Troubles with mathematical contents

Marco Facchin [**corresponding author**]
Istituto Universitario di Studi Superiori IUSS Pavia
Linguistics and Philosophy IUSS Center; Pavia, Italy
Palazzo del Broletto, Piazza della Vittoria n. 15, 27100, Pavia
marco.facchin@iusspavia.it
https://orcid.org/0000-0001-5753-9873

Abstract:

To account for the explanatory role representations play in cognitive science, Egan's deflationary account introduces a distinction between cognitive and mathematical contents. According to that account, only the latter are genuine explanatory posits of cognitive-scientific theories, as they represent the arguments and values cognitive devices need to represent to compute. Here, I argue that the deflationary account suffers from two important problems, whose roots trace back to the introduction of mathematical contents. First, I will argue that mathematical contents do not satisfy important and widely accepted *desiderata* all theories of content are called to satisfy, such as content determinacy and naturalism. Secondly, I will claim that there are cases in which mathematical contents cannot play the explanatory role the deflationary account claims they play, proposing an empirical counterexample. Lastly, I will conclude the paper highlighting two important implications of my arguments, concerning recent theoretical proposals to naturalize representations *via* physical computation, and the popular predictive processing theory of cognition.

Keywords: Representation, Content, Mathematical contents, Computation, Implementation, Adversarial examples.
9797 words all included

Declarations:
Founding: Not applicable
Conflict of Interests/competing interests: The author declares no conflict of interests.
Disclosure statement: The author reports there are no competing interests to declare.
Availability of data and Material: Not applicable.
Code availability: Not applicable.
Authors' contribution: Marco Facchin is the sole author of the paper.

Acknowledgments: Thanks to the members of the IUSS 2021 reading group: discussing with you gave me the idea to write this paper. Early drafts of this paper have been presented at the SIFA 2021 Conference in Noto and the AISC early career online event. Thanks to the audience of these conferences for their insightful questions. A special thanks goes to Joe Dewhurst for his extensive and insightful feedback on an earlier (and really poorly put together!) version of this paper.

1 - Introduction

Which factors turn the firing of some neurons into a representation? A prominent way to answer is by *naturalizing content*; i.e. pointing to a naturalistic relation holding between neural activity and worldly targets *in virtue of which* the former represents the latter. "Classic" proposals point to causal/informational factors (e.g. Dretske 1988; Fodor 1990), biological functions (e.g. Millikan 1984) or abstract notions of resemblance (e.g. O'Brien and Opie 2004). Despite their differences, all these proposals face the problem of *content (in)determinacy*: due to the number of causal intermediaries (e.g. Artiga and Sebastian 2018), the indeterminacy of biological functions (e.g. Hutto and Myin 2013, Ch.4) or the ubiquity of abstract resemblance relations (e.g. Sprevak 2001), these proposals all fail to provide unique and univocally determined representational contents.

Egan (2014; 2019; 2020a) suggests these failures motivate a different agenda. Rather than naturalizing content, we should aim at capturing the explanatory role content plays in cognitive science.[1] To this end, Egan's (2014; 2020a) deflationary account identifies two distinct kinds of content: *cognitive and mathematical*. Cognitive contents are contents usually understood (representations of worldly targets), which are not really represented "inside" cognitive systems, for they only belong to a *facultative gloss* layered over cognitive-scientific explanations. Conversely, mathematical contents are genuine ingredients of cognitive-scientific explanations, and so they really are represented "inside" cognitive systems. Or so, at least, Egan claims.

Here, I want to challenge this account, which I'll summarize in §2. Then, In §3, I will raise my first challenge, arguing that the deflationary account *does not* satisfy the desiderata it sets for itself: like "classic" accounts of content, it cannot provide well determined contents *while* staying natural in the relevant sense. In §4, I will raise my second challenge, claiming, *contra* Egan, that there is at least one family of cases (namely, adversarial example induced misclassifications) in which cognitive contents are not just a "facultative gloss", being needed to provide a satisfactory explanation. Lastly, in §5, I will highlight some implications of my challenges. To anticipate a bit, these implications concern the development of some "computation based" and deflationary-account inspired accounts of representations, with an attention to accounts or representations tied to the popular neurocomputational theory of predictive processing (e.g. Wiese 2017).

[1] Really, in *computational* cognitive science. Since anti-computationalists are typically also anti-representationalsits (e.g. Hutto and Myin 2013), I will ignore them here.

2 - Egan’s deflationary account of representations

A perspicuous way to understand Egan’s deflationary account of representations is to consider it as a reaction motivated by the failure of “classic” accounts. In particular, “classic” accounts fail to satisfy these extremely widely accepted desiderata:[2]

(1) Misrepresentation: A successful account of representations allows for misrepresentation to occur

(2) Determinacy: A successful account of representations assigns determinate contents to representational vehicles

(3) Empirical adequacy: A successful account of representations conforms to the actual practice of cognitive science

(4) Naturalism: A successful account of representations specifies, using non-intentional and non-semantic terms, at least sufficient conditions for a state or structure to bear a determinate content

(5) No pan-representationalism: A successful account of representations does not imply that many clearly non-representational things count as representations

Being extremely widespread, **(1)** to **(5)** need little introduction, so I’ll limit myself to the few remarks relevant for my arguments.

(1) and **(2)** are *constitutively* connected. The ability to misrepresent **(2)** *identifies* representations, setting them apart from mere states or objects (cf Dretske 1986). But misrepresentation requires **(1)** *determinate* contents: open-endedly disjunctive contents make misrepresentation, if not impossible, at least highly problematic. That’s why content determinacy is such a big problem for “classic” accounts.

Empirical adequacy **(3)** is not “just” a desideratum in its own right, it is *the essential desideratum* of the deflationary account. For the

[2] See Egan (2019: 248-249; 2020a 28-29). In her (2020a) Egan mentions *more* desiderata than I report. I omit those for the sake of brevity.

deflationary account *aims* at capturing the explanatory role of representations in cognitive science; hence its success is predicated on it conforming to the relevant empirical practice - the deflationary account *just cannot* be revisionary. Importantly, satisfying **(1)** and **(2)** is a prerequisite for satisfying **(3)**, for cognitive scientists *do* posit representations with well determined contents (e.g. Backer *et al*. 2021).

Condition **(4)** captures the widespread idea that content is not a basic ingredient of the world; contents depend on more basic features, in terms of which they can be explained. These features may, but *need not*, be the one "classic" accounts aiming to naturalize content rely on. Other features can do the trick too.[3] To satisfy **(4)**, the only thing that matters is that these features are not *already* semantic (or intentional, or contentful).

Lastly, **(5)** is a prerequisite for **(3)**, at least insofar cognitive scientists do not label *every* behavior-producing structure a representation (cf. Webb 2006). Moreover, it safeguards the explanatory power of representations: pan-representationalism trivializes the explanatory power of content, equating representations to mere causal mediators (see Ramsey 2007; Orlandi 2020).

