Troubles with mathematical contents

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.

1 - Introduction

Philosophers of cognitive science have long aimed at accounting for content in non-semantic, non-intentional terms (e.g. Cummins 1989; Shea 2018). In spite of their best efforts, the objective has not been met yet (see Ryder 2019 for a review).

Egan (2014; 2019; 2020a) claims this sad state of affairs motivates a different agenda. Her *deflationary account* mainly aims at capturing the explanatory role representations play in cognitive science.[1] To this end, she distinguishes two kinds of contents: cognitive and mathematical. Cognitive contents are contents usually understood, which, on Egan's account, are just a *facultative gloss* on cognitive-scientific explanations proper. Conversely, mathematical contents represent the argument and values of the functions cognitive systems compute, and thus are essential to, and sufficient for, cognitive-scientific explanations.

1 More precisely, in *computational* cognitive science. Yet, since non-computational cognitive science is typically also non-representational cognitive science (e.g. Kelso 1995), I drop the "computational" qualifier for ease of exposition.

Here, I examine Egan's proposal, arguing that mathematical contents do not satisfy the adequacy conditions on content Egan accepts, and that they are not *always* sufficient to provide satisfactory cognitive-scientific explanations.

To do so, I first sketch the adequacy condition Egan imposes, as well as her deflationary account (§2). I then argue that mathematical contents do not satisfy the adequacy conditions previously sketched (§3). In §4 I articulate my second claim, introducing adversarial examples to deep classifiers and showing that mathematical contents seem unable to account for them. A quick conclusion follows (§5).

2 - Egan's deflationary account of representations

Here, I first introduce the adequacy conditions on a theory of representation Egan accepts (§2.1), then sketch her account (§2.2) and present Egan's arguments to the effect her account satisfies the adequacy conditions (§2.3).

2.1 - The adequacy conditions

Egan (2019: 248-249; 2020a: 28-29) argues that any theory of representations must satisfy at least[2] the following *desiderata* in conjunction:

> (1) Misrepresentation: A successful account of content allows for misrepresentation to occur
>
> (2) Determinacy: A successful account of content assigns determinate contents to representational vehicles
>
> (3) Empirical adequacy: A successful account of content conforms to the actual practice of cognitive science
>
> (4) Naturalism: A successful account of content specifies, using non-intentional and non-semantic terms, at least sufficient conditions for a state or structure to bear a determinate content

2 Egan (2020) actually lists more. But they won't play any role in my argument, hence the omission.

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

These requirements are both minimal and well-known, hence their discussion will be brief.

Note first (1) and (2) are *constitutively* connected. The ability to misrepresent *identifies* representations, setting them apart from mere states or objects (e.g. Dretske 1986). But misrepresentation requires *determinate* contents: open-endedly disjunctive contents make misrepresentation, if not impossible, at least problematic. (3) seems a *desideratum* in its own right - at least insofar modern theories of content aims at capture how representational content figures in cognitive-scientific explanations (e.g. Shea 2018; Rupert 2018). (4) captures the widespread idea that content is not *fundamental*: it supervenes on more basic facts and features of the world. Hence, it should be explainable in terms of these more fundamental facts and features. Lastly, (5) is weakly connected to (3), at least insofar cognitive scientists do not label *every* behavior-producing structure a representation (e.g. 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).

According to Egan (2019: 249-250; 2020a: 29-31), no naturalistic account of representation thus far proposed satisfies (1) to (5), mainly because they fail to satisfy (1) and (2). She claims that causal-informational theories all fall prey to the disjunction problem (see Artiga and Sebastian 2018; Rosche and Sober 2019), that teleological theories deliver indeterminate contents, due to the indeterminacy of functions (see Fodor 1990) or quinean indeterminacy (see Cao 2020), and that similarity-based approaches fail because similarity has not the logical properties of representations, and similarities fail to deliver sufficiently determinate contents (see Segundo-Ortin and Hutto 2019).

In Egan's view, this state of affairs calls for a different approach to representation and content, which I sketch below.

2.2 - The deflationary account

As the deflationary account aims *just* at accounting for the role of representations in cognitive-scientific explanations (Egan 2020a: 43), these provide a natural starting point.

Egan (2010; 2014; 2017; 2020) argues that cognitive-scientific explanations are *function-theoretic*: they unveil the function F computed by a cognitive device S. Egan's favorite example is Marr's (1982) account of early vision, according to which the retina (S) computes a smoothing function (F) convolving a Laplacian operator with a Gaussian operator.

Function-theoretic explanations are *environment neutral*: they make no reference to environmental states of affairs (Egan 2010; 2014). Notice that this form of environment neutrality reaches *into the agent*, to the other computational systems S is connected to (if any). As a matter of fact, retinas take light intensities as inputs. But, were their inputs sound waves (or signals from another computational system) their function-theoretic characterization would not change: they would still be convolving Laplacian and Gaussian operators. Relatedly, function-theoretic characterizations are *domain general*: retinical computations (partially) constitute vision only when retinas are "wired up" to visual cortices in a certain way. Weren't retinas "wired up" to visual cortices that way, their computations wouldn't partially constitute vision - but they would still be the *same* computations.

Thus, the function-theoretic characterization of S tells *only* what function F it computes. Yet, it is genuinely explanatory: it deepens our understanding of S by specifying what it does in terms of a mathematical function we *already* understand. Plus, by being environment neural and domain general, it allows us to predict how S will behave in a wide range of circumstances, thereby boasting a significant counterfactual depth (Egan 1999; 2010; 2014; 2017; 2020a).

But what does it mean to say S computes F? According to Egan (2010; 2014; 2020a), it means that:

> (i) There is a *realization function* f_R mapping the microphysical states of S onto vehicle-types; &
>
> (ii) There is an *interpretation function* f_I mapping one-to-one the vehicle types identified by f_R onto the values and arguments of F, &
>
> (iii) For all argument - value pairs of F, if S is in a state (as identified by f_R) that f_i maps on a specific argument of F, then S is caused to enter in a state (as identified by f_R) that f_i maps on the corresponding value of F

Egan (2014; 2020) provides this simple example. Suppose S computes the *addition function* F. This means that (i): there is a way f_R to group S's states together in well defined vehicle types; & (ii) there is a 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''')=$n + m$.

