Philosophy of Science (2025), **00**, 1–21
doi:10.1017/psa.2025.10157

ARTICLE

Decoding Cognitive Neuroscience: A Defence of the Explanatory Role of Content

Johan Heemskerk

University of Warwick, United Kingdom
Email: j.heemskerk@warwick.ac.uk

(Received 10 April 2025; revised 07 August 2025; accepted 09 August 2025)

Abstract

Cognitive neuroscientists typically posit representations that relate to various aspects of the world, which philosophers call representational content. Anti-realists about representational content argue that contents play no role in neuroscientific explanations of cognitive capacities. In this paper, I defend realism against an anti-realist argument due to Frances Egan, who argues that for content to be explanatory it must be both essential and naturalistic. I introduce a case study from cognitive neuroscience in which content is both essential and naturalistic, meeting Egan's challenge. I then spell out some general principles for identifying studies in which content plays an explanatory role.

1. Introduction

Cognitive neuroscientists typically posit 'representations' which relate to various aspects of the world. Some cell may be said to represent edges (e.g. Marr 2010), or some high-dimensional neural activation space may be said to represent a complex object (e.g. DiCarlo and Cox 2007). Philosophers term the thing in the world which a representation represents the 'content' of that representation.

However, anti-realists argue that representational content does not play a role in neuroscientific explanation (e.g. Hutto and Myin 2014; Raja 2018). In this paper I focus on a particularly compelling argument due to Frances Egan. Egan argues that the mathematical function computed by a given 'state or structure' (read: representational vehicle) is sufficient to explain the cognitive capacity under investigation. Content attributions, Egan argues, merely serve various heuristic purposes, such as making the scientific theory perspicaciously related to the presumed pre-theoretic interests of readers.

I argue, contra Egan, that representational content is a genuine explanatory posit in some regions of cognitive neuroscience. I specifically address Egan because she provides helpful criteria for identifying explanatory relevance. I will argue that we

can accept Egan's (stringent) criteria, and show that – at least in certain studies – content meets them.

Egan offers two considerations in favour of her position: (i) content is not "essential" and (ii) content is not "naturalistic". However, she argues that being both essential and naturalistic are necessary for content's inclusion in the 'theory proper' of cognitive neuroscience (i.e. for content to be explanatorily relevant). Egan argues that content is not essential in the sense that it is not used to individuate representations in cognitive neuroscience. Rather, she maintains, representations are individuated with respect to the mathematical function computed by the system under investigation. Content is not naturalistic in the sense that there is no determinate scientific principle guiding content attribution – only the heuristic requirements of scientists.

I endorse Egan's criteria that content, to feature in the explanations of science, must be both essential and naturalistic. However, I introduce a case study in which content *is* both essential and naturalistic. I show that, in a landmark study on face recognition by Chang and Tsao (2017), representational states are individuated with respect to their contents in precisely the way required by Egan (section 3.1). I then show that content is determined by the naturalistic relations of encoding and decoding, which constrain content attributions not with respect to the heuristic requirements of researchers, but by the discriminatory and decoding capacities of the system itself (section 3.2).

I end by considering three principles that will hopefully aid theorists in isolating those regions of cognitive science in which content is used as part of the theory, not just as a heuristic gloss (section 4).

2. The problem

According to Egan, attributions of content in cognitive neuroscience are a "gloss", and not part of the "theory proper". She argues that content is attributed to a mental representation[1] on the basis of pragmatic choices by the scientist; she chooses, from the many environmental items that have some degree of co-variance with the representational state, the item[2] that best serves her communicative aims.

For example, she may attribute the content ⟨*face*⟩ to a neural representation within the fusiform gyrus. She may do this because we are primarily interested in the role that the neural representation appears to play in face recognition, and choosing the content ⟨*face*⟩ transparently situates the investigation within that context. She could have chosen any number of co-varying contents, including ⟨*face-like-shape*⟩ or ⟨*two-dots-above-a-curved-line*⟩, either of which specifies something that co-varies with the neural representation.[3] The only reason she chose ⟨*face*⟩, so Egan argues, is that it specifies the most intuitive environmental item, helping the reader understand the

[1] I follow Thomson and Piccinini (2018) in treating mental representations as a case of neural representation. Generally, I take neural representations to be high-level generalisations over neural activity. Specifically, neural activity can be modelled as a mediating state that provides information about the external environment to downstream systems.