Now, as indicated in §1, Egan's (2019a; 2020a) deflationary account is motivated by the fact that "classic" accounts fail to satisfy **(1)** and **(2)**. But what's the shape of the proposed alternative? Since her account mainly aims at accounting for the explanatory role representations play in cognitive science, cognitive-scientific explanations offer a natural starting point to illustrate it.

Egan (2010; 2014; 2017; 2020) construes cognitive-scientific explanations as *function-theoretic*: they unveil the mathematical function F computed by a cognitive device S. These explanations characterize only the input/output behavior of S, thereby sitting at Marr's (1982) computational level. And, indeed Egan uses Marr's computational account of vision as a prototypical example of a function theoretic explanation.

[3] As we will shortly see, features relevant to *computational implementation* may do the trick (see Coelho Mollo 2021; Piantadosi 2021).

According to Marr, retinas - the system S - contribute to vision by computing a smoothing function F convolving a Laplacian operator with a Gaussian operator. This, Egan claims, is a function-theoretic explanation, which tells us *all there is to know* about the role retinas play in vision.

But what does it mean to say a system S computes a function F? Egan (2010; 2014; 2020a) suggests that S computes F just in case:

(i) There exist a *realization function* f_R mapping, in a many-to-one fashion, the physical states of S onto a range of vehicle types; &

(ii) There exist an *interpretation function* f_I mapping, in a one-to-one fashion, the relevant vehicle types in **(i)** onto the values and arguments of F; &

(iii) For all argument - value pairs of F, if S is in a state that, according to **(i)** and **(ii)**, maps on a specific argument of F, then S is caused to enter in a state that, according to **(i)** and **(ii)** maps on the corresponding value of F

Less formally: **(i)** f_R identifies the computational state types (or vehicle types) tokened in S; **(ii)** f_I matches them to the arguments and values of F in a one-to-one fashion, and **(iii)** says that S computes F just in case the state-transitions in S "march in step" with the argument-value pairings of F. Egan (2014; 2020) provides this simple example. Suppose S computes the *addition function* F. This means that **(i)**: there is a function f_R grouping S's states together in well defined vehicle types; & **(ii)** there is a one-to-one mapping f_I from these vehicle types onto numbers, such that; **(iii)** if S is in a state s' (as identified by f_R) and f_I(s')=*n*, and then receives an input causing it to occupy state s'' and f_I(s'')=*m*, then S is caused to enter a state s''' and f_I(s''')=*n* + *m*.

An alternative way to spell out **(ii)** is by saying that f_I gives to the vehicle types identified by f_R their *mathematical contents*. Mathematical contents thus represent abstract objects; namely the arguments and values of the F computed by S.[4] Mathematical contents are also the explanatory factors highlighted by function-theoretic explanations, of which they are an essential component (Egan 2014: 122-123). They explain (but more of this in §4) by subsuming the behavior of a physical system under a mathematical function we *already* understand. They translate something unknown (a system's behavior) into something independently known (a mathematical function), allowing us to postdict and predict the behavior of the system in a wide range of possible circumstances (cf Egan 1999; 2010; 2014; 2017; 2020). Knowing function-theoretic characterization of a device S, we know how S behaves given some relevant input.

According to Egan (2014; 2019; 2020) function theoretic explanations are not *complete* explanations of our cognitive capacities. They only inform us that a system S computes a function F. But *why* does computing F constitute a cognitive capacity? To answer, we need to supplement the function-theoretic explanation with an *ecological component*: a series of assumptions about the environment S operates in, allowing us to understand *how* computing F contributes to cognition. To continue with Marr's retinical example: it is only because *in this environment* adjacent retinal cells receive (roughly) the same amount of light that computing a smoothing function allows sharp changes in illumination to "pop up", thereby allowing a system to detect edges.

Note how the ecological component invites *cognitive* contents into the picture: doesn't the ecological component suggest that retinas *represent* edges? Egan (2014; 2019: 254; 2020a) answers negatively: retinas *really* only represent mathematical contents, the representation of which allows a system in an environment such as ours to "use" retinas to detect edges. Yet, Egan concedes, that's quite a mouthful, and so we can *say for simplicity's sake* that retinas represent edges. Ascriptions of cognitive contents, then, allows us to summarize the explanatory job done by the ecological components in an understanding-friendly manner. They're a

[4] Which need not be *numbers*. If F is a function from vectors to labels (as many neural networks), the relevant mathematical contents will be vectors and labels, which are not numbers.

perspicuous, but really not-strictly-speaking-needed and thus *not ontologically committing*, summary of how *S*'s computing of F in a given environment contributes to the cognitive capacity under investigation.

Thus, to repeat, ascriptions of cognitive contents are *just, merely and only* ascriptions. Retinas *do not really* token representations of edges; they only really token representations of the inputs and outputs of F (Egan 2014; 2020a). Indeed, in her view, *no cognitive contents are really ever represented within cognitive systems*. Only mathematical contents are. So, there's really no *fact of the matter* about which cognitive contents cognitive systems really represent over and above our ascriptions (cf. Coelho Mollo 2020). Cognitive contents are only ascribed from the outside, based on our pragmatic and explanatory interests - roughly, based on how user-friendly the grip they afford over a system's behavior is. And thus the cognitive-content based talk is revealed to be just an informal, and strictly speaking facultative, "gloss" over genuine cognitive-scientific explanations (Egan 2014; 2019; 2020a). [5]

And it is precisely this "glossy" nature that allows Egan's (2014; 2020a) account to successfully face desiderata **(1)**-**(5)**, or at least the *relevant* desiderata in that list. As a matter of fact, we ascribe determinate contents: we say retinas represent edges, not that they represent the disjunction "edges or shadows". Thus, **(2) determinacy** is satisfied. But given their constitutive connection, **(1) misrepresentation** is satisfied too. Being built on several case studies, we can expect the deflationary account to satisfy **(3) empirical accuracy**. Egan claims her account is safe from pan-representationalims, satisfying **(5)**. After all, on her account cognitive contents are ascribed, and surely we don't ascribe cognitive contents to *everything*. And while the account fails **(4) naturalism**, for the explanatory interests grounding cognitive contents are as intentional and as contentful as it gets, failing **(4)** is now a "don't care" factor. Since cognitive contents are no longer *really* part of cognitive-scientific explanations, the naturalistic credentials of cognitive science are not

[5] However, it can play a *heuristic* role in guiding the empirical investigation of a device, if no function theoretic characterization of the device is available (Egan 2020a: 45-48).

under threat. The non-naturality of cognitive contents is quarantined in an "informal gloss" over the real scientific theory, and thus does not spread to the latter. The "glossy" nature of cognitive contents makes **(4)** irrelevant, allowing Egan's account to *ignore* it.

Even supposing Egan is right on *all* of the above, I can't help but notice that there's an important sense in which that *does not really matter*. For, thus far Egan has only shown that *cognitive contents* satisfy the relevant desiderata. But cognitive contents, the deflationary account mantains, *are not really there*. Only mathematical contents are. So it seems that, in order for the deflationary account to succeed, we must show that *mathematical* contents satisfy **(1)** to **(5)**.