The values and arguments of F onto which f_I maps the vehicle types f_R identifies are the *mathematical contents* represented by S. Notice that since function-theoretic explanations are environment neutral, mathematical contents are narrow; *so narrow they are independent from the other computational devices S interacts with*. Notice further mathematical contents need not represent *numbers*. They represent the arguments and values of F *whatever they are*. For example, if F is a function from vector to labels, then they represent vectors and labels. Notice lastly, mathematical contents are essential in function-theoretic explanations (Egan 2014: 122-123), and they determine the truth of such explanations: if S does not represent the mathematical contents needed to compute F, then a function-theoretic explanation suggesting that S computes F is false. Hence, we should be committed to them - they are essential to the truth of our best cognitive-scientific explanations.

As said above, Egan construes function-theoretic explanations as environment-neutral. Yet, computing a given function F *won't* contribute to cognition in all possible environments. Thus, cognitive-scientific explanations need to supplement the function-theoretic characterization of S with an *ecological component* (Egan 2019: 253-255).[3] Were the world different, a device

3 Which *is* a part of cognitive-scientific explanations proper.

convolving Laplacian and Gaussians operators wouldn't contribute to vision, but to some other cognitive capacity (or no capacity at all). Hence, the ecological component clarifies how computing F (partially) constitutes cognitive processing *intentionally understood*; that is, how it contributes to the cognitive capacity under investigation described in a familiar intentional lexicon (e.g. vision described as seeing what is where)

On Egan's (2014; 2019: 254; 2020a) view, the information the ecological component conveyes is often perspicuously "summarized" ascribing *cognitive* contents to the vehicles identified by f_R; that is, by saying they represent distal (environment-related) states of affairs. This is done, for instance, when we say that retinal outputs represent *edges*, rather than the result of the convolution of Laplacian and Gaussian operators. This relates the inner operations of S to the environment, thereby clarifying in a perspicuous manner how computing F contributes to cognition pre-theoretically understood:

> "cognitive content is the 'connective tissue' linking the sub-personal mathematical capacities posited in the theory and the manifest personal-level capacity that is the theory's explanatory target" (Egan 2019: 253)

Yet, in spite of their "connective" role, cognitive contents are *not part of cognitive-scientific explanations* proper: they are only a strictly speaking facultative gloss layered over them to perspicuously summarize their ecological component (Egan 2014; 2020a).

In Egan's view, this means that we shouldn't be committed to cognitive contents: they are *only ascribed* to vehicles to simplify our understanding of S's operations. Cognitive contents are *not* the upshot of some privileged vehicle-target naturalistic relation, and are *not* really represented within S. This also entails that there's not fact of the matter concerning which cognitive contents S *really* represents - we are free to ascribe them as we see fit, based on our explanatory and pragmatic concerns; for instance, to minimize the effort required to understand how a computational system works (Egan 2014; 2019; 2020a; Mollo 2020). They might also play a role in orienting the early stages of cognitive-scientific research, so as to discover the relevant function-theoretic characterization of a device (Egan 2020a: 45-48).

Thus, we shouldn't endorse any strong commitment regarding cognitive contents and their existence: they are just a nice linguistic ploy to track the inner goings-on of a computational system in a perspicuous and easy to understand manner.

2.3 - Satisfying the adequacy conditions

Egan (2020a; 2014) claims her account is ideally poised to satisfy (1) to (5). The pragmatic and heuristic nature of cognitive contents allows us to ascribe determinate cognitive contents to each vehicle. And once such content is fixed, misrepresentation can easily occur: if we ascribe to a state *s'* (identified by f_R) a cognitive content *x*, and *s'* is tokened when *x* is not the case, then that token of *s'* misrepresents. Thus, (1) and (2) are easily satisfied. (3) is satisfied too: the account is *prima facie* empirically accurate; and indeed Egan (2010; 2014; 2020a) amply refers to the empirical practice of cognitive science. (4) seems satisfied too. Sure, cognitive contents are non-natural (in the relevant sense), as they are partially determined by our epistemic/pragmatic concerns. But they are *not* part of cognitive science proper, hence they pose no threat to the naturalistic credentials of the latter (Egan 2020a:35). Lastly, as Egan writes:

> "Pan-representationalism is not a worry for the deflationary account, because it does not purport to offer a metaphysical theory of representation. It does not specify a general representation relation that holds independently of explanatory practice in cognitive neuroscience. " (Egan 2020a: 43)

Hence, (5) seems satisfied too.

I find this idyllic picture unpersuasive.

3 - Whence mathematical contents?

Egan's deflationary account takes mathematical contents to be *essential* to cognitive-scientific explanations: cognitive science *really is* committed to them. But, to the best of my knowledge, the deflationary account does not say *how* they are determined, leaving us wondering why the vehicles identified by f_R bear the mathematical contents they bear, and whether they satisfy (1) to (5).

So, what determines mathematical contents? Egan (2014: 213) seems to suggest that *no* naturalistic relation determines them. But then it seems that mathematical contents violate (4), and Egan's account wouldn't be adequate *given her own standards of adequacy*. Notice that, unlike cognitive contents, mathematical contents are taken to be essential to cognitive-scientific explanations. Hence their violation of (4) *does* threaten the naturalistic credentials of cognitive science.

Surely mathematical contents cannot be naturalized by means of causal/informational semantics: the arguments and values of F are *prima facie* mathematical objects, which cannot enter in any causal/informational relation with the vehicles identified by f_R. Perhaps it could be argued that these vehicles are *structurally similar* to the values and arguments of F or that they have the *teleological function* of representing them. The move is technically viable, but it forces a dilemma on the deflationary account. Recall that the deflationary account is motivated by the fact that "standard" theories of content fail to meet (1) and (2). If the deflationary account is correct on this point, resorting to such theories to determine mathematical contents will yield mathematical contents that violate (1) and (2). But if the deflationary account is wrong on that matter, then there is at least one perfectly naturalistic theory of content yielding determinate contents capable of misrepresenting, and it is not clear why the deflationary account should be preferred over that theory.

