H. Indeed, even a scientist who proposes that H explains E may not take that reason to be strong enough to justify belief.

Third, and perhaps most germane for the present discussion, "abduction" is not used as Hacking uses it in (Hacking, 1983). Although Hacking has very little to say about abduction, he does say this: "The idea of 'inference to the best explanation' is quite old. C.S. Peirce (1839–1914) called it the method of hypothesis, or abduction. The idea is that if, confronted by some phenomenon, you find one explanation (perhaps with some initial plausibility) that makes sense of what is otherwise inexplicable, then you should conclude that the explanation is probably right" (Hacking, 1983, p. 52). So, Hacking identifies abduction and IBE, whereas that is not the proposal here. Further, Hacking expresses this as a normative claim (presumably for scientists, and not just philosophers) about what one should conclude. In this paper, abduction is not understood that way. The concept of "abduction" is intended here to describe some instances of scientific reasoning without any normative commitment. It is used for sociology of science.

Hodgkin and Katz's reasoning was not just any old type of abductive reasoning. It was not like thinking that a deer walked through the snow here, because that explains the marks in the snow. That would be causal abduction, because the deer walking through the snow provides a causal explanation of the marks in the snow. Instead, Hodgkin and Katz's reasoning was compositional: the change in membrane potential during the course of the action potential was compositionally explained by the change in the number of charges inside and outside of the axon. The potential was composed, or made up of, the balance of internal and external charges.[13] A theory of abduction involves a theory of explanation. So, a theory of causal abduction involves a theory of causal explanation, and a theory of compositional abduction involves a theory of compositional explanation.[14]

The theory of compositional explanation proposes that, in the explanations Hodgkin, Huxley, and Katz invoked, the explanans entities ontologically determine the explanandum entities.[15] Thus, for Hodgkin and Katz, the explanation of the action potential by the movements of ions across the membrane presupposes that the movements of ions across the

---

8 N.B., Harman does not explain what he means by "warrant".

9 In many formulations of IBE, the proposal is that H potentially explains E. That refinement does not change the principal point here.

10 See Dellsén (2024).

11 Hacking (1989) takes abduction to introduce pursuit-worthy hypotheses.

12 Cabrera (2023) provides an overview of this line of thought, along with multiple references.

13 Some New Mechanists might count this as an instance of mechanistic constitutive explanation. See Craver (2007).

14 For a precursor to this taxonomy, see Psillos (2002), Schurz (2008), and Schupbach (2017).

15 This idea borrows from Jaegwon Kim's "realist" conception of explanation, according to which an explanation is a representation of an ontological dependence relation. See Kim (1994).

membrane ontologically determines the action potential. Apply this idea to the three-curve experiment. The proposal is that Hodgkin and Katz believed that

> For curve 1, x_1 φ_1-ing, x_2 φ_2-ing, ..., x_n φ_n-ing compositionally explain S ψ-ing
> For curve 2, x_1 φ_1-ing, x_2 φ_2-ing, ..., x_m φ_m-ing compositionally explain S ψ^*-ing
> For curve 3, x_1 φ_1-ing, x_2 φ_2-ing, ..., x_l φ_l-ing compositionally explain S ψ^{**}-ing,

and this means that Hodgkin and Katz were also committed to the view that

> For curve 1, x_1 φ_1-ing, x_2 φ_2-ing, ..., x_n φ_n-ing ontologically determine S ψ-ing
> For curve 2, x_1 φ_1-ing, x_2 φ_2-ing, ..., x_m φ_m-ing ontologically determine S ψ^*-ing
> For curve 3, x_1 φ_1-ing, x_2 φ_2-ing, ..., x_l φ_l-ing ontologically determine S ψ^{**}-ing.

Notice that, in the experiments under examination, the action potentials in these curves—these explananda—are spatio-temporal particulars. The images on the oscilloscope (redrawn in Figure 3) were taken by Hodgkin, Huxley, and Katz to indicate action potentials at specific times and places. Further, in the experiments under examination, the ions moving across the membrane—the explanantia—are also spatio-temporal particulars. The focus on the explanation of spatio-temporal particulars by other spatio-temporal particulars means that Hodgkin, Huxley, and Katz's explanations are *singular* compositional explanations. These singular compositional explanations are the basis for singular compositional abductive inferences in just the way that the (singular) causal explanation of the marks in the snow is the basis for a (singular) causal abductive inference.

The foregoing account of explanation and abduction helps articulate the rationale in the theory of confirmation that Hodgkin, Huxley, and Katz had in mind. Hodgkin and Katz believed that the early influx of sodium ions compositionally explains the experimentally measured initial rise of the action potential. But that means that they took the early influx of the sodium to ontologically determine the initial rise of the action potential. They also had the further thought that an influx of sodium could not explain the initial rise—it could not ontologically determine the initial rise—if it did not exist. Hodgkin and Katz's presupposition is that the compositional explanation that they offered presupposes the existence of the entities referred to in the sentences stating the explanans. Their explanation, thus, presupposed the existence of the entities in the explanans. It presupposes that there is an influx of sodium. Philosophers of science need not hold that all explanations work like this one. Instead, one need merely suppose that Hodgkin and Katz's explanation worked this way.[16]

The foregoing introduces a theory of confirmation, but there is also a theory of disconfirmation. In thinking about Bernstein's membrane theory, Hodgkin and Huxley supposed that, if Bernstein's theory were correct, there would be no overshoot. There was an overshoot, so there is reason to think Bernstein's theory is not correct. Consider the point more abstractly. In thinking about Bernstein's membrane theory, Hodgkin and Huxley supposed that if Bernstein's theory were correct, then some entities x_1 φ_1-ing, x_2 φ_2-ing, ..., x_n φ_n-ing ontologically determine S ψ-ing. But, when they performed an experiment to measure S ψ-ing, they did not find S ψ-ing. Instead, they found S ψ^*-ing. But, if x_1 φ_1-ing, x_2 φ_2-ing, ..., x_n φ_n-ing ontologically determine S ψ-ing, rather than S ψ^*-ing, then x_1 φ_1-ing, x_2 φ_2-ing, ..., x_n φ_n-ing must not be in place. In other words, there is reason to believe that the configuration of entities that Bernstein supposed to be in play in the action potential were not in play in the action potential.

