Troubles with mathematical contents

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

Abstract:

The deflationary account of representations purports to capture the explanatory role representations play in computational cognitive science. To this end, the account distinguishes between mathematical contents, representing the values and arguments of the functions cognitive devices compute, and cognitive contents, which represent the distal states of affairs cognitive systems relate to. Armed with this distinction, the deflationary account contends that computational cognitive science is committed only to mathematical contents, which are sufficient to provide satisfactory cognitive explanations. Here, I scrutinize the deflationary account, arguing that, as things stand, it faces two important challenges deeply connected with mathematical contents. The first depends on the fact that the deflationary account accepts that a satisfactory account of representations must deliver naturalized contents. Yet, mathematical contents have not been naturalized, and I claim that it is very doubtful that they ever will. The second challenge concerns the explanatory power of mathematical contents. The deflationary account holds that they are always sufficient to provide satisfactory explanations of cognitive phenomena. I will contend that this is not the case, as mathematical contents alone are not sufficient to explain why deep neural networks misclassify adversarial examples.

***Keywords**:* Representation, Content, Mathematical contents, Computation, Implementation, Adversarial examples.

***Wordcount**:* 9365, references, abstract and notes included

Declarations:
Founding: Not applicable
Conflict of Interests/competing interests: The author declares no conflict of interests.
Availability of data and Material: Not applicable.
Code availability: Not applicable.
Authors' contribution: Marco Facchin is the sole author of the paper.
Acknowledgments: This essay has been presented at the SIFA 2021 and the AISC online graduate conference. I wish to thank the audience of these conferences for their comments and suggestions. A special thanks goes to Joe Dewhurst, for his extensive and insightful feedback.

1 - Introduction

What are the factors turning a wrinkly clump of specialized cells like a brain into a representational system? "Classic" answers to this question attempts to *naturalize content*; i.e. spell out the answer pointing to some privileged naturalistic relation. Some such accounts center around causal/informational factors (Dretske 1988; Fodor 1990), others appeal to biological functions (Millikan 1984), and others still centered around abstract notions of resemblance (O'Brien and Opie 2004). Yet they all share a problem of *content determinacy*. They fail to specify conditions yielding *unique and well-determined* contents to representations. Causal/informational accounts cannot decide whether distal or proximal objects are represented (cf Artiga and Sebastian 2018). Theories based on biological functions face the terrible *disjunction problem*: the contents they deliver being often unruly and open-endely disjunctive (cf Hutto and Myin 2013, Ch.4). Resemblance-based theories have problems in determining contents too: resemblances are notoriously cheap (Sprevak 2001).

Egan (2014; 2019; 2020a) suggests these failures suggest a change of approach. Rather than striving to naturalize content, we should aim at capturing the role representations play in cognitive science.[1] To reach this goal, Egan's (2014; 2020a) deflationary account identifies two distinct kinds of content: *cognitive and mathematical*. Cognitive contents are contents usually understood (representations of distal worldly targets), and they pertain to a *facultative gloss* layered over cognitive-scientific explanations. They're not "inside" the systems cognitive science studies. Conversely, mathematical contents *really are* "inside" these systems, and are thus genuine components of cognitive-scientific explanations. Yet, they only represent abstract objects; namely the arguments and values of the functions cognitive systems compute.

I want to raise some problems for the deflationary account. Thus, after having introduced it (§2), I will argue that it fails to satisfy the desiderata it sets for itself (§3), and that it falls short of accounting for at least one interesting *explanandum* of contemporary cognitive science: adversarial examples-induced misclassifications (§4). To anticipate the moral I will draw from these objections (§5), I say this: that a thorough examination of the deflationary account might prove useful in two regards. On the one hand, it can help accounts of representations outside the "classical", content-naturalization based, tradition (Cohelo-Mollo 2021; Piantadosi 2021) to avoid some dangerous pitfalls. On the other hand, it is relevant for the currently most talked about neurocognitive theory predictive processing and the free-energy principle. For, many philosophers interested in representations and working within these frameworks have heavily relied on elements borrowed from the deflationary account (Wiese 2017; 2018; Ramstead *et al* 2020). The problem raised here, thus, might (dangerously) propagate to the philosophical literature on predictive processing.