As far as I can see, the only naturalistic way to determine mathematical contents left is to endorse a realistic (non observer-dependent) form of interpretational semantics (e.g. Cummins 1989; cfr. Egan 2014 117, 119). Roughly, the idea would be that of claiming that the vehicles in S (identified by f_R) carry the mathematical contents they carry because *the causal transitions in S are interpretable under F*; that is, the states of S are such that they can be interpreted[4] as standing for the arguments and values of F. Notice how this seems to be only a baroque unpacking of point (ii) above: to say that S satisfies F in that sense *just is* to say that there exists an interpretation function

4 Notice that what matters here is *only* that the states of S *would* support the ascription, *if made*. So there needs to be no *actual ascriber* - indeed, were an actual ascriber necessary, (4) would fail to obtain.

f_I mapping the relevant states of S onto the arguments and values of F, in a way such that the computational state transitions in S and the argument-value pairings of F march "in step".[5] Although this move could perhaps allow mathematical contents to satisfy (4), it exposes the deflationary account to numerous objections.

To start, every physical system satisfies at least a function F - namely the function that describes the system's internal dynamics (cfr. Scheutz 1999). Were satisfying a function F *sufficient* to bear mathematical contents, every physical system would bear some mathematical contents. But this is a form of pan-representationalism, which violates (5).

A similar result is obtained when noticing that, given an appropriately *ad hoc* and gendermarried f_R, each physical system can be cast as an inputless finite state automaton (Putnam 1988; Copeland 1996). This means that every physical system is interpretable under at least the transition function of that finite state automaton, thereby representing its arguments and values.

Worse still, given a specific f_R, certain systems can satisfy *more than one function* F. To see why, consider a system S under a specific f_R. Suppose that f_R identifies a set of vehicle types such that their relations satisfy (according to a specific interpretation function f_I) a function F: a finite version of addition in the range 1-9; i.e. a finite version of addition that takes a pair of numbers in the range 1-9 as arguments and yields a single number (in the range 2-18) as value. Now, if S satisfies F under f_I, it equally satisfies a function F* under f_I*, where F* is a function taking as arguments two members of the set of the first nine USA presidents and returning as value a member of the set of USA presidents from Adams (2nd president) to Grant (18th president).[6] If this is correct, and satisfying a function is *sufficient* for a vehicle to have a content, then the vehicles of S identified by

5 Yet, there's an important difference between Egan's account and interpretational semantics. On the deflationary account, the fact that the vehicles in S (as identified by f_R) represent (via f_I) the values and arguments of F *constitutes* the computational status of S: S computes F *because* it represents the appropriate values and arguments. Conversely, according to Cummins (1989: 91-91), S represents the values and arguments of F *because* S computes F: S's computational status is thus constitutive of its representational status. Thus, simply adding Cummins's interpretational semantics to the deflationary account *might* make it viciously circular (cfr. Piccinini 2004).

6 Just to give and example, if F* takes as arguments *Jackson* (7th president) and *Harrison* (9th president), it will yield *Lincoln* as value (16th president).

f_R seem to have multiple mathematical contents. Hence it seems that their mathematical content is not well-determined, marring (1) and (2).[7]

How could the deflationary account (conjoined with an interpretational semantics) be defended from these attacks? For starters, notice that, at least in this case, the problems concerning (5) cannot be dismissed by claiming that the deflationary account does not purport to single out a privileged metaphysical relation (or, for that matter, property) endowing vehicles with mathematical contents (cfr. Egan 2020a: 43). This is because, on the one hand, interpretational semantics *does* purport to identify such a relation or property (in terms of interpretability); and, ond the other had, weren't interpretational semantics in the business of identifying such a relation or property, mathematical contents would not be naturalized, and so it wouldn't be able to solve the problem with (4).

Restricting the scope of the deflationary account to the structures typically studied by cognitive science won't help with (5) either. First, it is presently not clear what these structures include: whilst certainly not mainstream, bacterial (e.g. Lyon 2015) and plant (e.g. Calvo *et al*. 2020) cognition have their advocates. Secondly, some structures studied by mainstream cognitive science which are typically taken to be non-representational are assigned some contents by the present account.For example, early subsumption architectures largely utilized finite state machines (Brooks 1999); hence, according to the present account, subsumption architectures turn out to represent the arguments and values of the transition functions fueling them. But subsumption architectures are paradigmatic examples of non-representational architectures. Hence the account does not seem to be empirically accurate as required by (3).

Cummins (1989: 101-102) claimed that the multiplicity and indeterminacy of contents poses no challenge as long as the *explanatory relevant* contents are provided. But this claim will not rescue the account. Indeed, claiming so *amounts to* earsing (2), and thus (1), from the list of adequacy

7 Recently, Papayannopoulos *et al*. (*forthcoming*) have described the two problems mentioned above as two different *kinds* of computational indeterminacy: *functional and interpretative*. Functional indeterminacy concerns how to carve the physical states of a device so that they correspond to the relevant computational states (i.e. which f_Rs are admissible for each device S). Interpretative indeterminacy concerns how to pair the computational states to their meanings (i.e. which f_Is are admissible for each device S).

conditions (see §2.1). But once these points have been erased, it is not clear why one should prefer the deflationary account over "standard" naturalistic accounts of representations - after all, the fact that such accounts fail to satisfy (1) and (2) is what *motivates* the deflationary account. Moreover, if representations *really* are identified by their ability to misrepresent it seems that we simply *cannot* erase point (1). But then, if misrepresentation really requires determinate contents, we cannot erase point (2) either.

It seems to me that what the deflationary account needs to face the problem sketched above is a way to restrict (i) the number of systems that compute and (ii) the number of functions F computed by each system. The account needs to find a way to restrict the number of systems *rightfully interpretable* under a F[8], as well as the number of Fs under which a system can be *rightfully interpreted*. Only if *few* systems compute *few* well-selected Fs (ideally one) the account can satisfy (1), (2) and (5).