And (spoiler alert), they don't seem able to satisfy **(1)** to **(5)** - or so I shall now argue.

3 - The indeterminacy of mathematical contents

Do mathematical contents satisfy desiderata **(1)** to **(5)** - or at least the *relevant* desiderata amongst them ?

Consider first **(4) naturalism**. Mathematical contents are either natural or non natural. If they are natural, **(4)** is satisfied. If they are not, they fail to meet all the desiderata, and so Egan's account falls short of her own standards of adequacy. And this failure matters. Unlike cognitive contents, mathematical contents are not quarantined in an "informal gloss". They are "really there": they're tokened within computational systems, and they're essential ingredients of our cognitive-scientific explanations. Their non-naturality *does threaten* the naturalistic credentials of cognitive science. This time **(4)** *cannot* be ignored. When it comes to the deflationary account theoretical stability, then, mathematical contents *better be* natural.

So, are mathematical contents natural? Egan's answer is a bit confusing. In some passages, she seems to *deny* that mathematical contents are natural (Egan 2014: 213). It's not hard to see why: "classic" content naturalizing relations all have a hard time accounting for representations of abstract and non existing targets.[6] Moreover, the deflationary account should be an *alternative* to "classic" ones, so it cannot rely on "classic" ones to naturalize mathematical contents. But it is hard to think that this is Egan's "official" view, at least insofar adopting this view *amounts to* conceding that **(4)** is not met, and thus that the deflationary account does not satisfy the relevant desiderata.

Perhaps this is why, in other passages, Egan (2014: 117, 119) seems to suggest that mathematical contents *are* naturalized by a minimalistic form of interpretational semantics. Her suggestion seems to be that the vehicle types (identified by f_R) represent the mathematical contents they represent *because* there is an interpretation function f_I associating vehicle types and contents in a one-to-one fashion. The vehicle types identified by f_R represent the mathematical contents they represent *in virtue of* the fact that they can be interpreted as values and arguments of F via f_I. Whilst naturalistic in the relevant sense[7], this approach fails to satisfy other relevant desiderata.

First, such an account leaves us in the dark about f_R. How are the relevant vehicle types identified? Usually representations are type identified by their contents (cf. Egan 2012: 256). But, clearly, that procedure is not available to us now, given that we want f_I to yield yet-uninterpreted vehicle tokens their mathematical contents - the relevant tokens must (logically) be type identified *before* f_I operates on them. So, in order for Egan's minimalistic interpretational semantics to get off the ground, we need an account of the realization function f_R, spelling out how the relevant types are identified. And such an account must satisfy certain relevant desiderata in its own right. An account of f_R must, for example, avoid pancomputationalism. It is remarkably easy to construe any

[6] Accounts of content based on abstract similarities have less problems in this regard: *abstract* similarities with *abstract* structures are easy to come by - indeed, perhaps *too* easy to come by, as often these accounts have to find some means to avoid content underdetermination (e.g. Cummins 1996). Needless to say, Egan does not hold that these means are successful.

[7] As Cummins (1989: Ch. 10) noticed, in order for some states to be *interpretable as* representing, there is no need of *someone actually interpreting* them. A system S may be interpretable as computing F even if no-one actually interprets it as such.

physical system as computing something (Putnam 1988; Searle 1992; Copeland 1996; Scheutz 1999). A good account of f_R will not allow for such construals: given how *cheap* possible interpretations are (see below), such an account would entail a form of pan-representationalism, leading to a violation of **(5)** and **(3) empirical adequacy**.[8] Moreover, an account of f_R must be such that it identifies *only one* set of computational vehicle types for each computational system. Otherwise, a system's computational identity would be unclear, and we would not be able to say whether a system S computes a function F *rather than* a different function F*. Lastly, a good account of f_R must be naturalistic, so as to avoid problems with **(4)**.

Secondly, even with an appropriate account of f_R at hand, the problem of content indeterminacy would loom large (Cummins 1989: 100-102). Given a system S and a well defined set of vehicle types (i.e. a well defined f_R), it will typically be possible to put them in a one-to-one correspondence with *multiple* sets of argument-value pairings. Indeed, for every device S computing a function F we can *always* build up some *ad hoc* function F* under which S is interpretable while keeping f_R constant. Take, for example, an imaginary device S and a f_R such that given f_R S can be interpreted as computing a limited form of addition: the inputs can be interpreted as numbers ranging from 1 to 9 and the outputs can be interpreted as numbers ranging from 2 to 18. Now, the same device, under the same f_R, can be interpreted as computing a function (isomorph to addition) from the first nine US presidents to the set of presidents from Adams (2nd president) to Grant (18th president). The same sets of states can be interpreted as realizing the addition function F(7;9)=16 *or* a function F*(Jackson;Harrison)=Lincoln.[9] So, given f_R, S is interpretable under *at least* two functions. But what do the vehicles tokened inside S

[8] Couldn't this problem be avoided by limiting the scope interpretability to the systems studied by cognitive science? That would restrict the number of systems to which the present account assigns contents, thereby avoiding the problems with **(5)** and **(3)**. But the move is ineffective, for current cognitive science studies *all sorts of systems*, including plants (Calvo *et al* 2020), Bacteria (Lyons 2015), subcellular mechanisms (Yakura 2019) and even certain materials (McGivern 2019, Tripaldi 2022). Surely saying that these systems represent *counts* as a commitment to panrepresentationalism.

[9] Of course, they're respectively the 7th, 9th and 16th US presidents. Notice also that such alternative interpretations *are legion*: we can always also interpret S as computing over presidents (or monarchs, or whatever) of *any* country!

really represent, the arguments and values of F or the ones of F* (or both)? There seems no *principled* way to choose. Therefore **(1)** and **(2)** fail to obtain.

There seems also to be a further problem. Egan seems to espouse a *semantic* account of computation: for S to compute F, S *must* represent the arguments and values of F; that is, certain relevant mathematical contents (cf. §2). On this account, S computes F *in virtue of* the fact that S represents the arguments and values of F. But now notice that, in order for Egan's account of content to work as a form of interpretational semantics, the dependency relation between computation and representation must be reversed. If Egan's account is a form of interpretational semantics, then clearly the identification of the relevant computational state types must *come logically prior to* and *constrain* any mathematical content endowing interpretation. Thus S represents what it represents *partially in virtue of* what it computes; that is, in virtue of the relevant computational state types tokened within, and constraining the interpretability of, S. Simplifying: on the semantic account of computation Egan espouses, S computes what it computes *in virtue of* what it represents. But on the interpretational semantics Egan seemingly espouses, S represents what it represents *in virtue of* what it computes.

Cummins (1989: 93) argued this tension should be solved by prioritizing computation: according to Cummins, we should say systems represent *because* they compute.[10] If we transpose this move in Egan's account, we gain several boons. First, we solve the problem above. Secondly, a suitably robust account of physical computation (and computational implementation) may restrict the number of systems that compute, thereby avoiding pancomputationalism, and thus helping to satisfy **(5) no pan-representationalism** and **(3) empirical adequacy**. Further, a suitably robust account may enable us to say that each computing system S computes few - ideally one and only one - well determined function F. In this way, it will help us restrict the *admissible* interpretations of each

[10] What if we *do not* trust Cummins and prioritize representation? Answer: then we need a substantial account of what makes certain states represent mathematical contents, and we're back to square one.

system, helping make their mathematical contents appropriately determined. Thus, it would be a step towards satisfying **(1)** and **(2)**, and thus **(3)**.