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

2 - Egan's deflationary account of representations

Egan's account is a reaction against "classical", content-naturalization based, accounts of representation. This reaction is motivated by the fact that "classical" accounts fail to meet a number of minimal and (almost) universally accepted *desiderata*[2], namely:

> **(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

Some clarifications. **(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. This is precisely why content determinacy is such a big problem for "classic" accounts.

Though surely a desideratum in its own right, empirical adequacy **(3)** is pivotal for the deflationary account. The deflationary account aims to capture the explanatory role of representations in cognitive science, so it *must* be empirically adequate to succeed. Notice also **(3)** is connected with **(1)** and **(2)**: cognitive scientists ascribe fairly well-determinate contents to representations (e.g. Backer *et al*. 2021). The satisfaction of **(1)** and **(2)** is a prerequisite for satisfying **(3)**.

Condition **(4)** captures the widespread idea that content supervenes on more basic facts and features of the world. Hence, it should be explainable in terms of these more fundamental facts and features. Notice that these features may, but need not, be the causal/informational, teleological or similarity based features mentioned in (§1). Content might, but need not, be naturalized by a "classic" content-naturalizing relation. It may be naturalized by other factors, such as computational implementation (Coelho-mollo 2021; Piantadosi 2021). What matters for **(4)** is only that the factors naturalizing content, whatever they may be, are not *already* semantic or intentional.

Lastly, **(5)** descends from **(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

[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.

trivializes the explanatory power of content, equating representations to mere causal mediators (see Ramsey 2007; Orlandi 2020).

Egan (2019; 2020a) argues that "classic" accounts *always* fail to satisfy **(1)** and **(2)**. Hence, they should be rejected. But what's the shape of the alternative? Since the deflationary account aims at spelling out the role representations play in cognitive scientific explanations, these explanations provide the natural starting point.

Egan (2010; 2014; 2017; 2020) construes cognitive-scientific explanations as *function-theoretic*: they unveil the mathematical function F computed by a cognitive device S. Function-theoretic explanations sit at Marr's (1982) computational level: they mathematically characterize the input-output behavior of a device. And indeed Marr's account of vision provides Egan's paradigmatic example of a function-theoretic representation. According to Marr we explain what retinas - the system S - do in vision by saying they compute a smoothing function F convolving a Laplacian operator with a Gaussian operator. 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 **(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.[3] Mathematical contents are also the explanatory factors highlighted by function-theoretic explanations, of which they are an essential component (Egan 2014: 122-123). Their explanatory power (but more of this in §4) consists in subsuming the

[3] 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.

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. They don't illuminate how computing F contributes to a cognitive capacity. To illuminate this, function theoretic explanations need to be complemented by an *ecological component*: a series of assumptions about the environment that clarifies how computing F contributes to the exercise of a cognitive capacity. The ecological component allows us to "make sense" of F; it makes us 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 to detect sharp changes in illumination value (edges), thereby contributing to vision.

But note how the ecological component has let *cognitive* contents into the picture: computing a smoothing function allows retinas to *detect edges*. Doesn't this mean that retinas represent edges? Egan (2014; 2019: 254; 2020a) answers negatively: ratinas *really* only represent mathematical contents. Yet, Egan concedes, rather than stating the function computed and the ecological component, we can *say for simplicity's sake* that retinas represent edges. It's simpler and easier to understand. Thus, we can "summarize" the job done by the ecological component *via* ascriptions of cognitive contents.