[2] I follow Shea in using 'item' as a technical term intended to be metaphysically neutral about the things that can be content – objects, processes, properties, and so on can all be items in this sense (Shea 2018, 76n).

[3] An image of two dots above a curved line is a stimuli sometimes used in facial recognition tasks (e.g. Tsao and Livingstone 2008, 10).

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

broad significance of the scientist's research. The content-gloss is a way of marketing one's research, with no theoretical principles – only *pragmatic* principles – governing its attribution.

Egan's argument presents a challenge to realism about representation in cognitive neuroscience. Specifically, a tenet of realism which states that "content [. . .] is not simply derived from our own interpretive activities" (Ramsey 2020, 56). If content were not part of the theory proper, but part of a gloss, it *would be* derived from our own interpretative activities. I will show that, in certain circumstances, content is part of the theory proper. As such, I aim to clear one hurdle in the way of representational realism.

The present defence of realism is conducted with the aim of demonstrating that content can feature in a causal explanation of the operation of the cognitive system. This places the current project alongside much recent literature on the realism/anti-realism debate. For example, Lee (2021), Gładziejewski and Miłkowski (2017), and Piccinini (2022) all argue that a realist account of content can deliver contents that are "causally relevant to the system" (Lee 2021, 12). The present account makes an advance by highlighting the precise role of encoding and decoding (see section 3.1.1). By closely following scientific practice, I align with the methodology of Thomson and Piccinini (2018), who argue that scientists frequently cite neural representations in their explanations, making the anti-realist position untenable. The present paper extends and complements such work by distinguishing between representational talk as a gloss and as a genuine explanatory posit. This ensures that any purported evidence of scientists employing representations can be verified as genuinely explanatory, directly addressing the concerns of anti-realists such as Egan.

Egan's argument also threatens contemporary psychosemantic methodology. Psychosemanticists aim to provide a naturalistic theory of representational content. A theory of representational content provides an account of the relation between a representation and its content – answering the question: in virtue of what does representation *X* have content *Y*? Contemporary theorists pursue a methodology that involves turning to scientific practice to uncover an implicit theory of content. As expressed by Ramsey, the methodology involves a critical examination of "the different ways cognitive scientists appeal to notions of representation in their explanations of cognition" (Ramsey 2007, xv). Such theorists hold, as Burge puts it, that there are elements of a theory of representation that cognitive science, "without being fully aware of its own accomplishment" (Burge 2010, 9), has discovered and that the philosopher seeks to uncover. However, if Egan is right, there *is* no implicit theory of content in cognitive science. Only heuristic principles govern content attribution. Those looking for theoretically informed content attributions are bound to be disappointed. So, there must be a response to Egan if contemporary psychosemantic research is salvageable. I hope to provide an adequate response that allows us to hold out hope for this methodology.

In the next section I reconstruct Egan's argument, and clarify the terms 'content', 'theory proper', and 'gloss'. In the remainder of the paper I respond to Egan's argument. I will assume throughout that Egan is right both about the requirement that content feature in the theory proper if it is to be explanatory, *and* Egan's criteria for inclusion in the theory proper. I focus on a case study from cognitive neuroscience in which, I argue, content attributions meet Egan's criteria for inclusion in the theory proper. I draw three general principles from this case study. These principles should

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

help guide us towards those parts of cognitive science in which content attribution is part of the theory proper. If content attribution does not meet these principles, we are likely to find that content is being used as a gloss.

2.1. Egan's argument

In this section I set out a reconstruction of Egan's argument. I then define some key terms used within the argument.

The argument goes as follows:

(1) Content is explanatory if and only if it is part of the theory proper of cognitive science.
(2) Content is part of the theory proper of cognitive science only if both of the following two conditions are met:
    (a) Content is treated as an essential part of the states or structures in question. This requires that scientific theories "individuate the states and structures they posit partly in terms of their content" (Egan 2018, 251).
    (b) Content is treated as naturalistic, meaning that it is determined by "a privileged naturalistic relation holding between a state/structure and the object or property it is about" (Egan 2018, 251).
(3) However:
    (a) Cognitive states and structures are not individuated in terms of content, rather they are individuated by the mathematical functions they compute.
    (b) "[S]ince pragmatic considerations typically *do* play a role in determining cognitive contents, these contents are not determined by a naturalistic relation" (Egan 2018, 255). Further, the indeterminacy problems faced by naturalistic accounts of content provide some reason to think that no single privileged naturalistic link between a state and an environmental item exists.
(4) Therefore, content is not part of the theory proper of cognitive science.
(5) Therefore, content is not explanatory.