---

[16] The existence of a change in membrane permeability is brought into the picture, since Hodgkin and Katz must postulate this as a cause of the influx of sodium ions. The change in membrane permeability to sodium, along with the sodium concentration gradient across the membrane, cause the influx of sodium ions. If there were no change in membrane permeability to sodium ions, the sodium ions would not enter the axon.

Section 1 reviewed an example of Hodgkin, Huxley, and Katz providing reasons against one theory of the action potential and reasons in favor of another. Section 2.1 has now sketched an abductive theory of confirmation and disconfirmation that attempts to characterize these reasons. It attempts to characterize the theory of confirmation and disconfirmation that Hodgkin, Huxley, and Katz had in mind. Section 2.2 will now articulate the entity realism that is implicit in their work.

2.2. Hodgkin, Huxley, and Katz's Entity Realism

In the Hodgkin, Huxley, and Katz experiments, the explanantia entities—the things singular compositional abductive confirmation is meant to confirm—are spatiotemporal individuals, rather than (natural) kinds. Thus, for Hodgkin, Huxley, and Katz to say that sodium ion influxes exist or are real is for them to say that there are individual sodium ion influxes that exist at specific times and places. This is Hodgkin, Huxley, and Katz's entity realism. This is to say nothing about natural kinds, assuming, of course, that a natural kind is something more than just the spatio-temporally particular individuals themselves. The rationale for Hodgkin, Huxley, and Katz's limited ontological commitment is that abductive inferences from experimental work might confirm the existence of individual sodium ion fluxes at specific places and times, but that further generalizations from those spatiotemporal particulars will likely be needed in order to justify the existence of kinds. Moreover, one would want a philosophical theory of what a kind is.

Further fleshing out this entity realism, "entities" will not be limited to spatio-temporally localized individuals, such as electrons, axonal membranes, or sodium ions. "Entities" will include activity instances, such as changes in membrane permeability, and property instances, such as the charges of individual sodium ions. Entities in these distinct ontological categories are needed to describe Hodgkin, Huxley, and Katz's work accurately. Hodgkin, Huxley, and Katz believed that individuals with property instances engage in activity instances. Their interpretation of the experiment presupposes that individual, positively charged sodium ions move from the extracellular space into the axoplasm.

This ontological framework forestalls an objection Alan Musgrave directed at Hacking's entity realism many years ago: "To believe in an entity, while believing nothing further about that entity, is to believe nothing, I tell you that I believe in hobgoblins (believe that the term 'hobgoblin' is a referring term). So, you reply, you think there are little people who creep into houses at night and do the housework. Oh no, say I, I do not believe that hobgoblins do that. Actually, I have no beliefs at all about what hobgoblins do or what they are like. I just believe in them" (Musgrave, 1996, p. 20). The entity realism here ascribed to Hodgkin, Huxley, and Katz is not the bare "property-free"/"activity free" entity realism Musgrave criticizes. Hodgkin, Huxley, and Katz did not merely believe in the existence of sodium ions. They also believed that each ion has a unit positive charge (a property instance) and that each passes through the axonal membrane (an activity instance).

Consider, now, a contrast between theory realism and entity realism. A theory realist aims to have a true theory of the world. An entity realist, however, aims to refer to things (spatio-temporal particulars) in the world. This formulation is at least similar to many formulations in the philosophical literature. It is not, however, exactly the one Hacking develops, but it faces the same challenge his formulation does: How is it that entity realism and theory realism do not collapse into each other? As Hacking makes the point, "The two realisms may seem identical. If you believe a theory, do you not believe in the existence of the entities it speaks about? If you believe in some entities, must you not describe them in some theoretical way that you accept? This seeming identity is illusory. The vast majority of experimental physicists are realists about entities but not about theories" (Hacking, 1982, p. 72). Unfortunately, neither Hacking's reply in the last sentence nor his later comments specify how one is to avoid entity realism and theory realism collapsing into each other.

A more positive approach to this question should specify something about what one means by a "theory". Years ago, for example, Musgrave proposed that "any collection of beliefs about X's counts as an X-theory" (Musgrave, 1996, p. 21). Insofar as this is a stipulative use of "theory", then a commitment to the sort of entity realism just described will bring

with it a commitment to theory realism. Nevertheless, one might still defend a distinction between entity realism and theory realism by proposing a use of "theory" that respects more common features of actual scientific and philosophical usage. If one is interested in capturing a more common usage for "theory", one will likely reject the set of beliefs,

> An electron weighs less than a ton.
>
> An electron is not a banana.
>
> John loves electrons.

as a theory of electrons. This would count as a "Musgrave theory", but presumably one wants a more restrictive account of what a theory is.

One quite minimal feature of more common uses of "theory" is that scientific theories, at the very least, aim to state general features of the world. They sometimes express or entail generalizations about the world.[17] Thus, the historian and philosopher of science can distinguish ontological commitments to spatiotemporal particulars, as found in the proposed formulation of entity realism, and ontological commitments made by general statements, empirical generalizations (or laws), and hence theories. This is, obviously, not a philosophical theory of what a theory is. Instead, it is merely a means to distinguish between scientists trying to develop a true theory and scientists trying to embrace referring expressions. There may be other, better ways to distinguish between entity realism and theory realism, but this one meshes well with the theory of confirmation attributed to Hodgkin, Huxley, and Katz.