The only way I see to do this is by invoking a strong (naturalistic and independently motivated) account of computational implementation. Such an account could restrict the number of computing systems to a selected few, thereby solving (or at least allaying) the problem with (5). Moreover, such an account would presumably impose some constraints on the number of functions computed, thereby solving (or at least allaying) the problems with (1) and (2). Notice that such a move imposes a further modification of the deflationary account (other than it being merged with interpretational semantics): in order to use an account of computational implementation to place constraints on mathematical contents, one must abandon the idea that computation is explained in terms of mathematical contents (as in Egan 2014; 2020a). But this seems a fairly small price to pay, and in previous incarnations of the deflationary account (Egan 1992; 1995) computation *was not* analyzed

8 Notice that this is exactly what Cummins (1989: 91-92) did by claiming that S computes F only if S runs *a program* to compute F. Although this restriction *could* be merged in the deflationary account, it would create a problem for (3): many systems playing a key role in the current empirical practice of cognitive science, such as artificial neural networks, do not appear to execute programs. Hence I will not discuss it further.

in terms of mathematical contents - which suggests that the deflationary has the conceptual resources to pay the price without risking bankruptcy.

Now, the natural question concerns *which* account of computational implementation can perform the desired task. Since I cannot consider every account of implementation here, I follow (Piccinini 2015) and distinguish three coarse-grained approaches to computational implementation, namely (a) "mapping" approaches, (b) semantic approaches and (c) mechanistic approaches. I now examine them in turn.

(a) "Mapping" approaches. The general idea behind "mapping" approaches to computational implementation is this: a system S implements a computational device C only if their state transitions "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 $I(s'')=c''$.

Notice this is just a *necessary* condition. If one takes it also to be *sufficient*, one lands on the "simple mapping account" of implementation (see Godfrey-Smith 2009). Otherwise, one could impose some further limitation on the implementation relation, for instance requiring that, when S transitions from *s'* to *s''*, *s' causes* S to enter in *s*'', or that such transitions must be counterfactual-supporting (see Piccinini and Maley 2021).

There are reasons to doubt that "mapping approaches" can provide the required constraints. It is well accepted that "mapping approaches" entail a form of limited pan-computationalism: that is, they all entail that each physical system implements at least a finite state automaton computing the identity function (Chalmers 1995; 2011).

Whilst not necessarily fatal for "mapping approaches" (crf. Sprevak 2019; Schweitzer 2019), limited pan-computationalism *is* necessarily fatal for the deflationary account, at least given the adequacy conditions in (§2.1). For, if every physical system implements a computational device, then, for every physical system S there is a realization function f_R grouping S's physical states in a

way such that they will "march in step" with the state transitions of a computational device C. And, if this is the case, it is then trivial to build an interpretation function f_I such that the vehicle types identified by f_R represent the argument and values of the function C computes (minimally, a function states of C to states of C). Hence, every physical system represents some mathematical content (i.e. the relevant states of C), and pan-representationalism is not avoided.

(b) Semantic approaches. Semantic approaches to computational implementation cluster together because they all endorse the idea that physical computation *essentially* consists in the manipulation of representations. The relevant kinds of representations and how they should be manipulated varies from account to account (cfr. Fodor 1975 with O'Brien and Opie 2008); but, in all cases, the idea that only *representational* systems can be *computational* systems seems to enable semantic accounts to allay, or even solve, the problem with (5). Surely not every physical system is a representational system.

Now, semantic approaches to computational implementation *presuppose* a solid notion of representation and, thus, of content. The deflationary account offers two distinct notions of content: mathematical and cognitive. None of the two allows the deflationary account of representations to merge with a semantic approach to computation successfully.

If a semantic account of computational implementation requires the notion of mathematical content, then clearly such an account cannot be mobilitated to constrain the way an interpretational semantics assigns mathematical contents - that would be circular. But a semantic account leveraging cognitive contents to spell out the notion of computational implementation is clearly incompatible with the deflationary account, for it would plunge cognitive contents at the heart of computational cognitive science. And that is antithetical to the stance on cognitive contents the deflationary account recommends.

(c) Mechanistic approaches. These approaches cluster together because they apply insight from (neo-)mechanist philosophy of science to unravel the nature of computational implementation. On

these views, a physical system implements a computational device only if it is a *mechanism with the function*[9] *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. 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 *medium-independent*: it must be sensitive only to the degrees of freedom of digit types, while ignoring any other feature of their tokens.

Adopting a mechanistic approach to implementation allows the deflationary account to satisfy (5). According to mechanistic approaches, few physical systems compute. Not all physical systems are mechanisms (e.g. the system composed by me and my left shoe isn't), not all mechanisms operate according to rules (e.g. a random number generator doesn't, see Piccinini 2010: 293), and not all mechanisms operating according to rules have the *function* to compute[10] or manipulate their inputs and outputs according to medium independent rules (e.g. my stomach systematically pairs the "input" it receives to the "output" it produces, but the physical stuff these inputs and outputs are made of matters a lot to the stomach rule-following behavior). Thus, by limiting the set of systems that can *rightfully* be seen as satisfying Fs to the ones a mechanistic approach recognizes as genuinely implementing a computation, the deflationary account can avoid pan-representationalism, thereby satisfying (5).[11] This is progress.

9 I will leave the relevant notion of function unspecified for two reasons. First, it is not relevant for my argument. Secondly, it changes across accounts of mechanistic implementation (cfr. Miłkowski 2013; Piccinini 2015), and I do not *need* to take a stance on which is the best mechanistic account.

10 The set of systems having such function varies as the relevant definition of function varies. I will stay neutral on the issue here.

11 Although it should be noted Egan (2017) has important reservations concerning mechanistic approaches, and so perhaps the overall picture might not be as rosy as I just suggested.