As Egan uses the term, **content** (which she refers to as *cognitive* content) is some distal item or items external to the representational state itself. Cognitive content is "domain-specific": the distal items taken to be the content are "properties or objects relevant to the cognitive capacity to be explained" (Egan 2018, 253). In other words, cognitive content must be part of an explanation of the functioning of the system in which the representational state that relates to that content is embedded.

Egan is a deflationist about content. As with deflationism about truth (Armour-Garb et al. 2023), Egan maintains that content does not feature in the explanatory account provided by the cognitive scientist. Egan refers to the explanatory account as the **theory proper**, a term she attributes to Chomsky (Egan 2014, 119).

Rather than provide a general definition of the theory proper, Egan lists three components of the theory proper, as she sees it, of computational cognitive neuroscience:

> [(i)] a specification of the function (in the mathematical sense) computed by the mechanism, [(ii)] specification of the algorithms, structures, and processes involved in the computation, as well as what I call [(iii)] the ecological component of the

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

theory – typically, facts about robust covariations between tokenings of internal states and distal property instantiations under normal environmental conditions. (Egan 2020, 11)

Egan maintains that these components collectively constitute a complete explanation of the cognitive capacity investigated by the cognitive scientist. They are "sufficient to explain the system's success (and occasional failure) at the cognitive task" under investigation (Egan 2020, 11). We can therefore minimally characterise the theory proper as that element of the scientific theory that is sufficient to explain the cognitive capacity under investigation. This definition is neutral on the question of how to characterise a scientific explanation more broadly. I return to (i)–(iii) in due course.

The theory proper is to be contrasted with the **gloss** cognitive scientists use when presenting their theory (Egan 2018, 254). The gloss plays no theoretical role beyond its (indispensable) use as a heuristic to convey the significance that the theory has within our ordinary understanding of the world, or to aid comprehension of otherwise difficult technical vocabulary and concepts. The gloss serves "to illustrate, in a perspicuous and concise way, how the computational theory addresses the intentionally characterized phenomena with which the theorist began" (Egan 2020, 12).

If Egan is right, we should be deflationists about representational content. Realism should be thrown into question. In particular, the tenet that I aim to defend, that "content [. . .] is not simply derived from our own interpretive activities" (Ramsey 2007, 56), becomes untenable. Content would be a mere gloss, derived from the scientist's purely pragmatic attempt to intuitively articulate the theory proper. In addition, contemporary psychosemantic methodology becomes a wild goose chase. Deriving a theory of content from cognitive science is impossible if cognitive science has no implicit theory of content.

My response to Egan's argument will be to demonstrate that, in an empirical case study, cognitive content is required, along with (i)–(iii) in Egan's characterisation of the theory proper, for a complete explanation of the cognitive capacity under investigation. Since I assume that the necessary conditions for inclusion in the theory proper set out in (a) and (b) in the above argument are right, my response requires demonstrating how, in the case study, content meets both conditions. Condition (a) is clarified in section 2.2, and (b) in section 2.3.

Showing that both conditions are met requires refuting both (c) and (d). In the remainder of the paper I will explicate Egan's arguments in favour of (c) and (d), then attempt to answer them. Of course, this will only establish that content meets the *necessary* conditions to feature in the theory proper. Egan specifies no *sufficient* conditions for inclusion in the theory proper, and I make no attempt to set out sufficient criteria. However, the three components of the theory proper that Egan lists, *plus* cognitive content, appear, upon inspection, to be the only elements of the theory provided by the scientist, suggesting – absent an argument to the contrary – that they are collectively sufficient for the explanation of the cognitive capacity under investigation.

2.2. *Clarifying content as 'essential'*

The first condition for content to be essential is that cognitive scientists must "individuate the states and structures they posit [. . .] partly in terms of their content" (Egan 2018, 251).

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

Scientists can individuate states and structures using various criteria, such as cortical location in Talairach space. However, Egan has in mind individuation that isolates a state or structure with respect to its explanatory role in a given cognitive capacity. That is, the state or structure must be individuated according to the role it plays in a given scientific explanation.