Consider now how the historian and philosopher of science may ascertain what scientists are trying to refer to. The proposal is that some scientists do this by trying to confirm hypotheses with terms that refer. Hodgkin, Huxley, and Katz did this when trying to confirm that movements of sodium and potassium ions generate action potentials. In other words, in attempting to confirm the hypothesis that there are changes in membrane permeability during an action potential, Hodgkin, Huxley, and Katz were referring to changes in membrane permeability. This is not to assume, in general, that whenever a scientist tries to confirm some hypothesis H containing a term T, the scientist must presuppose that T refers. That would be a hasty generalization from a single case. The weaker assumption is that when Hodgkin, Huxley, and Katz tried to confirm the influx of sodium during the action potential, they were trying to refer to an influx of sodium.

3. Hacking on Experimentation and Scientific Realism

Having presented some of Hodgkin, Huxley, and Katz's experimental work (Section 1) and some of their ontological and methodological presuppositions (Section 2), this section will draw attention to some implications for Hacking's views regarding entity realism and confirmation. At the top of the list of the implications is that, contrary to Hacking's contention, some scientists do use abductive reasoning to confirm the existence of unobserved entities.

To begin, consider the scope and ambitions of Hacking's account. Consider, first, the proposed scope of Hacking's entity realism. Sometimes he appears to describe the whole of science. At other times, he seems to limit his proposal to physics or experimental physics. So, for example, he writes, "Different sciences at different times exhibit different relationships between 'theory' and 'experiment'. One chief role of experiment is the creation of phenomena. Experimenters bring into being phenomena that do not naturally exist in a pure state. These phenomena are the touchstones of physics, the keys to nature, and the source of much modern technology" (Hacking, 1982, pp. 71-72). Cf. Hacking (1984, p. 155). Later, he claims that "The vast majority of experimental physicists are realists about entities but not about the-

[17] This view is a weaker version of the assumption that a theory is a set of laws or a set of law statements. It is in keeping with Nancy Cartwright's idea of theory as organizing scientific knowledge. See Cartwright (1983).

ories. Some are, no doubt, realists about theories too, but that is less central to their concerns. Experimenters are often realists about the entities that they investigate, but they do not have to be so" (Hacking, 1982, p. 72). Cf. Hacking (1984, p. 155, italics in original). For present purposes, the assumption is that Hacking's theory is meant to cover the case of mid-twentieth-century experimental physiology. After all, what non-ad hoc reason is there to limit the scope of Hacking's theory to experimental physics?

An important component of Hacking's discussion of experimentation and entity realism is the treatment of scientific work on parity and weak neutral currents. Thus, *prima facie*, Hacking shares with the present work the goal of characterizing scientific practice. Importantly, although Hacking evidently intends to describe scientific practice—although he sees himself as providing a bit of sociology of science—he also has larger aspirations. For example, he writes,

> Do I describe a merely sociological fact about experimentalists? It is not surprising, it will be said, that these good practical people are realists. They need that for their own self-esteem. But the self-vindicating realism of experimenters shows nothing about what actually exists in the world. In reply I repeat the distinction between realism about entities and realism about theories and models. (Hacking, 1982, p. 73)

Prima facie, in the final sentence, Hacking indicates that he does not intend to make a mere sociological observation about experimentalists. He, instead, believes scientific practice has some epistemological or metaphysical implications for entity realism. Then, later, he writes,

> Even if experimenters are realists about entities, it does not follow that they are right. Perhaps it is a matter of psychology: the very skills that make for a great experimenter go with a certain cast of mind that objectifies whatever it thinks about. Yet this will not do. The experimenter cheerfully regards neutral bosons as merely hypothetical entities, while electrons are real. (Hacking, 1982, p. 76)

Hacking also comments that,

> Bain was surely right a century ago, when assumptions about the minute structure of matter could not be proved. The only proof could be indirect, namely, that hypotheses seemed to provide some explanation and helped make good predictions. Such inferences, however, need never produce conviction in the philosopher inclined to instrumentalism or some other brand of idealism. (Hacking, 1982, p. 85). Cf. Hacking (1984, p. 169)[18]

In the final sentence, Hacking suggests that he is concerned to address instrumentalists or idealists. The work in this paper does not address the extent to which scientific practice warrants an epistemological or metaphysical conclusion. That requires further argumentation. Instead, this paper aspires to a more detailed and accurate account of scientific practice that is true to the case of Hodgkin, Huxley, and Katz. Getting an accurate historical accounting is the project here.

3.1. Hacking and the Ontology of Entity Realism

Turn, now, to Hacking's formulation of the distinction between entity realism and theory realism:

> I turn to the purely philosophical issue of scientific realism. Simply call it "realism" for short. There are two basic kinds: realism about entities and realism about theories. There is no agreement on the precise definition of either. Realism about theories says that we try to form true

[18] Hacking (1982) and Hacking (1984) overlap extensively, even having the same title. Moreover, there are extensive overlaps with Hacking (1983). This paper will not attempt to sort through the variations in Hacking's presentations.

> theories about the world, about the inner constitution of matter and about the outer reaches of space. ... Realism about entities—and I include processes, states, waves, currents, interactions, fields, black holes, and the like among entities—asserts the existence of at least some of the entities that are the stock in trade of physics. (Hacking, 1984, p. 155)[19]

In principle, Hacking might have provided parallel formulations of entity realism and theory realism of the sort introduced in Section 2.2. There is, in theory, that much agreement between the entity realism developed here and Hacking's entity realism.

That said, the account in Section 2 is not meant to embrace some of Hacking's views on the relative importance of theory realism versus entity realism. Recall Hacking's comment, quoted above, that, "The vast majority of experimental physicists are realists about entities but not about theories. Some are, no doubt, realists about theories too, but that is less central to their concerns" (Hacking, 1982, p. 72). Perhaps experimentalists are committed to both entity realism and theory realism; however, the present account does not weigh in on the centrality of these distinct concerns.