So, the question now is: *which* account of physical computation will deliver these boons? The answer, I fear, is "none"; for, accounts of physical computation can be clustered together in three big families of approaches (semantic, "mapping", and mechanistic; see Piccinini 2015; Piccinini and Maley 2020), and there are very general reasons to believe no approach *can, in principle* provide us these boons.

Consider, first, *semantic* approaches. Despite their variety[11] they all agree that representation is *necessary* for computation. This makes them unsuited to solve the problems of the deflationary account. For, presumably, representations require *contents*. If the content required is cognitive content, then the deflationary account would simply be false: being necessary for computation, cognitive contents would not be just a facultative gloss over real cognitive-scientific explanations. But if the content required is mathematical content, then the account would be circular: our account of physical computation would presuppose the kind of well determined and well distributed mathematical contents we'd like it to deliver.

Consider now "mapping" approaches. In general, they claim a system S implements a computational device *C* computing a function F (minimally, the transition function of *C*) just in case the physical state transition of S and the computational state transition of *C* "march in step", meaning that there is a one-to-one mapping *I* from a relevant subset of states of S onto the states of *C*, and, for all state transitions *c′* → *c″* of *C*, S transitions for *s′* to *s″* only if *I*(*s′*)=*c′* and *I*(*s″*)=*c″*. This is the *necessary* condition all mapping accounts share. If this condition is *also* taken to be sufficient, one reaches the "simple" mapping account (cf Godfrey-Smith 2009). Otherwise, one could robustify the account adding further necessary

[11] For a sample, see (Fodor 1975; O'Brien and Opie 2008; Shagrir 2001; Maley 2021)

conditions. These vary from account to account, and won't matter here (see Piccinini and Maley 2021 for a survey).

Two general reasons impede mapping approaches to deliver the desired boons. One is *computational indeterminacy*, which I will discuss below, when dealing with mechanistic approaches. The other is that *all* mapping approaches entail a limited form of pancomputationalism - they all entail that each and every physical system S implements an inputless finite state automaton C computing the identity function (cf Chalmers 1995; 2011).[12] Now, while this form of limited pancomputationalism need not be fatal for mapping accounts and may even be successfully dealt with in various ways (see Orlandi 2018; Sprevak 2019; Schweitzer 2019 for discussion), it poses a large problem when it comes to using "mapping" approaches to physical computation to deliver mathematical contents. For, if systems represent mathematical contents because they compute, *and all systems compute something*, then all systems represent some mathematical contents. And this is a form of panrepresentationalism. Thus, **(5) No panrepresentationalism** is not satisfied. But since **(5)** is a prerequisite for **(3) Empirical adequacy**, **(3)** fails too. But, as indicated in §2, **(3)** is central to the deflationary account. So, "mapping" approaches are not an option.

Consider, lastly, *mechanistic* approaches.[13] These approaches apply insights from (neo-)mechanist philosophy of science to unravel the nature of computational implementation. Roughly, they claim that a physical system implements a computational device only if it is a *mechanism with the function*[14,15] *to compute* (see Miłkowski 2013; Piccinini 2015). Roughly put, a mechanism of a phenomenon is a set of spatiotemporal components performing certain functions and having certain spatiotemporal relations,

[12] Alternatively: let C* be an inputless finite state automaton with a single state *x*. Let its state transition function F* be F*(*x*)=*x*. Lastly, let the mapping *I* be a mapping from all the states of any system S to *x*. Clearly given this mapping, any physical system S implements C*, and so computes F*.

[13] Assuming, for the sake of discussion, that they are compatible with the deflationary account. They may not be (cf. Egan 2017).

[14] I will leave the relevant notion of function unspecified because (a) it's not relevant for my argument and (b) which notion to use is a contested matter (cf Miłkowski 2013; Piccinini 2015) on which I need not take a stance.

[15] Dewhurst's (2018b) rejection of this functional constraint is an exception. For the (quite limited) purposes of this paper, I will only note that this leaves the proposal open to the kind of pancomputationalism-connected problems of "mapping" approaches.

such that they *constitute* the phenomenon under investigation (cfr. Piccinini 2010: 285). "Computing" is here understood as the manipulation of digits according to rules. Digits may be thought of as the minimal computationally-salient states manipulated by a device, which may be concatenated to yield more complex computationally salient states. The rule according to which a mechanism yields digits as output when "fed" some digit determines the mechanism's computational identity. Importantly, such a rule must be *medium-independent*: it must be sensitive only to the degrees of freedom of digit types, while ignoring any other feature of their tokens.

Now, being a robustified version of the "mapping" account (cf. Piccinini 2015), the mechanistic approach avoids pancomputationlism. Not every physical system is a computational system in the sense just sketched: for one thing, not every physical system has functions, let alone the function of computing. The mechanistic account thus avoids the problems with **(3)** and **(5)** sketched above.

Yet, the mechanistic approach struggles in identifying the computational identity of certain devices (cf Sprevak 2010; Piccinini 2015: 36-39; 127-130; Dewhurst 2018a, Fresco *et al.* 2021), which prevents it from being able to assign well–determined mathematical contents as required by **(1)** and **(2).** To see the problem, consider a computing device S operating on two digit types "@" and "#". S Takes two digits as inputs yielding one as output according to the following rule: it outputs @ *iff* both inputs are @s; else it outputs #. **Table 1** below summarized S's behavior.

Input$_1$	**Input$_2$**	**output**
@	**@**	**@**
@	**#**	**#**
#	**@**	**#**
#	**#**	**#**

Caption: **Table 1: The input-output table of S**

Table 1 looks similar to the truth table of the *logical conjunction*: a function from (pairs of) truth values to truth values. It is thus natural to think @s represent the truth value *true* and #s represent the truth *value* false. It thus seems that the mathematical contents carried by @s and #s are well determined. But the impression is misguided. Let @s carry the mathematical content *false* and #s carry the mathematical content *true*. Now **Table 1** looks like the (upside-down) truth table of the *inclusive disjunction*. So, do "@"s represent the truth value *true*, *false* or both? It seems there's no *principled* way to answer.

Worse still, the mathematical contents of @s and #s may be undetermined *even when the F S computes is determined.*[16] Consider a system S* displaying the computational behavior summarized in **Table 2**:

Input	**Output**
@	**#**
#	**@**

Caption: **Table 2: The input-output table of S***

S* takes one digit as input yielding one as output. If the input is a @, it yields a # and *vice versa*. It is natural to say S* computes the *logical negation function*, and there seems to be no other interpretation around.[17] But saying S* computes the negation function it is yet not enough to determine whether @s represent the truth value true or the truth value false. The relevant mathematical contents are left *undetermined*. So, **(2) determinacy** is not met, and since **(2)** is not met, **(1) misrepresentation** is not met too.[18]

An obvious objection to my point is this: whilst looking at S alone does not determine the mathematical contents of @s and #s, looking at how S

[16] Thus, attempts to restore computational determinacy such as the ones in (Dewhurst 2018a; Fresco and Miłkowski 2021; Dothey and Dewhurst 2022) are not going to solve the problem of indeterminacy I'm pointing at.