According to Egan, the *only* relevant explanatory principle of individuation is the mathematical function computed by the state or structure. For example, as Egan highlights, Marr's edge detector mechanisms "compute the Lapalacean of a Gaussian" input (Egan 2014, 122). Egan considers this the only explanatorily relevant individuation due to the second condition for a property to be essential: any change in that property (e.g. the mathematical function computed, or the content) must result in a change *to the system itself*.

In the first step of her argument, Egan notes that the mathematical function computed by a state or structure remains the same across a range of conditions under which cognitive content can change (Egan 2014, 122). So, cognitive content appears to change while the system remains unaffected in its mathematical processing.

In itself, change in content is not a problem for Egan. Indeed, the mathematical function computed by a cortical structure can change. Neural plasticity describes changes in behaviour of cortical regions depending on their inputs and outputs. If we were to remove a part of the cortex from the visual system (say, one which computes the Laplacean of a Gaussian) and place it in an entirely different system, a different mathematical function may be computed. For example, if the input distribution into the new region is no longer Gaussian, the cortical cells' response profiles can be altered to efficiently encode the new statistical distribution of their inputs (see, e.g., Laughlin 1981; Friston et al. 2006). In fact, it is not clear that it would make sense to describe the structure as computing the Laplacean of a Gaussian embedded in the very same cortical region if some environmental change leads to the input distribution becoming non-Gaussian.

This is the crux of Ramsey's own response to Egan; change in content does not by itself invalidate content ascription (Ramsey 2020, 74). However, crucially, Egan argues that any change in content must result in a change **to the system itself**. If the mathematical function computed were to change, the system itself would be impacted in some way. If the structure were to stop sharpening gradual luminance changes (which is what taking the Laplacean of a Gaussian does), the system (person) would bump into things constantly. It is precisely the restriction to changes to the system itself that isolates individuation conditions that are explanatorily salient. What the defender of the explanatory role of contents must show, then, is that a change in content results in a change to the system itself.

Unfortunately, Egan provides no general criteria for assessing whether a change constitutes a change to the system itself. So, I will attempt to provide a general characterisation that is intended to capture the sense in which such change is necessary for individuation to play a role in scientific explanation:

Change to the system itself: A change in state or structure type is a change to the system itself iff the change affects either (i) **what** the system does with respect to a given cognitive capacity (for example, face recognition) or (ii) **how** the system performs that cognitive capacity.

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

For example, for (i), a change in the content of some low-level representation hypothesised to be involved in face recognition must either disrupt, improve, annihilate, or otherwise alter, on a behaviourally observable level, the performance of face recognition. Or, for (ii), a change in content must force the system to adopt a different strategy to achieve the same previous level of performance, where this means: the previous scientific hypothesis regarding some contribution of a target representational state or structure (including a hypothesis at the mathematical level of description) to a given cognitive capacity must be invalidated or refined.

These two criteria involve changes to the system itself at various levels of grain, depending on precisely how information processing is disrupted. A change in content might be registered immediately by surrounding structures, might only be registered further downstream, or may never be registered, leading to a difference in the final level of capacity performance. They also ensure that any change is explanatorily salient within the confines of the scientific hypothesis. So, the criteria specify the explanatory aim of the cognitive scientist: to discover how some cognitive capacity is achieved, on varying levels of grain. This, I submit, captures Egan's reasoning behind the restriction to changes to the system itself.

With (a) clarified, in section 3.1 I attempt to counter (c) in Egan's argument. There, I present a case study in which representational states or structures *are* individuated in terms of their content, in such a way that if the content of that representational state were to change, there would be a change to the system itself (in the sense specified above). First, in the next section I clarify (b).

2.3. Clarifying 'naturalistic relation'

In this section I attempt to clarify premise (b) in Egan's argument: "since pragmatic considerations typically *do* play a role in determining cognitive contents, these contents are not determined by a naturalistic relation" (Egan 2018, 255). Further, the indeterminacy problems faced by naturalistic accounts of content provide some reason to think that no single privileged naturalistic link between a state and an environmental item exists.

In short, I will suggest that we understand the relevant type of determinate, naturalistic relation to be any relation that is **non-pragmatic** and **sufficiently determinate** to be explanatorily salient. I introduce each notion in turn.