It is important to note that Hacking also makes some strong modal claims that have not been explored here. Hacking proposes that "Discussions about scientific realism or antirealism usually talk about theories, explanation, and prediction. Debates at that level are necessarily inconclusive. Only at the level of experimental practice is scientific realism unavoidable—but this realism is not about theories and truth" (Hacking, 1984, p. 154). This paper does not defend the idea that debates over scientific theoretical realism are inconclusive. Much less does it defend the idea that these debates are necessarily inconclusive. Nor does the paper claim that at the level of experimental practice scientific realism is unavoidable. This paper sets aside all these strong modal claims in favor of weaker non-modal descriptive claims: Hodgkin, Huxley, and Katz were entity realists who supported their realism with abductive reasoning.

3.2. Hacking on the Confirmation of Entities

Turn now to confirmation. For the historian and philosopher of science seeking to understand experimental practice, it is worthwhile distinguishing between the idea that experimentation generates evidence (for some entity X) and the idea that experimentation generates knowledge (of some entity X). The distinction is worthwhile insofar as scientists sometimes interpret an experiment as providing some evidence for some hypothesis, even when that evidence is not sufficient for knowledge. In general, scientists may agree that a single experiment generates evidence but differ over whether the available evidence generates knowledge of the hypothesis. The discussion that follows is to be interpreted in terms of the weaker concept of evidence, rather than the stronger concept of *compelling* evidence sufficient for generating knowledge.

Hacking appears to commit to three theses regarding the confirmation of entities:

> (A) Experimentalists do not take experimentation on entities to provide evidence for the reality of entities.
>
> (B) Experimentalists do not take abductively inferring entities to provide evidence of the reality of entities.
>
> (C) Experimentalists take the manipulation of the entities to provide evidence for the reality of entities.[20]

---

[19] Hacking often uses "we" to refer to both philosophers and scientists. The intent of this paper is to characterize scientific reasoning without assuming that it is the same as philosophical reasoning. Thus, "we" is not used like this in the paper.

[20] One might interpret Hacking to be making stronger claims, not about what constitutes evidence for the reality of entities, but what constitutes knowledge of the reality of entities. Given that the more important aim of the present paper is to understand Hodgkin, Huxley, and Katz's work, rather than exhaustively survey Hacking's options, this interpretation is not pursued here. One might also interpret Hacking

These are the conclusions Hacking would have us draw from his case study of the electron and neutral bosons. These theses indicate what experimentalists take and do not take to be evidence for the reality of entities. Unfortunately, Hacking's exposition is not always as clear and univocal as these theses suggest.

Hacking claims that "Experimenting on an entity does not commit you to believing that it exists. Only manipulating an entity, in order to experiment on something else, need do that" (Hacking, 1983, p. 263).[21] The scientific review in Section 1 indicates that Hacking has overgeneralized. Contrary to thesis (A), Hodgkin and Huxley used the results of their experiments in 1939 and 1945 to support the existence of an overshoot/reversal of the action potential. Contrary to thesis (B), Hodgkin and Katz reasoned abductively about the results of their three-curve experiment to establish the existence of an inward flux of sodium ions during an action potential.[22] Finally, contrary to thesis (C), Hodgkin and Huxley believed in the action potential and its overshoot before Hodgkin and Katz used them in their 1949 interpretation of the action potential. One of the upshots of Hodgkin and Huxley's 1945 paper was that the overshoot was not an artifact. Moreover, if Hodgkin and Katz had not accepted the action potential and its overshoot as real, rather than an artifact, they would not have used it as a basis for their three-curve experiment in 1949.

The interpretation of Hacking's comment that "Experimenting on an entity does not commit you to believing that it exist" is not absolutely unassailable. Maybe by "commit" Hacking means to imply some sort of compulsion to believe in an entity that is not found in the Hodgkin, Huxley, and Katz case. But what is this compulsion? And maybe the "you" is meant only to philosophers reading the paper, but not experimental physiologists. But why would the attitude of philosophers matter to how one is to understand Hodgkin, Huxley, and Katz's reasoning? These are not questions pursued here, since the primary concern is with understanding Hodgkin, Huxley, and Katz's entity realism and their experimental reasoning, rather than an exhaustive critique of Hacking's entity realism.

When it comes to thesis (B), Hacking is at times decidedly equivocal.[23] Much of his reasoning is framed in terms of the importance of an explanation. In an early chapter in *Representing and Intervening*, Hacking claims that,

> There are times when we feel a great gain in understanding by the organization of new explanatory hypotheses. But that feeling is not a ground for supposing that the hypothesis is true. Van Fraassen and Cartwright urge that being an explanation is never a ground for belief. I am less stringent than they: it seems to me like Peirce to be merely a feeble ground. In 1905 Einstein explained the photoelectric effect with a theory of photons. He thereby made attractive the notion of quantized bundles of light. But the ground for believing the theory is its predictive success, and so forth, not its explanatory power. Feeling the key turn in the lock makes you feel that you have an exciting new idea to work with. It is not a ground for the truth of the idea: that comes later. (Hacking, 1983, p. 53)

In the first two sentences, one might think that Hacking's view is that explanation is not evidence in favor of a hypothesis. But then again, he could be read as maintaining that an explanation generates both evidence and a feeling, but that it is not the *feeling* that provides the evidence. This allows an explanation to be evidence in favor a hypothesis. This second interpretation is consistent with what Hacking says next: unlike van Fraassen and Cartwright, but like Peirce, he

---

as making normative claims, such as that experimentalists should not take experimentation on entities to provide evidence for the reality of entities. These normative claims go beyond the present concern with getting an accurate historical accounting of Hodgkin, Huxley, and Katz's work.

21 Hacking also claims that "The experimentalist does not believe in electrons because, ... they 'save the phenomena.' [Thesis (B)] On the contrary, we believe in them because we use them to create new phenomena, such as the phenomenon of parity violation in weak neutral current interactions." [Thesis (C)]

22 Further, as noted above in footnote 12, they also performed another abductive inference to the conclusion that action potentials involve the existence of a change in membrane permeability.