But it is not progress enough, for at least some such mechanisms can be interpreted under two Fs; hence they can be assigned multiple mathematical contents (cfr Piccinini 2015: 36-39; 127-130; Dewhurst 2018, Fresco *et al*. 2021). An example taken from (Sprevak 2010) illustrates the point.

Suppose S is a computing mechanism able to manipulate two input digits to produce one output digit. Both input and output digits belong to one of two types: "@"s and "#"s. S manipulates them as follows: when both inputs are "@"s, it outputs a "@", otherwise, it outputs "#". This is a clear rule a mechanism can follow, which systematically pairs the inputs and outputs of the mechanism. Yet, this input/output pairing is interpretable under two different argument/value pairings, hence under two different Fs. So, it is compatible with two different assignments of mathematical contents. According to an interpretation f_1, "@"s represent the truth-value *true* and "#"s represent the truth value *false*. Conversely, according to f_1^{-1} "#"s represent the truth-value *true* and "@"s represent the truth value *false*. Thus, according to f_1, S computes the conjunction function, whereas according to f_1^{-1} it computes the inclusive disjunction function. According to both f_1 and f_1^{-1} the vehicles processed by S have mathematical contents: in both cases, they represent the truth-values that are arguments and values of two simple logical functions. So, it seems that the contents of "@"s and "#"s are not well determined: they are interpretable under two different mathematical functions, and thus they represent *both* "true" and "false". But then (2) fails to obtain. Given that (2) and (1) are constitutively connected, (1) fails to obtain too.

Notice that I'm not claiming that the computational identity of the mechanism is indeterminate (cfr. Sprevak 2010; Fresco *et al.* forthcoming). I'm persuaded that the computational identity of the device *is* determinate, and in fact it can be easily expressed as a pairing of input and output digits as in **table 1**[12]:

12 Using the useful distinction of Papayannopoulos *et al*. (*forthcoming*) we can say that here S is interpretatively, *but not functionally*, indeterminate: in this example, S *does* compute a well determinate function, *if by function we mean only a determinate input-output digit pairing*. But that input-output digit pairing is interpretable under too many Fs; hence the mathematical contents are still indeterminate. Notice also that taking the relevant F to be the well determined function pairing input and output digits *will not* yield determinate mathematical contents. Indeed, it will yield no contents at all. For that function is defined over digits. And digits are not *represented by* the mechanism during computation; rather, they are *instantiated in* the mechanism during computation. Another way to see the point is this: according to the mechanistic account, these digits are causally efficacious. But, according to the deflationary account,

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

Table 1: The input-output table of S

What I'm claiming is that a mechanism pairing inputs and outputs in this way can be interpreted under *two* logical functions: conjunction and disjunction. Notice that the same holds true even for a computational device whose computational identity is not considered uncertain in the same way (Fresco *et al*. 2021: 6). Consider a device exhibiting the input-output behavior described in **table 2**

Input	**Output**
@	#
#	@

Table 2: a device with a well defined computational identity can still have indeterminate mathematical contents

The computational identity of this device is not indeterminate: the device computes the logical negation function (Sprevak 2010; Fresco *et al*. 2021). But both "@" and "#" can be interpreted as *both* truth values.[13]

At this point, it seems natural to think that we can "discover" the *correct* interpretation by looking around the computational device, to see *how* it cooperates with other such devices. By so doing, it could be established whether it computes the conjunction or disjunction function, and which truth value "@"s and "#"s respectively represent (cfr. Dewhurst 2018; Fresco and Miłkowski 2021; Fresco *et al*. 2021). For example, were the device found to cooperate with a second device S' computing as in **table 3**:

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

Table 3: the input-output behavior of S'

content is not causally efficacious (Egan 2014; 2020). Ergo, these digits are not contents.

13 Again, using the distinction of Papayannopoulos *et al*. (*forthcoming*), the device is *interpretationally*, but not *functionally*, indeterminate.

Knowing that S operates together with S', it would be natural to interpret "@"s as "true" and "#"s as "false"; after all, S' really seems to be computing the "if… then" function.

Yet, this way of proceeding does not solve the problem, and runs against the spirit of the deflationary account. It runs against the spirit of the deflationary account because it makes mathematical contents *too wide* to be the denizens of function theoretic explanations: they would partially depend on the internal environment of a computational device. And it does not solve the problem because although S' is surely naturally interpret*ed* as computing the "if...then" function, it is also interpret*able* under the opposite assignment of truth value. Surely, under such assignment of truth values S' does not compute an *interesting or useful* function, and this is why we would naturally interpret it as computing "if...then". But unless mathematical contents are determined by our observations and interests (which would make them non natural), S' remains interpretable under the opposite assignment of truth values. Thus, even if the mechanistic account of computation can allay the problems with (5), it cannot ally the problems with (1) and (2).

It thus seems right to conclude no account of computational implementation can constrain mathematical contents enough to allow them to meet (1) to (5) in conjunction.

Perhaps, then, the deflationary account is best served leaving interpretational semantics behind. Yet, it is not clear what could substitute for it. The only open option I see remaining to naturalize mathematical contents is to endorse a form of semantic primitivism, claiming that there are basic semantic facts concerning mathematical contents, and that these facts are unanalyzable and yet perfectly naturalistic (cfr. Burge 2010).

But this option is *extremely* unappealing. For one thing, unless some satisfactory account of these primitive semantic facts is provided, the move sounds like a bluff allowing one to use mathematical contents without providing an account for them (cfr. Piccinini 2015: 35). Secondly, and relatedly, it is not clear why one shouldn't *also* endorse a form of primitivism about cognitive

contents. I can see no principled reason as to why one should hold that there are primitive semantic facts about mathematical contents but no primitive semantic facts about cognitive contents.

In conclusion, it seems correct to say that, as things stand, mathematical contents do not appear naturalizable. Hence they violate (4), and Egan's account violates her own standard of adequacy.

4 - Are mathematical contents explanatory powerful?

According to the deflationary account, the mathematical contents posited by the function-theoretic characterization of a device, together with the ecological component of the theory, are sufficient to fully explain cognitive phenomena.

I'm unpersuaded.