While we lack consensus on a general notion of what it takes for something to be naturalistic,[4] I will take the relevant contrast to be pragmatism in Egan's sense. I claim that to be part of the theory proper, content must be minimally naturalistic in the sense that it is not determined by principles that invoke communicative heuristic values. In addition, the relation should be spelled out in terms that are generally scientifically acceptable, such as the mathematical functions Egan herself takes to individuate the representational states in question. As I spell out in section 3.2, understood in the right way, the relations of encoding and decoding provide scientifically acceptable principles for isolating sufficiently determinate contents.

I claim that, rather than require that the naturalistic relation isolate just one content, we should insist only that content is *sufficiently* determinate to feature in an

[4] As Papineau (2021) writes, naturalism "has no very precise meaning in contemporary philosophy".

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

explanation. I agree with Ramsey that "some degree of indeterminacy is to be expected and should be tolerated" (Ramsey 2023). However, I do argue that content should be specified in terms that isolate phenomena in just the same way that operationalised terms isolate phenomena for the purposes of empirical testing. General considerations of indeterminacy should not place a priori constraints on scientific theorising except insofar as they violate this requirement; content need be only as determinate as any other empirically testable scientific posit.

For example, Egan rightly considers the mathematical function computed by a state or structure to be part of the theory proper. However, as is well understood, multiple mathematical functions can be used to model the behaviour of the same neural state or structure (for a discussion, see Fresco et al. 2025). Some indeterminacy is benign provided we do not ascribe mathematical functions that mischaracterise the behaviour of the system. The same is true of content.

We should instead take Egan's challenge as pushing the question of whether content, as used within the theory proper, is sufficiently determinate to serve the purposes of the explanation on offer. In section 3.2 I argue that encoding and decoding, correctly understood, meet this requirement. Further, in section 3.2.1 I argue that employing operationalised technical terminology to isolate content is required to achieve the appropriate level of determinacy.

3. Case study: Chang and Tsao (2017)

In the following sections, I present a case in which two representational states share a mathematical function but are individuated with respect to their content. This constitutes a counterexample to Egan's claim in (c), that cognitive states and structures are not individuated in terms of content but by the mathematical functions they compute. I argue that Chang and Tsao (2017) types individuate two representational states in virtue of their content even though they both compute the dot product of two vectors. Crucially, a change in this content results in a change to the system itself (argued for in detail in section 3.1.1).

3.1. Content is essential

Chang and Tsao (2017) attempt to discover how faces are represented in the primate brain. In the course of doing so, they seek to uncover two things: which mathematical function is computed by face cells, *and* what those face cells represent, i.e. their content. *Both* questions require an answer in order to discover how face recognition is enacted by the system.

Chang and Tsao discover two types of cell involved in facial recognition. One type processes "shape" content, predominantly found in the anterior medial (AM) of the inferotemporal cortex (IT), and another type processes "appearance" content, predominantly found in the middle lateral/middle fundus (ML/MF) of the IT. As will be crucial for the argument in section 3.2.1, Chang and Tsao provide technical terms for "shape" and "appearance". For simplicity, I focus on their definition of "shape".

In order to arrive at their target phenomena, first "a set of landmarks were labelled by hand" on images of 200 faces from "an online face database" (1015), as shown in figure 1.

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

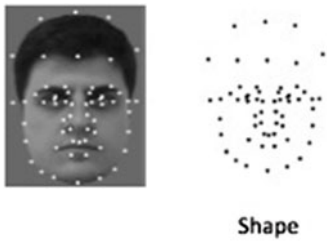

Figure 1. Labelling landmarks by hand.
Source: Chang and Tsao (2017), Cell Press. Reprinted with permission.

Each set of extracted landmarks forms a "set of 200 shape descriptors". Chang and Tsao "performed principal components analysis (PCA)" in order to "extract the feature dimensions that accounted for the largest variability in the database, retaining the first 25 PCs for shape" (1015). PCA answers the question: when two faces differ by some amount (with respect to shape), which aspects of the shape contribute most to the difference? The principal components are newly generated features that each combine various parts of the shape landmarks. For example, the first principal component of the shape descriptor "involved changes in hairline, face width, and height of eyes" (1015). Each principal component is a complex arrangement of landmarks, defining points in specific positions relative to one another. For example, we can understand the term 'hairline' to refer to a set of points forming a line positioned towards the top of a space relative to the points that define the bottom (which we could call the 'chin', for example).