Yet, Egan holds, these ascriptions are *just* ascriptions. Whilst retinas *really detect* edges, they do not *really represent* edges. No edge-representations are *really* tokened within retinas; only representations of the relevant values and arguments are (Egan 2014; 2020a). Indeed, in her view, *no cognitive contents are really ever represented within cognitive systems*. Only mathematical contents are. So, there's no *fact of the matter* about which cognitive contents cognitive systems really represent (cf. Coelho Mollo 2020). Cognitive contents are only ascribed from the outside, based on our pragmatic and explanatory interests - roughly, based on the intuitive grasp they afford over a system's behavior. And thus the cognitive-content based talk is revealed to be just and informal, and strictly speaking facultative, "gloss" over genuine cognitive-scientific explanations (Egan 2014; 2019; 2020a.) [4]

Egan (2014; 2020a) holds that realizing the pragmatic and "glossy" nature of cognitive contents allows us to satisfy the relevant desiderata. We surely *do ascribe* determinate cognitive contents. We don't say that retinas represent edges *or* shadows *or...*;

[4] 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)

we say they represent *edges* - full stop. Thus **(2)**, content determinacy, is satisfied. But **(2)** and **(1)** - misrepresentation - are constitutively connected. Once content is well determined, misrepresentation is entirely possible, and **(1)** is satisfied too. Being built on several case studies, we can expect the deflationary account to be empirically accurate, so **(3)** should be satisfied. Egan claims her account is safe from panrepresentationalims, satisfying **(5)**. After all, on her account cognitive contents are ascribed, and surely we don't ascribe cognitive contents to *everything*. Lastly, the account is not naturalistic, so **(4)** is not met. But, Egan (2020a) claims naturalism 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 under threat. The non-naturality of cognitive contents is quarantined in an "informal gloss" over the theory, and thus does not spread to the latter.

Egan thus holds her account allows cognitive contents to satisfy **(1)** to **(5)**. But, I want to note, there is an important sense according to which, *this does not matter*, at least if the deflationary account is correct. After all, according to the deflationary account, cognitive contents are not *really* genuine components of cognitive science! Yet *mathematical contents* are. And it's not clear whether they satisfy **(1)** to **(5)**. I claim they don't.

3 - Mathematical contents don't satisfy the desiderata

Do mathematical contents satisfy *desiderata* **(1)** to **(5)**? To answer, start with **(4)**. Mathematical contents are either natural or non natural. If they are natural, **(4)** is satisfied. If they are not, they fail to meet 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". Their non-naturality *does threaten* the naturalistic credentials of cognitive science. So, are mathematical contents natural?

Egan (2014: 213) *replies with a qualified no*: no "classic" content-naturalizing relation manages to naturalize mathematical contents. 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. Causal/informational theories, for example, just seem unable to account for this case. Further, the deflationary account is proposed as an *alternative* to "classic" theories of representations based on content naturalization. And, if mathematical contents were naturalized by a "classic" relation, the deflationary account would fail to be a real alternative. That would be almost a *pragmatic* contradiction (cf Egan 2014; 2020). But, if mathematical contents *must* satisfy **(4)** and "classic" theories of content do not succeed in naturalizing them, *what does*?

Egan (2014: 117, 119) seems to endorse 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. This is to say: the vehicle types identified by f_R

represent what they represent *because* they can be interpreted as arguments and values of some function F. Whilst such an approach *is* naturalistic in the relevant sense[5], it 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), and physically different states or objects are clustered together into a vehicle type because of their common content. But, the vehicles representing mathematical contents cannot be type-identified this way: for f_I to match vehicle types and mathematical contents one-to-one, the relevant vehicle types must have already been identified. And their identification matters. If there is no constraint on how these vehicles are identified, then pretty much *every* system is interpretable as computing *some* function (cf. Putnam 1988; Searle 1992; Copeland 1996 Scheutz 1999). And since *interpretability* is all that matters for mathematical content to be there, panrepresentationalism follows. Thus, **(5)** is unsatisfied. And since panrepresentationalism is connected to empirical adequacy, **(3)** seems unsatisfied too.[6]

Secondly, even supposing that there is a somewhat restrictive way to identify vehicle types, a problem with content determinacy looms large. 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. Take an imaginary device S and a f_R. Assume 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 F(7;9)=16 *or* a function F* F*(Jackson;Harrison)=Lincoln.[7] So, given f_R, S is interpretable under *at least* two functions. But what do the vehicles tokened inside S *really* represent? Numbers, presidents, or both? Do they represent the arguments and values of any function isomorph to addition in the 1-9 range? If interpretability is all that matters, then presumably they represent *all these mathematical contents*. But then their content is not well-determinate in any ordinary sense of the term: well-determined contents represent *one and only one clear target*. So, content determinacy fails. And since content determinacy is constitutively connected with misrepresentation, it fails to. *Desiderata* **(1)** and **(2)** are thus not satisfied.