23 See also footnote 16.

believes that an explanation can provide a weak ground.[24] Now the view appears to be that explanation does provide evidence; but only weak evidence. But then, following the comments about Einstein, Hacking explicitly denies that explanatory power provides ground, that is, evidence. Instead, it is predictive success that provides confirmation. Then, he returns to the idea that the feeling of understanding that comes with explanation does not provide confirmation, again leaving open the possibility that there is something about explanation *other than the feeling of understanding* that does provide confirmation. But, then shortly after, he comments that "Even if, contrary to what I have said, explanation were a ground for belief ..." (Hacking, 1983, p. 54). As noted, Hacking is thoroughly equivocal here.

In a later paper, Hacking is less equivocal but less direct in presenting his view. He writes,

> Inference to the best explanation is often traced back to C. S. Peirce and what he called the method of hypothesis (and also retroduction, abduction). In maturity Peirce did not think that the method of hypothesis gave ground for belief, but only for further investigation, testing, and induction. (Hacking, 1989, p. 568)

It is true that inference to the best explanation is often traced back to Peirce. Harman, for example, claimed that "'The inference to the best explanation' corresponds approximately to what others have called 'abduction'" (Harman, 1965, p. 88), and this is widely understood to be an implicit reference to Peirce. This is not, however, to say that IBE and abduction are the same. As noted above, Harman believed that IBE is *warranted* abduction, so for Harman at least, there is a difference. Moreover, the difference is supposed to be that IBE is supposed to provide warrant, so *prima facie* it is supposed to provide evidence. By contrast, judging from his subsequent comments, Hacking means to identify IBE and abduction, and then understand both, as Peirce did, as *not* providing ground/ evidence/ confirmation/ warrant.[25]

Given that Hodgkin, Huxley, and Katz's experimental work conflicts with the foregoing reading of Hacking's view, one might wonder if Hacking does, in fact, really deny that scientists use abduction for confirmation. Indeed, one might think that all Hacking really wants to say is that the evidence from manipulation is better than evidence from abduction. This is, after all, what he claims: "The best kinds of evidence for the reality of a postulated or inferred entity is that we can begin to measure it or otherwise understand its causal powers. The best evidence, in turn, that we have this kind of understanding is that we can set out, from scratch, to build machines that will work fairly reliably, taking advantage of this or that causal nexus" (Hacking, 1983, p. 274).

If Hacking were to retreat to the weaker view that manipulation creates better evidence than abduction, that would be a major concession.[26] It would concede the main point of this paper. If he were to allow that abductive reasoning provides confirmation (that is weaker than confirmation by manipulation), he would thereby risk opening the floodgates to many more unobservable entities, not just the relatively fewer unobservable entities that can be manipulated for further experimental ends. Bear in mind that he supports an anti-realism in astrophysics, where he takes manipulation for experimentation to be impossible.[27]

The upshot here is that—admitting the various ambiguities in Hacking's writing—Hodgkin, Huxley, and Katz's work challenges all three of Hacking's descriptive methodological theses. These mid-twentieth-century experimental physiologists used experiments on action potentials to confirm the existence of the overshoot. Further, they used abductive

---

[24] Note that many Peirce scholars believe that Peirce thought that abduction serves merely to introduce some hypothesis, rather than to confirm it. See Cabrera (2023) for discussion and references.

[25] Note, as well, that even if in maturity Peirce did not think that the method of hypothesis gave ground for belief, that leaves open the issue about whether Peirce was right. One of the morals of Sections 1 and 2 is that Hodgkin, Huxley, and Katz, at least, believed that abductive reasoning provides evidence.

[26] Were Hacking to retreat to this view, it should invite closer questioning of the reasons he has for it.

[27] See Hacking (1989).

reasoning to confirm the existence of an early sodium influx during an action potential.[28] Hodgkin, Huxley, and Katz were, thus, entity realists who confirmed the existence of certain entities using abduction.

5. Conclusion

The examination of a portion of Hodgkin, Huxley, and Katz's work from 1939 to 1949 reveals that they embraced entity realism about action potentials, changes in membrane permeability, and ion fluxes. They also believed in some of these entities for abductive reasons. This conclusion is at odds with Hacking's combination of entity realism with a rejection of abductive reasoning. Moreover, this conclusion is not on the philosophical radar in current work on entity realism.

There are, of course, any number of further issues to be explored. For one thing, we have reviewed only one instance of entity realists embracing abduction. How many more are there, if any? What normative conclusions might one draw from the cases? Presumably, the more instances of highly respected scientific work, such as Hodgkin, Huxley, and Katz's work on the action potential, the stronger the case for thinking that philosophers of science should take seriously the view that abductive reasoning can support entity realism. Finally, there are many details of Hacking's philosophy of experiment that one might wish to pursue. Why think that manipulationist evidence is stronger than abductive evidence? Why think that theory realism is necessarily inconclusive?

REFERENCES

Aizawa, K. (2025). *Compositional Abduction and Scientific Interpretation: A Granular Approach*. Cambridge University Press.

Aizawa, K. (Forthcoming). Singular Compositional Abduction and Van 't Hoff's *Voorstel*. In Y. Shan (Ed.), *Epistemology and Philosophy of Science*.

Aizawa, K., & Headley, D. B. (2022). Abduction and composition. *Philosophy of Science*, *89*(2), 268-282.

Aizawa, K., & Headley, D. B. (2025). Scientific constitutive abduction. *European Journal for Philosophy of Science*, *15*(2), 28. https://doi.org/10.1007/s13194-025-00655-3

Bird, A. (2022). *Knowing Science*. Oxford University Press.

Cabrera, F. (2023). Inference to the Best Explanation: An Overview. In L. Magnani (Ed.), *Handbook of Abductive Cognition* (pp. 1863-1896). Springer International Publishing. https://doi.org/10.1007/978-3-031-10135-9_77

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press.

Craver, C. F. (2007). *Explaining the Brain*. Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780199299317.001.0001

Davey, K. (forthcoming). There are no bad lots, only bad formulations of inference to the best explanation. *British Journal for the Philosophy of Science*.