Each individual cell is thought to encode, on average, "6.1 feature dimensions" (1015) where each feature is a principal component. So, when Chang and Tsao discuss the shape content of a particular cell, they have in mind a highly constrained range of shapes within the parameters set by the feature dimensions the cell is responsive to. In other words, it is not that *any* shape is encoded – the shape is a very specific arrangement of landmarks for each principal component, and a specific set of principal components for each range a cell encodes.

Chang and Tsao find that ML/MF cells and AM cells differentially contribute to face recognition by independently processing different aspects of faces. This has the benefit, Chang and Tsao hypothesise, of allowing the independently processed aspects of faces to be used flexibly for a large number of tasks that require just one of the two types of content, either shape or appearance. Their findings challenge accounts according to which sparse responsivity indicates that the cells under investigation represent particular individuals (such as 'grandmother cell' accounts).

Crucially, Chang and Tsao observe that "the fundamental difference between ML/MF and AM lies *in the axes being encoded (shape versus shape-free appearance), not in the coding scheme*" (1020; emphasis added). The coding scheme is characterised as a mathematical function – specifically, the cells "taking a dot product between an incoming face and a specific direction in face space defined by the cell's STA" (1020), with the "incoming face" expressed as a 50-dimensional vector and the "face space defined by the cell's STA" another vector which, roughly, characterises the response profile (STA) of the cell.

Specifically, the dot product is taken between the following vectors: the incoming 50-d vector defining the input, and the 50-d vector defining the cell's STA. The STA, or

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

"spike-triggered average", of the cell is the *average stimulus* that the cell responds to (Chang and Tsao 2017, 1015). Before the STA is found, the stimulus range is "parameterized" (1022) along the dimensions of shape and appearance, divided into 25 shape metrics and 25 appearance metrics. Each new stimulus is generated by software that randomly assigns values for each of these 50 parameters. It is possible to find, by looking at the spike-rate of any given cell in response to an input, the response profile to the parameterized stimuli, averaged over a range of inputs. We thereby define a 50-d vector giving the axis of the cell, found by deducing the gradient of the average tuning curve of the cell to each input parameter. The cell's axis tells us precisely which parts of the stimuli the cell is responding to, and how strongly. Some cells are tuned primarily to shape properties (e.g. position of the nose relative to the eyes), while others are tuned primarily to appearance properties (e.g. texture and hue of the skin).

The scalar output of each cell is a result of computing, for both sets of cells (AM and ML/MF), the same dot product function. But, while both types of cell perform the *very same mathematical function*, they are individuated with respect to the *input* they encode. The input they encode is distal, either the shape or features (appearance) of external stimuli (specified using technical terminology, outlined in section 3.2.1). Simply put, the scientists individuate the state or structure under investigation with respect to content, *not* with respect to mathematical function computed, which is the same in both cases.

It is not possible to individuate the two types of representational state on a purely mathematical basis. Both types of cell process the very same type of 50-d vector inputs in which the vector values are determined by a quantification of the parameters of the input values. Both types of cell perform a dot product function between their axis and input vectors.

A prima facie mathematical difference between each cell appears to be given by differences between the 50-d axis vectors that model the STA of each cell. Each axis vector contains different values in each position of the vector. While this is not obviously a difference in the mathematical function computed (which remains the computation of a dot product), it may arguably be included in a fine-grained individuation of the mathematical function computed (the dot product taken between specific vectors).

However, there is no *mathematical principle* determining *which* values each axis vector contains. The only *principle* that determines the axis vector values is the observed response of the cell to the corresponding content that the 50-d input vector models.

Without invoking the fact that the values within the vectors model the response profile of the cell to externally specified input values, we leave the cognitive capacity, face recognition, unexplained. Without specification of the input determining the axis vector values, we appear to have an arbitrary principle of individuation that tells us very little about the contribution of each cell to the target cognitive capacity. But values in the 50-d axis vector are *not* arbitrary, nor are they determined by any mathematical principle. They are determined by which aspect of the external input each value within the vector corresponds to.

Indeed, nor is it the case that the states or structures are individuated according to cortical location. While ML/MF cells primarily represent shape, appearance

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

representations were also found in the same location, and shape representations were found in AM, which is otherwise dominated by appearance representations (Chang and Tsao 2017, 1015).