[5] 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.

[6] 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). Surely saying that these systems represent *counts* as a commitment to panrepresentationalism.

[7] Of course, they're respectively the 7th, 9th and 16th US presidents.

Worse still: given Egan's minimalistic interpretational semantics, her account may be circular. Surely, a good reason as to why a system S is interpretable as computing F is that it *actually computes* F. So, S is interpretable as computing F *because* S actually computes F. However, Egan's account of computation is semantic. In her view (conditions **(i)** to **(iii)** above) S computes F *because* S represents the arguments and values of F. But, if the minimal interpretational semantic Egan proposes is correct, S represents these arguments and values *because* it is interpretable as computing F. And it is interpretable in that way *because* it really computes F! We're caught in a circle.

Cummins (1989: 90-91) noticed the problem first, and argued it should be solved by prioritizing computation: according to Cummins we should say systems represent because they compute - not the other way around. Here, the move is advantageous for several reasons. A suitably robust account of physical computation (and computational implementation) can restrict the number of systems that compute, thereby avoiding pancomputationalism. Given the link between computation and mathematical contents, it will *also* avoid panrepresentationalism, making the deflationary account empirically adequate (or at least, more empirically adequate than it is). These are at least steps towards the satisfaction of **(5)** and **(3)**. Further, a suitably robust account of physical computation should enable us to say any physical system S computes few (ideally one) function F. It can thus restrict the ways in which any system can be rightfully (or at least non-otiously) interpreted. Thus, if mathematical contents are really grounded in interpretability, a robust account of physical computation may make, or contribute to make, mathematical contents determinate and able to misrepresent, thereby contributing to the satisfaction of *desiderata* **(1)** and **(2)**.

Yet, I fear no account of physical computation will deliver these boons. 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 as to why each approach *cannot, in principle*, provide us with the desired results.

Consider, first, *semantic* approaches. Such approaches can take a huge number of forms (Fodor 1975; O'Brien and Opie 2008; Shagrir 2001; Maley 2021), but they all cluster together because they take representation to be *necessary* for computation. Whilst individual accounts vary, they all agree with the "no computation without representation" motto. And this makes these approaches unsuited to function as a platform to deliver well-determined and well-distributed mathematical contents. For, presumably, representations require *contents*. Thus computation requires content. If the content required is cognitive content, then the deflationary account would simply be false: if computation requires cognitive content, then cognitive content is not a "facultative gloss" on computational (function-theoretic) explanations, but sits at their very heart. But if the content required is mathematical content, then we're caught in a loop: our account of physical computation would presuppose the kind of well-determined and well-distributed mathematical contents it is supposed to deliver.

Consider now "mapping" approaches. In general, mapping approaches 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 $c' \rightarrow c''$ 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 (Godfrey-Smith 2009). Otherwise, one could robustify the account adding further necessary conditions. These vary from account to account, and won't matter here (see Piccinini and Maley 2021 for a survey).

There are two reasons as to why "mapping" approaches will not deliver the desired boons. One has to do with computational indeterminacy, and I will discuss it thoroughly when dealing with mechanistic approaches. The other is that all "mapping" approaches entail a form of limited pancomputationalism. Minimally, they're all forced to concede that every physical system S implements an inputless finite state automaton C computing the identity function (cf Chalmers 1995; 2011).[8] 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)** is not satisfied. Since *desideratum* **(3)** (i.e. empirical adequacy) is connected to **(5)**, it would fail to be satisfied too.

Consider, lastly, *mechanistic* approaches.[9] 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*[10] *to compute* (see Miłkowski 2013; Piccinini 2015). Roughly put, a mechanism (in the relevant sense) responsible for a phenomenon is a set of spatiotemporal components performing certain functions and having certain spatiotemporal relations, 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 "feed" some digit determines the mechanism's computational identity. Importantly, such a rule must be

[8] 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*.