To start, how should mathematical contents explain? Contents can be said to explain in various ways. Some (e.g. Dretske 1988; O'Brien 2015) think that contents explain by acting as causes of a specific kind. But the deflationary account denies this, suggesting that only vehicle tokens have causal powers (Egan 2014; 2020a).

Others suggest that contents are explanatory powerful because they allow us to find generalizations that we wouldn't be able to find considering only the physical features of vehicles and vehicle types (Dennett 1991; Cao 2012). To explain why I react in the same way to a physical letter and an e-mail, the relevant thing to do is to appeal to the content of these mails, rather than seeking for some physical property shared by their vehicles. Yet, according to the deflationary account, f_1 establishes a *one-to-one* correspondence between vehicle types and contents (see point (ii) above). Hence, there's no generalization based on content which is not captured in a generalization based on vehicle types, as Egan (2020b) concedes.

Others (e.g. Gładziejewski and Miłkowski 2017; Shea 2018) suggest contents explain by accounting for the *success or failures of systems*. The deflationary account seems to endorse this view too. For instance, Egan writes:

> "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)

Notice that, on this view, contents do not *just* explain mistakes, but also *patterns* of mistakes (cfr Shea 2018). To elaborate on Egan's example: suppose the mathematical content of the difference vector orcentrating motor control should be x. Suppose that, instead of x, the system represents a different value x', such that $x < x'$. The fact that x' is larger than x explains why the hand overshoot. And the magnitude of the mismatch between x and x' explains the magnitude of the overshooting - had the system tokened a vehicle representing a value $x'' > x'$, the overshooting would have been larger.

Notice that for contents to explain *patterns* of successes and failures, there needs to be some intelligible correlation between the mathematical contents and the successes and failures. In the motor control toy-example given above, the correlation is simple and linear: the higher the value of x', the larger the overshoot. In more realistic cases, the correlation can be way more complex (e.g. highly non-linear). And yet, if mathematical contents are to explain patterns of failure, it seems that there must be *some* correlation between the mathematical contents represented and the failures of the system in which they are represented.

Now, there seem to be some cases in which mathematical contents and the ecological component alone seem to be insufficient to yield this kind of explanation. Thus, consider *adversarial examples to deep classifiers*.

Deep classifiers are a class of artificial neural network performing classification. Roughly put, their task is that of outputting a probability distribution over labels, given an input (Skansi 2018).[14] Just as the "classic" artificial neural networks of the 80's, deep classifiers consist of a finite set of

14 Notice this is a coarse-grained function-theoretic characterization of these networks.

hierarchically arranged processing units ("neurons") systematically connected by means of weighted connections ("axons"). Each "neuron" is a computational device in its own right, computing an *activation function*, pairing the input received by each neuron with the output each neuron produces. Importantly, the activation function is an *hyperparameter* of the machine learning model: a variable in the model whose value cannot be extracted from the data, and must thus be handcrafted by the network's creator.[15] "Axons", in contrast, are less computationally active: they just *weigh* the signals traversing them (amplifying or dampening it), so as to provide each "neuron" the right sort of input. These weights are the *parameters* of the machine learning model: variables whose value can be extracted from the data, and that "tell" the model where the boundaries between labels lie in the network's activation space. Hence notice that deep classifiers (and artificial neural networks more generally) richly trade in mathematical contents: each neuron *computes* an activation function, and the weights are the parameters of the *function* the network computes.

What sets apart deep classifiers from "classic" artificial neural networks is that deep classifiers are *deep*, meaning that they have more than a single hidden layer of processing units. Moreover, they are not *internally homogeneous*: they include different kinds of "neurons" computing different activation functions. Lastly, they are not *uniformly connected*: each "neuron" does *not* receive input from all the other "neurons" in the previous layer, but only from a selected few of them. Without entering in too much mathematical detail (which are not needed for the argument)[16], these features make deep classifiers significantly more computationally powerful than their "shallow" counterparts.

Adversarial examples to deep classifiers (see Yuan *et al*. 2019 for a survey) are either slightly altered input patterns which a well trained classifier *misclassifies* with high confidence (e.g. Su *et al*. 2019) or images unrecognizable to humans that well trained networks classify with high confidence (e.g. Nguyen *et al*. 2015). Adversarial examples are puzzling because they reveal an

15 Other hyperparameters include the learning rate, the topology of the network and the number of processing units.

16 For non-mathematical introductions to deep classifiers, see (Buckner 2019; Mitchell 2019). See also (Skansi 2018) for an easily accessible mathematical introduction to the subject.

unexpected weakness of the (currently) best machine learning model. They are also puzzling because, albeit some deep neural networks accurately model *some* aspects of human discrimination (see Yamins and DiCarlo 2016; Rajalingham *et al*. 2018), humans do not seem to be susceptible to adversarial examples - indeed, sometimes adversarial examples are often defined as data that "fool" state-of-the-art deep classifiers *but not humans*.[17] It is thus clear that adversarial examples call for an explanation: why does a machine capable of human-level performance commit such mistakes? Why don't *we* commit them? Answering these questions is important both to build better machines and to better understand the human cognitive system. Yet, it is very hard to see *how* these questions could be answered using mathematical contents (and ecological facts) alone.

First, consider the following question: why are deep classifiers fooled by adversarial examples? If mathematical contents (and ecological component) do indeed explain success and failures, it should be relatively easy to answer this question - after all, we do know the relevant mathematical contents deep classifiers trade in: we know which loss function they minimize, what activation function each neuron computes, their learning rate, and there are various network analysis techniques that can be used to further finesse this knowledge of a trained network. The fact that, knowing all this, we still are unsure as to why deep classifiers are susceptible to adversarial examples suggests that mathematical contents are not so explanatory powerful after all.

Moreover, it is hard to see how mathematical contents could explain the *patterns* of failures when it comes to classifying adversarial examples. This is because adversarial examples are *transferable*: if an example *E* "fools" a classifier *C*, then *E* is also likely to "fool" *in the same way* a second classifier *C'*, even if *C* and *C'* have different parameters and hyperparameters (e.g. they have a different network topology, use different activation functions and have been trained on different datasets, see Szegedy *et al* 2013: 5).