In summary, this section establishes that, as a matter of fact, Chang and Tsao individuate the states or structures under investigation with respect to content. In the next section I outline precisely how this individuation is explanatorily relevant in virtue of changes in content resulting in changes to the system itself.

3.1.1. A change to the system itself

In this section I spell out how scientists ensure that their content attributions are such that a change in content results in a change to the system itself. I introduce a worry articulated by, among others, de-Wit et al. (2016), which, I argue, aligns with Egan's own concern. In short, de-Wit et al. argue that experimenters typically fail to take account of what downstream systems can access from the information purportedly provided by neural activation. Rather, experimenters tend to focus on what *they* can decode from neural activation, not what the *brain itself* can decode. Fortunately, they suggest a solution: experimenters should take into account the decoding capacity of downstream cortical regions. I argue that if we constrain content attributions according to the decoding capacity of the cortex, we will isolate contents a change in which results in a change to the system itself. Thus, de-Wit et al.'s solution to their problem provides an answer to Egan's challenge. I show that Chang and Tsao are aware of this restriction, and explicitly hypothesise about what the brain can decode.

As discussed by a number of researchers (e.g. Grootswagers et al. 2018; Ritchie et al. 2019; Kriegeskorte and Diedrichsen 2019; de-Wit et al. 2016; Schyns et al. 2020; Baker et al. 2022), what researchers can decode from neural activity may be more than the brain can decode. For example, in one study Grootswagers et al. found that "only a subset of information that is decodable" to the experimenter is statistically related to eventual output, suggesting "a partial dissociation between decodable information and information that is relevant for behaviour" (Grootswagers et al. 2018, 258).

More generally, as de-Wit et al. write, "[m]uch modern cognitive neuroscience implicitly focuses on the question of how we can interpret the activations we record in the brain (experimenter-as-receiver), rather than on the core question of how the rest of the brain can interpret those activations (cortex-as-receiver)" (de-Wit et al. 2016, 1415).

This is a concern, according to such researchers, because unless we make the restriction to decodable information, we do not actually know what information the brain itself is using to perform behaviours thought to be supported by that information. I argue that these researchers' concerns closely resemble Egan's; if they are right, cognitive neuroscientists often describe what we can decode from a system, rather than what the *system itself* can decode. Given all *our* background knowledge, *we* can gain a great deal of information from the firing of a single neuron. For instance, our neural scanning machines may be capable of non-linear decoding of a cell. In a case in which downstream areas can only perform linear decoding, we are able to recover more information than the system itself.

However, if we use what *we* can decode to isolate the content of the representation, that content may change while the system itself fails to register it. As such, it would

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

mean that we could not causally relate the content to the performance of the cognitive capacity under investigation. To see why, consider what encoding and decoding involve. Encoding can be summarised as follows:

X **encodes** *Y* iff there is some mathematical function *f* that takes inputs from *Y* and converts them into outputs in *X* (e.g. $f(y_i) = x_i$), where *X* and *Y* are random variables for message sequences with alphabets (ranges of values) $y_{(1-n)}$ and $x_{(1-n)}$.[5]

When *decoding* a sequence of codewords, the receiver essentially undoes the encoding to retrieve the original message. In terms of cognitive capacities, this involves translating the channel input into messages that enable the system to interface with the environmental item information about which was encoded.

In telecommunications, the original message is typically known; the problem is not to work out what was sent, the problem is how to most efficiently encode the source string. However, in cognitive neuroscience, theorists need to find out *what* is encoded. This can be found by discovering what can be *decoded* and reverse-engineering what is *encoded*. However, if the original message is characterised by what *experimenters* can decode without restriction, the message retrieved will be a message the system itself (potentially) cannot retrieve. To be more precise: the values modelling the input may include values not included in a model of the input when restricted to the decoding capacity of the system.

In such cases, changing the non-cortically-retrieved values of the external item will have no impact at all on the system's capacity to interface with the environmental item which its cognitive capacity involves. There will be no corresponding values in the decoded string modelling elements of the neural response. So, a change in this unreadable content will make no difference to the system itself, either immediately or downstream. It cannot extract that information; it may as well not exist for the system itself.