[17] I'm assuming that we have some compelling independent reason to interpret the device as a logic gate. In fact, as a reviewer correctly noticed, without such an assumption, the computational identity of S* would not be determined; e.g. if "@" represents 2, and "#" represents ½, then the device is computing $F(x) = 1/x$.

[18] Note, also, that thus far I've assumed that the relevant digits are given and that their identity can be easily defined. But that is not the case, and there's an indeterminacy problem there too, see (Papayannopolus *et al* forthcoming).

is embedded in a larger computational device, and how it cooperates with other computational mechanisms, will.

The objection fails on several grounds. First, *even if* looking at how S contributes to a large computational system were sufficient to determine the mathematical contents of @s and #s, S *need not* be embedded in a larger computational system in order for it to compute. So, albeit the move may determine the mathematical contents of some computational systems; namely the ones embedded in larger computational systems, it fails to determine mathematical contents *in general*. For example, the mathematical contents of *individual* logic gates would remain indeterminate. A, perhaps limited, problem with content determinacy would still persist.

Second, observing how S is embedded in a larger system *does not* yield well determined mathematical contents. Consider a system M constituted concatenating S and S* as follows: S takes two inputs, yields an output that function as S* input, and then S* yields the final output. The behavior of M is summarized in **table 3**:

Input$_1$	**input$_2$**	**S**	**S***
@	@	@	#
@	#	#	@
#	@	#	@
#	#	#	@

Caption: **Table 3: The input-output table of M**

If @s represent *false* and #s represent *true*, M computes the *nor* (not or) function. Under the opposite assignment of truth values, M computes *nand* (not and). The mathematical contents in M are thus as undetermined as the ones in S and S*. Note that, *in principle*, no amount of added computational machinery will make the mathematical contents of @s and #s determinate. It will *always* be possible to "swap" the truth values and see the entire device as computing a function. Maybe the function will not be interesting or useful. But it would still be a *function* - and so we are *still* left to choose among two competing assignments of mathematical contents.

So, the minimalistic form of interpretational semantics Egan endorses does not really seem defensible: just like "classic" account of representations, it leaves (mathematical) contents indeterminate and unable to misrepresent.

Is there any other option? Maybe we could hope in *semantic primitivism*; i.e. the view that there are natural primitive (non-analyzable) semantic facts concerning mathematical contents (cf. Burge 2010). But semantic primitivism fails to satisfy the relevant desiderata **(1)** to **(5)**. In fact, it fails to satisfy **(4)**. Semantic primitivism *asserts* that content is natural in the relevant sense. But it does not offer an *account* of content in more basic terms as required by **(4)** (cf. Piccinini 2015: 35). Indeed, it can't *coherently* offer such an account, for any such account *entails* that content is not primitive. And even leaving this problem aside, it seems that primitive semantic facts are *epiphenomenal*: they make no difference to the computational behavior of a system. Recall system S, whose computational behavior is summarized below in **table 1 bis**:

Input$_1$	**Input$_2$**	**output**
@	@	@
@	#	#
#	@	#
#	#	#

Caption: **Table 1 bis: The input-output table of S (Again)**

Suppose that, as a matter of primitive semantic fact, S is an *And Gate*: it computes the *conjunction* function. So, as a matter of primitive semantic fact, @s represent *true* and #s represent *false*. Still, I *could* use S as an *Or Gate* in an appropriate system. I could even build a system for the purpose of using S as an *Or Gate*. It seems that the "primitive semantic facts", whilst sufficient to give us well determined mathematical contents, *make them irrelevant* to the actual functioning of a device. They lose their explanatory power. And, thus robbed of their explanatory power, they start to look like a simplificatory *gloss* summarizing the physical/causal behavior of physical systems.

So, at present, there seems to be no satisfactory way to naturalize mathematical contents. It seems that all avenues to naturalization force us to pay too high of a price. To naturalize mathematical contents is to forego *many* of the desiderata in the **(1)-(5)** list. And to keep them non-naturalized *is* to taint the naturalistic credentials of cognitive science.

But the troubles for the deflationary account are not over.

4 - Deflating the explanatory power of deflated representations

The deflationary account claims cognitive scientific explanations consist *only* in the function theoretic characterization of a device plus an ecological component. I want to argue that this is not the case: sometimes, *cognitive* contents are necessary too. Or, more prudently: mathematical contents and ecological component are not always sufficient to explain.

To see why that is the case, I must first clarify how mathematical contents are supposed to explain. On the one hand, it seems that mathematical contents explain allowing us to *predict* and *postdict* the behavior of a computational system. If I know that S computes F, I know how S *would* behave were it to receive an input *i;* namely since F(*i*)= *o*, S will produce *o*. On the other hand, mathematical contents allow us to describe the behavior of the system in terms of successes and failures, and to account for these successes and failures. If S computes F, S *fails* every time it does not output *o* =F(*i*) in response to *i;* and we can say S's failure is due to it having *miscomputed* F. In Egan's own words:

> "In attributing a competence to a physical system—to add, to compute a displacement vector, and so on—function-theoretic models support attributions of correctness and mistakes. Just as the normal

> functioning of the system—correctly computing the specified mathematical function—explains the subject's success at a cognitive task in its normal environment, so a malfunction explains its occasional failure. [...] One's hand overshooting the cup because the motor control system miscalculated the difference vector is a perfectly good explanation of motor control failure" (Egan 2017: 158)

Explanations of successes and failures are always explanations of *patterns* of successes and failures (as amply clarified in Gładziejewski and Miłkowski 2017; Shea 2018). This can be easily illustrated elaborating on Egan's example above: the hand overshoot because the device outputted a vector *v** larger than of *v* (the one it should have outputted). And here's the relevant pattern of failures: the *larger v**, the more severe the overshoot. And the larger *one specific component* of *v**, the more severe *the overshoot in a specific direction*. And where *v** smaller than *v*, then the system would not have overshoot: it would have under-shoot. That's the relevant pattern of successes and failures explained by mathematical contents.

Crucially, this explanation *requires* a *systematic correlation* between mathematical contents and outcomes: the *larger v**, the *more severe* the overshoot. This correlation may (and it typically will be) more complex and less linear, but it *needs* to be there. If that correlation is absent, then mathematical contents do not do the desired explanatory work, and the deflationary account fails to capture the way in which representations are used in cognitive science. Now, there is at least one case in which such a correlation between mathematical contents and successes and failures seems to be absent. Mathematical contents seem thus unable to play the explanatory role the deflationary account assigns them. Worse still, that role seems to be played by *cognitive* contents.[19] The counterexample is provided by *adversarial examples induced misclassification* (AEIM).