Dellsén, F. (2024). *Abductive Reasoning in Science*. Cambridge University Press. https://doi.org/10.1017/9781009353199

Eronen, M. I. (2019). Robust realism for the life sciences. *Synthese*, *196*(6), 2341-2354.

[28] They also used abductive reasoning to confirm the existence of a change in membrane permeability to sodium.

Hacking, I. (1982). Experimentation and scientific realism. *Philosophical Topics*, *13*(1), 71-87.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press.

Hacking, I. (1984). Experimentation and scientific realism. In J. Leplin (Ed.), *Scientific realism*. University of California Press.

Hacking, I. (1989). Extragalactic reality: The case of gravitational lensing. *Philosophy of Science*, *56*(4), 555-581.

Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review*, *74*(1), 88-95.

Hodgkin, A. (1994). *Chance and Design: Reminiscences of Science in Peace and War*. Cambridge University Press.

Hodgkin, A. L., & Huxley, A. F. (1939). Action potentials recorded from inside a nerve fibre. *Nature*, *144*(3651), 710-711.

Hodgkin, A. L., & Huxley, A. F. (1945). Resting and action potentials in single nerve fibres. *The Journal of Physiology*, *104*(2), 176.

Hodgkin, A. L., & Katz, B. (1949). The effect of sodium ions on the electrical activity of the giant axon of the squid. *The Journal of Physiology*, *108*(1), 37-77.

Khalili, M. (2023). A dialogue among recent views of entity realism. *Philosophy of Science*, *90*(4), 901-921.

Khalili, M. (2024). Entity Realism Meets Perspectivism. *Acta Analytica*, *39*(1), 79-95.

Khalili, M., & Kazemi, A. (2025). Entity Realism about Implicit Attitudes. *Topoi*, *44*(4), 877-888.

Kim, J. (1994). Explanatory knowledge and metaphysical dependence. *Philosophical Issues*, *5*, 51-69.

Lipton, P. (2004). *Inference to the Best Explanation*. Routledge.

Musgrave, A. (1996). Realism, truth and objectivity. In R. S. Cohen, R. Hilpinen, & R.-Z. Qiu (Eds.), *Realism and Anti-Realism in the Philosophy of Science: Beijing International Conference, 1992* (pp. 19-44). Springer.

Nanay, B. (2019). Entity realism and singularist semirealism. *Synthese*, *196*(2), 499-517.

Nanay, B. (2022). Entity realism about mental representations. *Erkenntnis*, *87*(1), 75-91.

Patton, L. (2025). Does Entity Realism Hold Up? *The Monist*, *108*(4), 337-352.

Psillos, S. (2002). Simply the best: A case for abduction. In A. Kakas & F. Sadri (Eds.), *Computational logic: Logic programming and beyond* (pp. 605-625). Springer.

Schupbach, J. N. (2017). Inference to the best explanation, cleaned up and made respectable. In K. McCain & T. Poston (Eds.), *Best Explanations: New Essays on Inference to the Best Explanation* (pp. 39-61). Oxford University Press.

Schurz, G. (2008). Patterns of abduction. *Synthese*, *164*(2), 201-234.

Ken Aizawa is Professor of Philosophy at Rutgers University, Newark. He is a historian and philosopher of science currently examining the role of abductive confirmation in science. He is the author of *Compositional Abduction and Scientific Interpretation: A Granular Approach* from Cambridge University Press.

Address: Department of Philosophy, Rutgers University, Newark. 404 Conklin Hall, 175 University Avenue, Newark, NJ 07102. USA. Email: ken.aizawa@gmail.com. ORCID: https://orcid.org/0000-0002-7143-4901

Theoria

An International Journal for Theory, History and Foundations of Science

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Summary

Volume 41/3, September 2026, pp. 345-469

Articles

Karl Gustav Bergman (Universitat Pompeu Fabra), «Are proper functions normative?», *Theoria*, 2026, 41/3, 347-366

This paper addresses the question: Are proper functions normative? It draws a distinction between *bearer* normativity and *agent* normativity and gives three arguments for the claim that proper functions are normative in the former sense. It ends by suggesting that two of these arguments can be generalized to the agentive case.

Esteban Withrington (UNAM), «Millikan on picturing and correspondence», *Theoria*, 2026, 41/3, 367-377

Ruth Millikan characterizes her account of sentential truth as a pictorial theory inspired by Tractarian ideas of isomorphism. I argue that her account is best developed in terms of systematic correlation, without requiring relations of structural resemblance between sentences and reality.

Santiago Arango-Muñoz (Universidad de Antioquia) y Sebastián Mejía-Rendón (Universidad Nacional de Córdoba), «Metacognitive feelings as mental affordances», *Theoria*, 2026, 41/3, 378-395

This paper argues that metacognitive feelings are mental affordances. Building on McClelland's Mental Affordance Hypothesis, we show that they make agents sensitive to their own cognitive states while motivating and regulating specific mental actions. We conclude that the Mental Affordance Hypothesis should be expanded to incorporate metacognitive feelings as internal affective affordances for mental action.

Marco Facchin (University of Antwerp), «Illusionism as a theory of mental content», *Theoria*, 2026, 41/3, 396-433

This programmatic paper introduces content illusionism in the philosophical debate over mental content. To this end, the paper first introduces the doctrine of content illusionism (§2), offers some arguments in favor of it (§3), deflects the most obvious objections against it (§4), and sketches a minimal mechanism responsible for the illusion of content (§5).

Murali Ramachandran (Thapar University), «The opacity of knowledge: An externalist's case for widespread KK-failure», *Theoria*, 2026, 41/3, 434-444

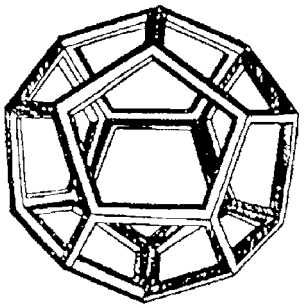
Epistemic externalists take knowledge to be underpinned by facts that subjects may not be in a position to know. We contend knowers themselves are generally aware of such facts that their knowledge depends on. On this basis we argue for widespread violations of the KK-principle: hardly anything known (about the world) is known to be known!