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

[10] 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.

medium-independent: it must be sensitive only to the degrees of freedom of digit types, while ignoring any other feature of their tokens.

Now, whilst the mechanistic approach is a robustified version of the "mapping" account (cf. Piccinini 2015), it avoids pan-computationlism. 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 2018, Fresco *et al*. 2021). And this prevents it from being a platform to assign well-determined mathematical contents able to misrepresent. So, **(1)** and **(2)** remain unsatisfied. Let me explain.

Let S be a computing mechanism, 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**
@	@	@
@	#	#
#	@	#
#	#	#

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 conjunction*. So, do "@"s represent the truth value *true*, *false* or both? Their mathematical content is not well determined.

This problem reaches deeper still. For the mathematical contents of @s and #s may be undetermined *even when the function computed is determined*.[11] Consider a system S* displaying the computational behavior summarized in **Table 2**:

Input	**Output**
@	#

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

#	@

Table 2: the computational behavior of S*

S* takes one digit as input yielding one as output. If the input is a @, it yields a # and *vice versa*. It thus seems natural to interpret S* as computing the *logical negation function*, and there seems to be no other interpretation around. 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)** is not met, and since **(2)** is not met, **(1)** is not met too.[12]

An obvious objection is this: whilst looking at S alone does not determine the mathematical contents of @s and#s, looking at how S is embedded in a larger computational device, and how it cooperates with other computational mechanisms, will determine the contents of @s nad #s.

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 system to compute*. S can compute alone. And if computation is what determines mathematical contents, there are still cases in which mathematical contents are indeterminate. 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***
@	@	@	#
@	#	#	@
#	@	#	@
#	#	#	@

Table 3: the computational behavior 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, useful or intelligent. But functions need to be interesting, useful or intelligent to be functions!

[12] 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).

Perhaps one could object to my analysis that computation is *necessary, but not sufficient*, for mathematical contents. Maybe an extra ingredient is needed. Maybe not *all* computational devices represent the arguments and values of the function they compute. Maybe. All of this is surely *possible*, and, whilst I cannot imagine what this "extra ingredient" may be, I've read too much Dennett (1991a) to take a failure of my imagination to be an indicator of a metaphysical impossibility. So, I challenge defenders of the deflationary account to identify the "extra ingredient". I contend this challenge will not be met - but unless someone tries to meet it, there is not much to discuss.

So, the minimalistic form of interpretational semantics Egan endorses is "on hold". And "classic" content naturalization relations (e.g. causal/informational ones) seem unable to do the required job, as highlighted above. What's left? As far as I can see, all that is left is a *semantic primitivism*; that is, the view that there are natural primitive (non-analyzable) semantic facts concerning mathematical contents (cf. Burge 2010). But semantic primitivism is unappealing. For one thing, absent an account of such primitive semantic facts, the primitivist strategy is more a bluff than an account of content (Piccinini 2015: 35). Secondly, it seems that these primitive semantic facts are *epiphenomenal*: they make no difference to the computational behavior of a system. Recall, bfiely system S, whose computational behavior is summarized below in **table 1 bis**:

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

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*. Or, if you prefer, I could bring S with me while I travel through possibile words, until I find a word W* in which, as a matter of primitive semantic facts, @s represent *false* and #s represent true, and use S as an *Or Gate* there. It seems that the "primitive semantic facts", whilst sufficient to give us well-determinate mathematical contents, *make them irrelevant* to the actual functioning of a device. They lose their explanatory power. And, thusly robbed of their explanatory power, it is not clear why we should look at mathematical contents as something more than 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. It is thus tempting to keep them unnaturalized: just erase desideratum **(4)** out of the list.

Whilst the move may be legitimate, there would still be problems for the deflationary account. Even with non-natural mathematical contents, the deflationary account fails to capture the role representations play in cognitive scientific explanation. Below, my counterexample.

4 - Deflating the explanatory power of deflated representations

On the view the deflationary account proposes, cognitive scientific explanations consist of two ingredients: (I) the function-theoretic characterization of the device unders scrutiny, and (II) the ecological component. To the extent that function-theoretic characterization illuminate the mathematical contents a device represents, it seems the deflationary account commits to the view that mathematical contents (with the ecological component) *explain* cognitive phenomena.