[19] Importantly, as a reviewer noticed, the explanatory role of cognitive contents may be *way more* widespread than the case study below could suggest. For, saying that one of my brain regions miscomputed a function F explains a failure of mine *only if* we take the values and arguments of F to be values *representing* task relevant parameters. Saying, for example, that I miscomputed the square root of 75 *does not* explain my failing to grasp a cup *unless* we don't take "75" to *be*, for example, an estimate of depth. But if we do so, then it's not clear *in what sense* I'm not really representing depth, but only the number 75.

AEIM is a phenomenon concerning *deep classifiers*: a specific class of deep neural networks. These devices have a clear computational profile: they compute a probability distribution over class labels, given an input vector (cf. Buckner 2018; Mitchell 2019; Skansi 2018). Thus their function-theoretic characterization is well known. Simplifying *a lot*, for my purpose here deep classifiers can be considered as the shallow neural networks of the '80s: what changes is just their scale and the number of computational layers. Thus, deep classifiers compute by transforming vectors, spreading activations through successive layers of "neurons". Each neuron yields an output (a vector component) based on the *activation function* it computes. All neurons in a layer thus collectively define the output vector of that layer. That output is then "funneled thought" weighted connections, which modify it proportionally to their weights, thus yielding the input for the next layer. The process is repeated until the last layer (called output layer), is reached.

Deep classifiers richly trade in mathematical contents. The weighted connections store the *parameters* of the model the classifier uses to classify its inputs. Neurons *compute* activation functions. They have (numeric) *bias*. The network also represents its own learning rate - a number "telling" the network how much to update its parameters. All these things, as well as the network topology (number of neurons and connections and how they're disposed) are the *hyperparameters* of the model, and they influence the classification (and thus the computation) too.

Suppose now a deep classifier *C* correctly classifies an input vector *v*. An *adversarial example* to *C* is a slightly modified version *v** of *v* that *C* misclassifies with very high confidence, *despite the fact that v and v* are identical to human eyes*. If, for example, *v* and *v** are images, their difference may be of just one pixel (e.g. Su *et al.* 2019).[20] And, of course,

[20] This is a *huge* simplification. An "alternative" family of adversarial examples is constituted by "senseless" (to human) vectors which the machine classifies with high confidence (cf. Nguyen *et al.* 2015). See (Yuan *et al.* 2019) for an up to date survey on adversarial examples.

when *v** fools *C*, we witness an instance of adversarial-example induced misclassification (AEIM).

AEIMs *call* for an explanation for a number of reasons. Deep classifiers are some of our best neurocognitive models of classification, *especially* when it comes to human visual classification (Yamins and DiCarlo 2016; Rajalingham *et al.* 2018). But *we are immune from AEIMs*![21] There's thus a significant difference between us and some of our best models of us. Understanding what this difference is is pivotal *both* to build better models *and* to understand ourselves.

Yet the explanatory schema the deflationary account proposes seems unable to account for AEIMs. For one thing, the discovery of adversarial examples and AEIMs was a surprise (cf. Szegedy *et al.* 2013). It was not expected (nor predicted) given the function-theoretic characterization of deep classifiers. So, they were *not* predicted given the function-theoretic characterization of deep classifiers. Contrary to what the deflationary account claims, mathematical contents were not enough to predict the relevant behavior of deep classifiers.

Further, *in spite of the fact that we do possess all the relevant function-theoretic knowledge about deep classifiers*, AEIMS still stands in need of an explanation. We don't yet know why such small perturbations in the input lead to such big changes in the outputs. And in fact, there is seemingly *no correlation* between mathematical contents and the outputs produced by AEIMs. Given that such a correlation is necessary in order for mathematical contents to explain, then it should be concluded mathematical contents do not explain (in the relevant sense).

To see why no such correlation holds, notice that adversarial examples are *transferable*. If an adversarially perturbed vector *v** fools a classifier *C*, then *v** is likely to fool also a different classifier *C* in the exact same way*. AEIMs are thus in an important way not random. There is a clear *pattern* in the failure they induce - a pattern that *prima facie*

[21] At least, in normal conditions. Time pressured humans *may* be fooled by adversarial examples (Elsayed *et al.* 2018).

looks like a primary explanatory target. And yet the pattern stands in no discernible correlation with mathematical contents: *ceteris paribus*, identical errors should correlate with identical (or at least relevantly similar) mathematical contents. And yet, when it comes to deep classifiers, *identical* errors correlate with different mathematical contents, for different classifiers are *bound to* have different mathematical contents. Indeed, not only adversarial examples are transferable across classifiers with different *hyperparameters* (such as different topologies, number of layers, biases, learning rate or activation functions, see Szegedy *et al* 2013), even architecturally identical classifiers trained on the exact same training set with the same training regime will encode different parameters (cf Churchland 1992: 177-178), thereby representing different mathematical contents. Thus, the situation looks like this: on the one hand, a tight and clear pattern of AEIMs; on the other, mathematical contents that appear to vary *ad libitum*. This clearly prevents the two from correlating in any intelligible way.

Worse still (for the deflationary account), the explanations currently proposed *massively* involve cognitive contents. Consider the following two proposed explanations.[22]

Proposed explanation #1. Ilyas *et al.* (2019), start by mathematically defining *features* (the properties guiding classification). Then they mathematically define a subclass of features: *useful features* (i.e. features that *correctly* guide the classification). This subclass is then (again, mathematically) divided into two disjoint subsets: robust and non-robust. Robust useful features correctly guide classification even *after* the adversarial perturbation has been applied. Non-robust ones *do not*. Thus adversarial induced misclassification is due to the classifier reliance on *non-robust useful features*.

Proposed explanation #2: (Zhou and Firestone 2019) tested human subjects in a variety of classification tasks using adversarially perturbed

[22] I use them *just* as examples. I do not want to imply they're the only, or even the best, explanations. See also (Engstrom *et al*. 2019; Bucker 2020) for discussion.

images, asking the human participants to pick up the label they think a machine would assign to the image. Strikingly, they found that in all the experiments (using a variety of adversarially perturbed images in a variety of experimental paradigms) participants were able to choose "like a deep classifier" with a percentage of success well above chance. This led Zhou and Firestone to suggest that adversarial examples induce misclassifications because networks do not discriminate between appearing *like* something and appearing *like being* something (e.g. a plush toy might appear *like* a tiger, but it does not appear *to be* a tiger).

The explanation offered by Zhou and Firestone is clearly based around cognitive contents: distal targets being represented *as similar to* or *as being* other distal targets. The explanation offered by Ilyas and colleagues mentions cognitive contents too, although in a roundabout way. In fact, their mathematical definition of features is intended to capture the "folk" definition of features as representations of salient distal properties (cf Hinton 2014; Olah *et al*. 2018). Further, robustness and non-robustness are defined relative to *a human-selected notion of similarity*. And such a notion is plausibly based on how we represent things *as being alike*.