17 There is some experimental evidence suggesting that time-pressured humans are susceptible to adversarial examples too (Elsayed *et al*. 2018), and so perhaps a definition of adversarial examples in terms of "they fool machines, but not humans" is not entirely accurate. At any rate, it seems safe to say that, even if adversarial examples can fool humans, they do not fool us *as catastrophically* as they fool deep classifiers.

The fact that adversarial examples are transferable seems to prevent the obtaining of any correlation between mathematical contents and success/failure in classification. If two different clarifiers C and C' are fooled in the *same* way by the *same* adversarial examples in spite of the fact that the mathematical contents they represent can be radically different (as a result of the fact that e.g. C and C' have different learning rate, deploy different activation functions, have different architectures, different connectivity or a different number of weights/parameters), then the chances of finding any discernible correlation between mathematical contents and successes/failures appears to be vanishingly slim.

Moreover, even if C and C' were two identical architectures, some of their mathematical contents would still be different, for their weights/parameters would still be randomly initialized prior to training; and even if training forces the weights to converge, it does not force them to be *identical* (Churchland 1992: 177-178). Yet, C and C' would still be "fooled" in the *same* way by the *same* adversarial example, in spite of this difference in mathematical content. If these reflections are on the right track, mathematical contents can vary *without* thereby varying the way in which deep classifiers react to adversarial examples. But this means that variations of mathematical contents *are not* reflected in variations of successes and failures, and so the two appear to be uncorrelated. To account for *patterns* of successes and failures, mathematical contents and successes and failures must be correlated. Hence, it appears that mathematical contents cannot *explain* in the relevant sense of the term.

Notice, further, that according to the deflationary account, the tokening of a vehicle bearing an incorrect mathematical content should explain *miscomputation* (Egan 2017). Misclassification induced by adversarial examples surely qualifies as a case of miscomputation: a network "fooled" by an adversarial example returns an incorrect probability distribution over labels, and so it seems correct to say that it has miscomputed such a distribution. But it seems to me far less clear that such a miscomputation is a result of the tokening of a vehicle bearing an *incorrect* mathematical content.

As far as I can see, when a network misclassifies an adversarial examples it is *not* the case that some of its neurons enter in a state that does not agree with (i.e. it is wrong in respect to) their activation function. Nor the blame can be easily shifted to the values of weights or the hyperparameters of the network, as the notion of a weight or hyperparameters carrying an incorrect mathematical content seems a bit mysterious. As pointed out before, different networks will have weights carrying slightly different numerical values, and (assuming parity of performance), I see no good reason to consider one of these value assignments as the correct one. And it seems to me that a similar point holds for the hyperparameters: their value can change across architectures, but, if the various architectures are equally (or comparably) good at the classification task, I see no good reason to consider a set of hyperparameters as the one carrying the "true" mathematical contents.

Now, perhaps the reflections articulated above are a bit unfair. After all, according to the deflationary account, what provides a complete explanation of a system's success and failures are mathematical contents *together with the ecological component of the theory*. And, thus far, I've been silent on the ecological component. So, one could legitimately contend that considering the ecological component would allow us to see *how* mathematical contents explain adversarial-examples induced misclassification.

Whilst entirely legitimate, I find it hard to make sense of the contention above, for it is far from clear what is the *ecology* of an artificial neural network. For the most part, artificial neural networks "live" within our computers, and the only access to the external world they are provided with is given by the input patterns they are administered. But that input pattern simply is the input vector, which is part of the mathematical contents networks need to represent. So, it seems that, in the case at hand, there is no ecological component *over and above* mathematical contents.

A defender of the deflationary account might then perhaps answer by claiming that since there is no ecological component *over and above* mathematical contents, the case of adversarial examples-induced misclassification fails to be a compelling counterexample against the deflationary account.

The idea would be that of claiming that since in the case at hand there is no separate ecological component, it does not fall within the explanatory scope of the deflationary account. This move strikes me as technically legitimate but practically suicidal, as would significantly diminish the explanatory power of the deflationary account. Indeed, making such an argument *just is* admitting that the deflationary account cannot explain how artificial neural networks (and other systems "inhabiting" only computers such as software agents) work. This seems a too high price to pay.

Thus far, I have argued that adversarial examples-induced misclassification cannot be explained by mathematical contents and ecological component (if present) alone. Observing how adversarial examples-induced misclassification is currently explained seems to back up my claim. Here, I will look at two proposed explanations to substantiate my claim. Notice that I'm using these explanations *only* as illustrative examples. I do not wish to imply that they are the *best*, or *correct*, or *only possible* explanations.[18] I wish only to claim that they are possible explanations, and that they do *not* rely on mathematical contents and ecological component (if at all present) *alone*.

The first explanation is provided by Ilyas *et al*. (2019).[19] Their explanation requires a bit of set up. First, they carefully define features mathematically, and then they define a subclass of features (useful features) as the features that reliably correlate with the true label of the input. Having done so, they divide useful features into two subclasses: robust and non-robust. Robust features remain invariant under adversarial perturbation (i.e. still correlate with the right label), whereas non-robust ones "flip" their label under perturbation. On the view Ilyas and colleagues suggest, adversarial misclassification is due to the networks' reliance on such non-robust features to carry out classification tasks.

Although thus far the explanation proposed is mainly mathematical, Ilyas and colleagues (2019:2) are clear in stating that their mathematical definition of features is intended to capture the "folk" definition of features as representations of salient distal properties (cfr Hinton 2014; Olah *et*

18 Indeed, I chose two hardly compatible explanations to remain as neutral as possible on the matter.
19 See also (Engstrom *et al*. 2019; Bucker 2020) for discussion of this proposal.

al. 2018), and they state, in an equally clear manner, that non-robust features are non-robust *only under a human-selected notion of similarity* (Ilyias *et al*. 2019: 10). This, they claim, makes adversarial examples-induced misclassification an human-centric phenomenon: *we* see networks being "fooled" because they rely on different (and non-robust, given our notion of similarity) *properties* to carry out the classification task. They also argue that adversarial examples-induced misclassification could be avoided endowing networks with human priors. Notice how all of this is presented (and naturally interpreted) in terms of cognitive contents: features are representations of distal properties, the human-selected notion of similarity holds among distal objects, and priors are representations of subjective assignments of probabilities to distal events.