Ken Aizawa (Rutgers University), «A case of entity realists using abductive confirmation», *Theoria*, 2026, 41/3, 445-460

In their physiological work circa 1950, Alan Hodgkin, Andrew Huxley, and Bernard Katz revealed themselves to be entity realists who used abduction to experimentally confirm the existence of unobserved entities. This runs counter to Ian Hacking's proposal that experimentalists confirm unobserved entities by manipulating them, rather than by using abduction.

THEORIA

AN INTERNATIONAL JOURNAL FOR THEORY, HISTORY AND FOUNDATIONS OF SCIENCE



Ὁ Θεὸς ἀριθμετίζει

VOL. 41

STABLISHED IN 1952 — SECOND PERIOD
BY Miguel Sánchez-Mazas (†)

THEORIA is associated with the *Society of Logic, Methodology, and Philosophy of Science in Spain* (SLMFCE) and the *Spanish Society for Analytic Philosophy* (SEFA)

Publishers:

Center for Analysis, Logic, and Legal Informatics (CALIJ)

University of the Basque Country Press

EHU

Donostia-San Sebastián / ISSN 0495-4548 e-ISSN 2171-679X

Vol. 41/1

Articles

Luis Estrada-González, Christian Romero-Rodríguez (Institute for Philosophical Research, UNAM), «How we learned to stop worrying and love tonk», *Theoria*, 2026, 41/1, 5-21.

Leonardo Flamini (University of Johannesburg), «On asking», *Theoria*, 2026, 41/1, 22-35.

Sergio Cermeño-Aínsa (University of Girona), «Perceptual reference, object files and Molyneux's Question», *Theoria*, 2026, 41/1, 36-56.

Peter J. Riggs (Australian National University), «Time dilation and rates of the passage of time», *Theoria*, 2026, 41/1, 57-70.

Eduard Dàvila (University of Valladolid), «A critical analysis of cosmological typicality and the anthropic principle», *Theoria*, 2026, 41/1, 71-86.

Francesc Roig-Loscertales (UNED), «¿Hay un sesgo de financiación en la investigación biomédica?», *Theoria*, 2026, 41/1, 87-102.

Vol. 41/2

Articles

Valentine Simpson (Universitat de València), Matheus Valente (UNED), «The philosophy of José Luis Bermúdez: an introduction», *Theoria*, 2026, 41/2, 113-121.

Víctor M. Verdejo (Pompeu Fabra University), «Thoughts about oneself to share in context: meeting Bermúdez's challenge», *Theoria*, 2026, 41/2, 122-135.

Matheus Valente (UNED), «Asymmetries between 'you' and 'I'», *Theoria*, 2026, 41/2, 136-156.

Derek H. Brown (University of Glasgow), «From immediate perception to perceptual belief», *Theoria*, 2026, 41/2, 157-178.

Shaun Gallagher (University of Memphis, University of Wollongong), «Why a perceptual zero-point amounts to nothing», *Theoria*, 2026, 41/2, 179-195.

Carlota Serrahima (University of Valencia), «Self-consciousness, ψ and ϕ», *Theoria*, 2026, 41/2, 196-214.

Daphne Bernués Moss, Marta Jorba Grau (Universitat Pompeu Fabra), «Bermúdez's view on inner speech: a critical assessment», *Theoria*, 2026, 41/2, 215-230.

Daniel Dohrn (Università degli Studi di Milano), «Can quasi-cyclical preferences be rational?», *Theoria*, 2026, 41/2, 231-244.

Sarah A. Fisher (Cardiff University), «Discursive deadlock and the social (media) life of frames», *Theoria*, 2026, 41/2, 245-261.

Marko Jurjako (University of Rijeka), «How general are Marr's levels of analysis? An assessment of Bermúdez's view», *Theoria*, 2026, 41/2, 262-282.

José Luis Bermúdez (Texas A&M University), «Self, inner speech, rationality and explanation: replies to Valente, Verdejo, Brown, Gallagher, Serrahima, Bernués and Jorba, Dohrn, Fisher and Jurjako», *Theoria*, 2026, 41/2, 283-338.

Vol. 41/3

Articles

Karl Gustav Bergman (Universitat Pompeu Fabra), «Are proper functions normative?», *Theoria*, 2026, 41/3, 347-366.

Esteban Withrington (UNAM), «Millikan on picturing and correspondence», *Theoria*, 2026, 41/3, 367-377.

Santiago Arango-Muñoz (Universidad de Antioquia), Sebastián Mejía-Rendón (Universidad Nacional de Córdoba), «Metacognitive feelings as mental affordances», *Theoria*, 2026, 41/3, 378-395.

Marco Facchin (University of Antwerp), «Illusionism as a theory of mental content», *Theoria*, 2026, 41/3, 396-433.

Murali Ramachandran (Thapar University), «The opacity of knowledge: An externalist's case for widespread KK-failure», *Theoria*, 2026, 41/3, 433-444.

Ken Aizawa (Rutgers University), «A case of entity realists using abductive confirmation», *Theoria*, 2026, 41/3, 445-460.

Theoria

An International Journal for Theory, History and Foundations of Science

Vol. 41

Luis Estrada-González, Christian Romero-Rodríguez (Institute for Philosophical Research, UNAM), «How we learned to stop worrying and love tonk», *Theoria*, 2026, 41/1, 5-21.

Leonardo Flamini (University of Johannesburg), «On asking», *Theoria*, 2026, 41/1, 22-35.

Sergio Cermeño-Aínsa (University of Girona), «Perceptual reference, object files and Molyneux's Question», *Theoria*, 2026, 41/1, 36-56.