Notice that these ascriptions *cannot* play a heuristic role in orienting the research for a function-theoretic characterization (Egan 2020a: 46-48). We *do possess* the relevant function-theoretic characterization. Deep classifiers are not *objets trouvé* whose computational profile must be discovered. They're artificial systems we create for the purpose of computing a mathematical function we already know - in the case at hand, a probability distribution over labels, given an input vector. So, in the case of AEIMs, cognitive contents are not just "heuristic patches" we use while we wait for the relevant function theoretic characterization to come. They must play a *deeper* explanatory role.

One could object that I've been too focused on mathematical contents. Maybe the *ecological component* holds the key to explain AEIMs. Maybe yes, but it is hard to see what the ecological component may be in the case at hand. The only window on the world available to classifiers is their input data. And it seems that it can be altered too without compromising the transferability of AEIMs (cf Szegedy *et al* 2013) - changing training sets *does not* appear to change how classifiers respond to adversarially perturbed vectors.

Alternatively, one might object that I've mischaracterized the explanatory role of mathematical contests. Cognitive contents are said to explain in many ways. Maybe mathematical contents can explain in multiple ways too. A popular way in which cognitive contents are said to explain is by *being causes* of a system's behavior (e.g. Dretske 1988; O'Brien 2015). But the deflationary account prevents mathematical contents from playing this explanatory role. On the deflationary account, contents have no causal powers (Egan 2014; 2020a). Another popular way in which contents are said to be explanatory powerful is that of allowing us to grasp patterns we would otherwise fail to grasp (e.g. Dennett 1991). The possible physical manifestations of, say, my request to open the window is unruly and possibly open endedly disjunctive. I can request to open the windows by asking it. Or by sending an email to the person closer to the window. Or by making gestures. To explain why, in all these cases, a person reacted by closing the window, the best thing to do is to appeal to the *content* of these gestures/mail/soundwaves. But mathematical contents cannot play this explanatory role either: the relation between them and vehicle types is one-to-one (see point **(ii)** in §2). So there's no pattern holding among contents that is not *also* a pattern holding at the level of vehicles.

Notice, to conclude, that the problem raised here is *independent* from the naturalistic credentials of mathematical contents. Even if mathematical contents were to be naturalized, the explanatory problems of the deflationary account would not be solved. *Pace* the deflationary account, cognitive contents do not seem to be just a gloss. Therefore, not **(3)**: the deflationary account is not empirically accurate.

5 - Concluding remarks

In this paper, I raised two distinct challenges to Egan's deflationary account. But why care about the problems the deflationary account faces? I think there are two reasons to care about them.

One has to do with a recent, and still young, movement trying to account for representations in terms of computation (Coelho Mollo 2021; Piantadosi 2021). There is something *right* in Egan's view: the repeated and constant failure of "classic" accounts of content *does* call for alternative accounts of representations and content. And, whilst still at present quite amorphous and disorganized, that movement is trying to answer this specific call in a way that, at least in spirit, is highly sympathetic to Egan's deflationary account - after all, her account too makes computation central to representation. By pointing out some problems with mathematical contents I hope to help this movement avoid some pitfalls. Although in a purely negative way, I'm still helping that movement to grow. Or at least, that is one of the reasons that led me to write this paper.

The other has to do with one of the most discussed neurocomputational theories right now, namely predictive processing/the free-energy principle (for introductions, see Hohwy 2013; Clark 2016). There's a huge debate concerning the representational credentials of the theory (see Sims and Pezzulo 2021). If the arguments I've provided here are correct, the anti-representationalist side has a potent weapon in its hands. For, a number of prominent accounts of representation within predictive processing elaborate upon Egan's cognitive/mathematical content distinction, claiming that cognitive contents are grounded upon, and determined by, mathematical ones (most explicitly the point is made in Wiese 2017; 2018; Ramstead *et al* 2020). But if the arguments presented above are on the right track, then mathematical contents are non-natural and indeterminate, in a way that makes the "cognitive" contents non-natural and indeterminate, in a way that largely favors non-representational interpretations of predictive processing.

References

Artiga, M., & Sebastian, A. S. (2018). Informational theories of content and mental representation. *Review of Philosophy and Psychology*, *11*, 613-627.

Backer, B., Lansdell, B., Kording, K. (2021). A philosophical understanding of representations for neuroscience. ArXiv: 2102.06592.

Buckner, C. (2019). Deep learning: a philosophical introduction. *Philosophy Compass*, *14*(10), e12625.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artifacts for deep learning. *Nature Machine Intelligence*, *2*, 731-736.

Burge, T. (2010). *The Origins of Objectivity*. New York: Oxford University Press.

Calvo, P., *et al.* (2020). Plants are intelligent, here's how. *Annals of Botany*, *125*(1), 11-28.

Chalmers, D. J. (1995). On implementing a computation. *Minds and Machines*, *4*, 391-402.

Chalmers, D. J. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, *12*(4), 325-359.

Churchland, P. (1992). *A Neurocomputational Perspective*. Cambridge, MA.: The MIT Press.

Clark, A. (2016). *Surfing Uncertainty*. New York: OUP.

Coelho Mollo, D. (2020). Content pragmatism defended. *Topoi*, *39*(1), 103-113.

Coelho Mollo, D. (2021). Why go for a computation-based approach to cognitive representations. *Synthese*, *199*(3-4), 6875-6895.

Copeland, J. (1996). What is computation?. *Synthese*, *108*(3), 335-359.

Cummins, R. (1989). *Meaning and Mental Representation*. Cambridge, MA.: The MIT Press.

Cummins, R. (1996). *Representations, Targets, and Attitudes*. Cambridge, MA.: The MIT Press.

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

Dewhurst, J. (2018a). Individuation without representation. *The British Journal for thePhilosophy of Science*, *69*(1), 103-116.

Dewhurst, J. (2018b). Computing mechanisms without proper functions. *Minds & Machines*, *28*(3), 569-588.

Doherty F. T. & Dewhurst, J. (2022). Structuralism, Indiscernibility and physical computation. *Synthese*, *200*(3), 1-26.

Dretske, F. (1986). Misrepresentation. In Bodgan R. (Ed.). *Belief: Form, Content and Function*. (pp. 17-36). New York: Oxford University Press.

Dretske, F. (1988). *Explaining Behavior*. Cambridge, MA.: The MIT Press.

Egan, F. (1999). In defense of narrow mindedness. *Mind&Language*, *14*(2), 177-194.

Egan, F. (2010). Computational models: a modest role for content. *Studies in History and Philosophy of Science Part A*, *41*(3), 253-259.

Egan, F. (2012). Representationalism. In E. Margolis, S. Samuels, & P. Stich (Eds.), The Oxford handbook of philosophy of cognitive science (pp. 250–272). Oxford University Press

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, *170*(1), 115-135.

Egan, F: (2017). Function theoretic explanation and the search for neural mechanisms. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (pp. 145-163). New York: Oxford University Press.

Egan, F. (2019). The nature and function of content in computational models. In M. Sprevak, M. Colombo, (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 247-258). New York: Routledge.