As the second example, consider the experimental work described in (Zhou and Firestone 2019). They 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) - an explanation clearly based on *cognitive*, rather than mathematical, contents.

Hence, it seems that, *contra* the deflationary account, currently available explanations of adversarial-induced misclassification do *not* rely exclusively on mathematical contents and the ecological component of cognitive theories.

Perhaps a defender of the deflationary account might contend that these appeals to cognitive content are just "loose talk" deployed to make the real, purely mathematical, easier to grasp. But this surely does not seem to be the case: both publications appear to be aimed at *experts*, which can

easily grasp a purely mathematical explanation. Moreover, Zhou and Firestone's paper contains very little mathematics, and only in the "methods" section. Their explanation is provided in purely informal, and cognitive content-involving, terms

A defender of the deflationary account might further contend that cognitive contents appear in these explanations only because the empirical research on adversarial-example-induced misclassification is still young, and we do not yet possess a complete function-theoretic understanding of these cases (cfr. Egan 2020a: 45-47). Yet, although adversarial examples have been discovered only recently (Szegedy *et al*. 2013), we *do* possess a complete function-theoretic characterization of artificial neural networks. Unlike *natural* neural networks, artificial neural networks are models *we build*. We determine their hyperparameters (such as activation function, network topology, learning rate and the like). And, after training, we can access their parameters and know their mathematical values. So, we do not have to reverse-engineer them, and we do not have to guess which function they compute from behavioral data. Hence, when describing the operations of artificial neural networks, cognitive contents cannot be *only*: "a temporary placeholder for an incompletely developed computational theory" (Egan 2020a: 34).

Lastly, a defender of the deflationary account might contend that these explanations are incorrect, and that the correct explanation will resort only to mathematical content (and the ecological component, if any). But in order for this argument to have some bite, it seems we need some positive reasons to think the explanations sketched above are misguided, and/or at least a sketch of a purely mathematical explanation. Lacking these, the objection seems only the assertion of one's faith in the explanatory prowess of the deflationary account.

5 - Concluding remarks

Here, I have scrutinized Egan's deflationary account of representations. I have argued that the mathematical contents that constitute its explanatory core fail to meet the relevant *desiderata* the

account accepts, and that they do not *always* appear able to provide satisfactory cognitive-scientific explanations.

I wish to conclude, however, by stating that this essay *shouldn't* be read as providing reasons to endorse "standard" naturalistic and reductionist theories of representation. These have monopolized the discussion on cognitive representations from the 80's on (e.g. Cummins 1989; Shea 2018), and the persistent uncertainty about their fortunes (Ryder 2019) *really* motivates the search for alternative approaches. Sadly, the discussion of these alternative approaches has thus far been aimed either at dismissing them in favor of "standard" ones, or at showing that such approaches end up collapsing onto "standard" ones (e.g. Sprevak 2013; Ramsey 2020).[20] This, I think, prevented further exploration of the theoretical space lying between "standard" representationalism and straightforward eliminativism of cognitive representations. Here, I have tried to explore this space, by assessing Egan's deflationary account on its own terms. Although the results of my explorations have been mostly negative, the deflationary account most likely covers only *a fraction* of the conceptual space separating "standard" representationalism from anti-representationalism. It is thus still possible that a thorough exploration of such a space will deliver us a workable, "non-standard" account of representations.

References

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

Brooks, R. (1999). *Cambrian Intelligence*. Cambridge, MA.: The MIT Press.

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.

20 This dismissive attitude has been so prominent that one can easily interpret (Mollo 2020) as making the point that *there really are* alternative approaches worth discussing.

Cao, R. (2012). A teleosemantic approach to information in the brain. *Biology and Philosophy*, *27*(1), 49-71.

Cao, R: (2020). New labels for old ideas: predictive processing and the interpretation of neural signals. *Review of Philosophy and Psychology*, *11*, 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.

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

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

Dennett, D. (1991). 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. (1992). Individualism, computation and perceptual content. *Mind*, *101*, 443-459.

Egan, F. (1995). Computation and Content. *The Philosophical Review*, *104*(2), 181-203.

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. (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.

Egan, F: (2020b). Content is pragmatic: comments on Nicholas Shea's Representation in Cognitive Science. *Mind & Language*, *35*(3), 368-375.

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.

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

Kelso, S. (1995). *Dynamic Patterns*. Cambridge, MA.: The MIT Press.

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

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

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

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

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

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. (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. (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/

Piccinini, G. (2004). Functionalism, computationalism and mental content. *Canadian Journal of Philosophy*, *34*(3), 375-410.

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.

Ramsey, W. (2020). Defending representation realism. In J. Smortchkova, K. Dolega, T. Schlicht (Eds.), *What Are Mental Representations?* (pp. 55-78), New York: Oxford University Press.

Roche, W., & Sober, E. (2019). Disjunction and distality: the hard problem for purely probabilistic causal theories of mental content. *Synthese*, https://doi.org/10.1007/s11229-019-02516-y

Rupert, R. (2018). Representation and mental representations. *Philosophical Explorations*, *21*(2), 204-225.

Ryder, D. (2019). Problems of representation II: naturalizing content. In S. Robyns, J. Simons, P. Calvo (Eds.), *The Routledge Companion to Philosophy of Psychology* - 2nd edition (pp. 251-279). New York: Routledge.

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.

Segundo-Ortin, M., & Hutto, D. (2019). Similarity-based cognition: radical enactivism meets cognitive neuroscience. *Synthese*, *198*, 5-23.

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. (2013). Fictionalism about neural representations. *The Monist*, *96*(4), 539-560.

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.

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

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.