How do mathematical content explain? The answer seems twofold. On the one hand, they explain by 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, S would produce the output *o* where *o*=F(*i*). 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* 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)

One important point, often stressed when it comes to *cognitive* content-based explanations, is that explanations of failures and successes always account for *patterns* of failures and successes (cf Gładziejewski and Miłkowski 2017; Shea 2018). The point 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*. It's possible to elaborate further: were *v** smaller than *v*, the system would not have overshoot: it would have under-shoot. Note that in order for this kind of explanation to work, there must be a *systematic correlation* between mathematical contents and failures: 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.

I claim that, in the case of *adversarial examples induced misclassification* (AEIM), no such correlation can be found. Hence AEIMs constitute a direct counterexample to the explanatory ambitions of the deflationary account.

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 and Skansi 2018). Thus their function-theoretic characterization is well known. Simplifying *a lot*, deep classifiers can be considered as "good old 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. The input vector is spread through successive layers of units or 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.

It is evident that 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).[13] And, of course, when *v** fools C, we assist to 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*![14] 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.

[13] 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.

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

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. Nor knowing the function-theoretic characterization allows AEIMs to be explained. In fact, they currently stand in need of an explanation. Worse still, the explanations currently proposed *massively* involve cognitive contents. Consider the following two proposed explanations.[15]

> *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 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 appealing *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 content. *Prima facie*, something can appear as something else *only if* we represent it as a distal target which it is not. But representing distal targets involve cognitive, rather than mathematical, contents. 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: *contra* Egan (2020a 46-48), these ascriptions of cognitive contents cannot have *only* the heuristic role of orientating the research for a function-theoretic characterization. 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.

[15] 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.

And they must play such a role, for there is seemingly *no correlation b*etween mathematical contents and 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, consider that adversarial examples are *transferable*. If an adversarial 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* looks like a prima 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, cf 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.

One could object 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).

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 (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 (Dennett 1991b). 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. 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 account is not empirically accurate.

5 - Concluding remarks

I've argued that the deflationary account of representation faces several severe problems. But aside from "*deflationary account bad*", what's the lesson to be learned? I want to point out two.

One concerns what currently is the most discussed neurocomputational theory, namely predictive processing (and the free-energy principle). Several attempts to account for representational content in that frameworks have argued that *cognitive contents are grounded into mathematical contents* (cf. Wiese 2017; 2018; Ramstead *et al* 2020). If the arguments presented here are correct, these authors should reconsider their views: mathematical content is so riddled with problems that it cannot ground naturalistically respectable cognitive contents. Or, minimally, if these authors want to stand their ground and defend the idea that cognitive contents are rooted into mathematical ones, they should tell us where mathematical contents "pop out from". Given the link between mathematical contents and computational implementation, this is an important challenge to meet, especially because how predictive processing is physically realized in the brain is, to put it mildly, obscure (cf. Walsh *et al*. 2020; Cao 2020).

The other concerns a recent, and, as far as I can see, disorganized trend in thinking about representations in cognitive science. The trend seems aimed at providing a naturalistic account of representations based on a naturalistic account of computation (cf Cohelo-Mollo 2021; Piantadosi 2021). These accounts should be applauded, at least insofar as they explore a territory which has traditionally been left unexplored by "classic" accounts of representations. Yet, they may suffer from over-idealization: they either do not consider any actual computational system (Cohelo-Mollo 2021) or they restrict their analysis to few, quite specific, ones (Piantadosi 2021). This might prove detrimental to the explanatory power of these accounts: by focusing on few, well-selected examples, they may become too narrow to account for important cognitive phenomena, just like the deflationary account has proven to be too narrow to account for adversarial example induced misclassification.

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.

Cao, R. (2020). New label for old ideas. *Review of Philosophy and Psychology*, *11*(3), 517-546.

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.

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

Cohelo-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.

Dennett (1991a). *Consciousness Explained*: Little brown.

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

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

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.

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 compare of the core visual objects recognition behavior of human, 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.

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.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. Annals of the New York Academy of Sciences, 1464(1), 242-268

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.