Peter J. Riggs (Australian National University), «Time dilation and rates of the passage of time», *Theoria*, 2026, 41/1, 57-70.

Eduard Dàvila (University of Valladolid), «A critical analysis of cosmological typicality and the anthropic principle», *Theoria*, 2026, 41/1, 71-86.

Francesc Roig-Loscertales (UNED), «¿Hay un sesgo de financiación en la investigación biomédica?», *Theoria*, 2026, 41/1, 87-102.

Valentine Simpson (Universitat de València), Matheus Valente (UNED), «The philosophy of José Luis Bermúdez: an introduction», *Theoria*, 2026, 41/2, 113-121.

Víctor M. Verdejo (Pompeu Fabra University), «Thoughts about oneself to share in context: meeting Bermúdez's challenge», *Theoria*, 2026, 41/2, 122-135.

Matheus Valente (UNED), «Asymmetries between 'you' and 'I'», *Theoria*, 2026, 41/2, 136-156.

Derek H. Brown (University of Glasgow), «From immediate perception to perceptual belief», *Theoria*, 2026, 41/2, 157-178.

Shaun Gallagher (University of Memphis, University of Wollongong), «Why a perceptual zero-point amounts to nothing», *Theoria*, 2026, 41/2, 179-195.

Carlota Serrahima (University of Valencia), «Self-consciousness, ψ and ϕ», *Theoria*, 2026, 41/2, 196-214.

Daphne Bernués Moss, Marta Jorba Grau (Universitat Pompeu Fabra), «Bermúdez's view on inner speech: a critical assessment», *Theoria*, 2026, 41/2, 215-230.

Daniel Dohrn (Università degli Studi di Milano), «Can quasi-cyclical preferences be rational?», *Theoria*, 2026, 41/2, 231-244.

Sarah A. Fisher (Cardiff University), «Discursive deadlock and the social (media) life of frames», *Theoria*, 2026, 41/2, 245-261.

Marko Jurjako (University of Rijeka), «How general are Marr's levels of analysis? An assessment of Bermúdez's view», *Theoria*, 2026, 41/2, 262-282.

José Luis Bermúdez (Texas A&M University), «Self, inner speech, rationality and explanation: replies to Valente, Verdejo, Brown, Gallagher, Serrahima, Bernués and Jorba, Dohrn, Fisher and Jurjako», *Theoria*, 2026, 41/2, 283-338.

Karl Gustav Bergman (Universitat Pompeu Fabra), «Are proper functions normative?», *Theoria*, 2026, 41/3, 347-366.

Esteban Withrington (UNAM), «Millikan on picturing and correspondence», *Theoria*, 2026, 41/3, 367-377.

Santiago Arango-Muñoz (Universidad de Antioquia), Sebastián Mejía-Rendón (Universidad Nacional de Córdoba), «Metacognitive feelings as mental affordances», *Theoria*, 2026, 41/3, 378-395.

Marco Facchin (University of Antwerp), «Illusionism as a theory of mental content», *Theoria*, 2026, 41/3, 396-433.

Murali Ramachandran (Thapar University), «The opacity of knowledge: An externalist's case for widespread KK-failure», *Theoria*, 2026, 41/3, 433-444.

Ken Aizawa (Rutgers University), «A case of entity realists using abductive confirmation», *Theoria*, 2026, 41/3, 445-460.

GUIDELINES FOR AUTHORS

THEORIA will only consider original papers in the areas of competence of our editorial board, never exceeding 10.000 words. We will try to make a decision on your paper within eight weeks.

We will only accepts electronic submissions. For further Instructions, please, visit http://www.ehu.es/theoria

EDITORIAL PROCEDURE

THEORIA is a peer-reviewed journal. The editors, together with the competent members of the editorial board, will screen the submissions and decide whether to send them to referee. Every paper will be then assessed by two (blind) reviewers. On this basis, the editors will make the final decision. The list of reviewers is regularly published in the journal. Our archives in the editorial platform are open to research agencies (such as the Spanish FECYT) for external verification of the quality of the review process.

COPYRIGHT NOTICE

THEORIA is published under a Creative Commons Licence. Authors who publish with this journal agree to the following terms:

- Authors retain copyright and grant the journal right of first publication with the work simultaneously licensed under a Creative Commons Attribution License that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.
- Authors are able to enter into separate, additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal.
- Authors are permitted and encouraged to post their work online (e.g., in institutional repositories or on their website) prior to and during the submission process, as it can lead to productive exchanges, as well as earlier and greater citation of published work.

SUBSCRIPTION AND BACK ISSUES

THEORIA is published 3 times a year. Annual subscription rates for 2010 are: €50 for individuals and €70 for institutions. Single Issues can be ordered for €20 each. A Cumulative Index covering the numbers 21-50 is available for €3. For ordering information or any enquiry concerning subscriptions, please contact: publicaciones@ehu.eus.

Publishers

This journal is published jointly by the University of the Basque Country and the Center for Analysis, Logic, and Legal Informatics (CALIJ) in Donostia/San Sebastián, Spain.

CALIJ

EHU

THEORIA is associated with
the Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España (SLMFCE)
and the Sociedad Española de Filosofía Analítica (SEFA)

THEORIA

An International Journal for Theory, History and Foundations of Science

Vol. 41/3 • Septiembre 2026

Dep. Legal: SS - 820/85 ISSN 0495-4548 e-ISSN 2171-679X

Articles

Karl Gustav Bergman
Are proper functions normative? _____ 347

Esteban Withrington
Millikan on picturing and correspondence _____ 367

Santiago Arango-Muñoz, Sebastián Mejía-Rendón
Metacognitive feelings as mental affordances _____ 378

Marco Facchin
Illusionism as a theory of mental content _____ 396

Murali Ramachandran
The opacity of knowledge: An externalist's case for widespread KK-failure _____ 434

Ken Aizawa
A case of entity realists using abductive confirmation _____ 445

Summary _____ 462