Egan, F. (2020a). A deflationary account of mental representations. In J. Smortchkova, K. Dolega, T. Schlicht (Eds.), *What Are Mental Representations?* (pp. 26-54), New York: Oxford University Press.

Elsayed, G. *et al.* (2018). Adversarial examples that fool both computer vision and time-limited humans. *arXiv preprint*, 1802.08195.

Engstrom, L., *et al.* (2019). A discussion of 'adversarial examples are not bugs, they are features'. *Distill*, https://distill.pub/2019/advex-bugs-discussion/

Fodor, J. (1975). *The Language of Thought*, Cambridge, MA.: Harvard University Press.

Fodor, J. (1990). *A Theory of Content and Other Essays*. Cambridge, MA.: The MIT Press.

Fresco, N., *et al.* (2021). The Indeterminacy of Computation. *Synthese*. https://doi.org/10.1007/s11229-021-03352-9.

Fresco, N., & Miłkowski, M. (2021). Mechanistic computational individuation without biting the bullet. *The British Journal for the Philosophy of Science*, *72*(2), 431-438.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: causally relevant and distinct from detectors. *Biology and Philosophy*, *32*(3), 337-355.

Godfrey-Smith, P. (2009). Triviality Arguments Against Functionalism. *Philosophical Studies. 145*(2): 273–295.

Hinton, G. (2014). Where do features come from?. *Cognitive Science*, *38*(6), 1078-1101.

Hohwy, J. (2013). *The Predictive Mind.*New York: OUP

Hutto, D., & Myin, E. (2013). *Radicalizing Enactivism*. Cambridge, MA.: The MIT Press.

Ilyas, A., *et al*. (2019). Adversarial examples are not bugs, they are features. *arXiv*: 1905.02175

Lyon, P. (2015). The cognitive cell: bacterial behavior reconsidered. *Frontiers in Microbiology*, 6:264.

Maly, C. (2021). The physicality of representation. *Synthese, 199*, 14725-14750.

Marr, D. (1982). *Vision*. Henry Holt: New York.

McGivern, P. (2019). Active materials: minimal models of cognition? *Adaptive Behavior*, *28*(6), 441-451.

Miłkowski, M. (2013). *Explaining the Computational Mind*. Cambridge, MA.: The MIT Press.

Millikan, R. G. (1984). *Language, Thought, and other Biological Categories*. Cambridge, MA.: The MIT Press.

Mitchell, M. (2019). *Artificial Intelligence: a Guide for Thinking Humans*. London: Penguin.

Nguyen, A., *et al*. (2015). Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (pp. 427-436).

O'Brien, G. (2015). How does the mind matter? Solving the content causation problem. In T. Metzinger, J. M. Windt (Eds.), *Open MIND*: 28(T). Frankfurt am Main, The MIND Group. https://doi.org/10.15502/9783958570146.

O'Brien, G., & Opie, J. (2004). Notes towards a structuralist theory of mental representations, in H. Clapin; P. Staines & P. Slezak (eds.), Representation in Mind: New Approaches to Mental Representaion (pp. 1-20). Oxford: Elsevier.

O'Brien, G., & Opie, J. (2008). The role of representation in computation. *Cognitive Processing*, *10*(1), 53-62.

Olah, C., *et al.* (2018). The building blocks of interpretability. *Distill*, *3*(3): e10. https://distill.pub/2018/building-blocks/

Orlandi, N. (2018). Perception without computation? In M. Sprevak, M. Colombo (eds), *The Routledge Handbook of the Computational Mind* (pp. 410-423). New York: Rutledge

Orlandi, N. (2020). Representing as coordinating with absence. In J. Smortchkova, K. Dolega, T. Schlicht (Eds.), *What Are Mental Representations?* (pp. 101-135), New York: Oxford University Press.

Papayannopoulos, P., *et al.* (*forthcoming*). On two different kinds of computational indeterminacy. *The Monist*. Preprint at: http://philsci-archive.pitt.edu/19622/

Piantadosi, S. T. (2021). The computational origin of representation. *Mind and Machines*, *31*, 1-58.

Piccinini, G. (2010). The mind as neural software? Understanding functionalism, computationalism and computational functionalism. *Philosophy and Phenomenological Research*, *81*(2), 269-411.

Piccinini, G. (2015). *Physical Computation: a Mechanistic Account*. New York: OXford University Press.

Piccinini, G., & Maley, C. (2021). Computation in physical systems. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (summer 2021 edition),
https://plato.stanford.edu/archives/sum2021/entries/computation-physicalsystems/ last accessed 19/06/2021

Putnam, H. (1988). *Representation and Reality*. Cambridge, MA.: The MIT Press.

Rajalingham, R. *et al.* (2018). Large-scale, high-resolution comparison of the core visual objects recognition behavior of humans, monkeys and state-of-the-art deep artificial neural networks. *Journal of Neuroscience, 38*(33), 7255-7269.

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

Ramstead, M. D. *et al.* (2020). Is the free-energy principle a formal theory of semantics? *Entropy, 22*(8), 889.

Scheutz, M., (1999). When physical systems realize functions…. *Minds and Machines*, *9*(2), 161-196.

Schweitzer, P. (2019). Triviality arguments reconsidered. *Minds and Machines*, *29*(2), 287-308.

Searle, J. (1992). *The Rediscovery of the Mind*. Cambridge, MA.: The MIT Press.

Shagrir, O. (2001). Content, computation and externalism. *Mind, 110*, 477-500.

Shea, N. (2018). *Representation in Cognitive Science*. New York: Oxford University Press.

Sims, M., & Pezzulo, G. (2021). Modeling ourselves: what the free energy principle reveals about our implicit notion of representation. *Synthese, 199*(3), 7801-7833.

Skansi, S. (2018). *Introduction to Deep Learning: from logical calculus to artificial intelligence*. Springer.

Sprevak, M. (2010). Computation, individuation and the received view on representation. *Studies in History and Philosophy of Science Part A*, *41*(3), 260-270.

Sprevak, M. (2011). Review of William M. Ramsy *Representation Reconsidered. The British Journal of Philosophy of science. 62*(3) 669-675.

Sprevak, M. (2019). Triviality arguments about computational implementation. In M. Sprevak, M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 175-191). New York: Routledge

Su, J. *et al.* (2019). One pixel attacks for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation. 23*(5), 828-841.

Szegedy, C., *et al.* (2013). Intriguing properties of neural networks. *arXiv preprint*: 1312.6199.

Tripaldi, L. (2022). *Parallel Minds*. Cambridge, MA.: The MIT Press.

Wiese, W: (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, *16*(4), 715-736.

Wiese, W. (2018). *Experienced Wholeness*. The MIT Press.

Webb, B. (2006). Transformation, encoding and representation. *Current Biology*, 16(6), R184-R185.

Yakura, H. (2019). A hypothesis: CRISPR-Cas as a minimal cognitive system. Adaptive Behavior, 27(3), 167-173.

Yamins, D., & DiCarlo J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356-365.

Yuan, X. *et al.* (2019). Adversarial examples: attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2805-2024.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature communications*, *10*(1), 1-9.