In order to satisfy the decoding requirement, studies must minimally demonstrate the plausibility of the brain (being modelled as) performing mathematical operations over values of the input. This is stressed by Ritchie et al. (2019), who suggest that simply linking the representation to behaviour is explanatorily insufficient. As they write, while "connection to behaviour supplies valuable evidence, we still think that it is not enough to warrant inferences to representational content". I agree; ideally, the specific mathematical decoding strategy would be identified.

The focus on decoding extends the existing realist literature; while most studies do emphasise the role of downstream systems in content determination, they typically refer merely to behavioural 'guidance' (e.g. Shagrir 2012, 539; Thomson and Piccinini 2018, 195; Millikan 2021, 2461; Gładziejewski and Miłkowski 2017, 342; Piccinini 2022, 5). Guidance is a much looser notion than decoding proper, since it merely specifies an increase in probability of success given the input (Gładziejewski and Miłkowski 2017). However, in principle, many correlated external items can guide an organism in this sense, even if the organism can only decode information about a limited subset of those items. The probability that a toad successfully feeds may be increased by the presence of flies, but it may only be able to decode the presence of little black dots (Neander 2017). On the other hand, as I have elaborated the notion, decoding connects most clearly to mechanistic explanation, since it provides a detailed picture of the

[5] See Stone (2015, 26–27) for more in-depth analysis.

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

"part-to-part interactions" (Mougenot and Matheson 2024) between values of the external item and the cognitive operations performed over the values of the representation.

If content attributions are constrained by decoding capacity, encoded information from the environmental item will be decoded by the system, and any change in the decoded values will impact either the performance of the cognitive capacity itself, or the strategy the system uses to enact the same capacity. In other words, the source message from the external item will be retrieved by – and used to enact those capacities by – downstream systems. So, any change in that message will have a downstream impact. I illustrate this with a simple thought experiment.

Imagine if the $\langle shape_t \rangle$[6] content of ML/MF cells were to change given a change in the environment. There are two possibilities: face recognition is either (partially or wholly) disrupted, or face recognition is not disrupted.

Imagine now that $\langle shape_t \rangle$ becomes $\langle sound_t \rangle$. Some external change, *per mirabile*, takes place such that in all cases in which $shape_t$ was present, $sound_t$ is now present (all faces become flat like mannequins, but each emits a particular sound). Moreover, receptors upstream of ML/MF are sensitive to $sound_t$, *and* downstream decoders can access $sound_t$ information. In other words, we have a genuine case of $\langle shape_t \rangle$ content being replaced by $\langle sound_t \rangle$ content.

If face recognition is disrupted, this is because the system cannot use this specific sound profile to enable face recognition (as we would expect), even if ML/MF cells are (somehow) responsive to sound, and downstream areas can (somehow) decode the sound (i.e. retrieve the encoded values). A change in content, in this case, has led to a straightforward change to the system itself in the sense that the performance of the cognitive capacity in question has been annihilated.

If face recognition is *not* disrupted, this is because the system *can* (somehow – perhaps the specific sound cue is highly correlated with the presence of faces, allowing the system to track whether a face is present – i.e. perform face recognition) use $\langle sound_t \rangle$ to recognise faces. However, the way in which the system is able to perform this cognitive capacity will be different from the way it performs the same capacity for $\langle shape_t \rangle$ content. The scientific theory of the encoding and decoding strategies used would look quite different. As we saw, Chang and Tsao hypothesise a specific linear encoding and decoding strategy for shape content involving "taking a dot product between an incoming face and a specific direction in face space defined by the cell's STA" (Chang and Tsao 2017, 1020). This mathematical function must change if the cell's STA is now defined by a different external sound profile. This is a change to the system itself in terms of the implementation of the encoding and decoding strategies, despite the fact that performance of the cognitive capacity is not disrupted.

So, do Chang and Tsao themselves restrict content attribution to decoding capacity? Yes and no. They *are* explicit in their *hypothesis* that the system itself linearly decodes shape input. It is true that their study focuses primarily on what they, as researchers, could decode from the cells. However, this is with the aim of demonstrating the feasibility of such a decoding strategy for the system itself – they

[6] The subscript 't' is used to indicate that this is a technical concept for a particular shape profile – see section 3.2.1.

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

"show that it is possible to decode any human face using just 200 face cells from patches ML/MF and AM" (Chang and Tsao 2017, 1024). Indeed, their own results are remarkable – from reading numerous single-cell recordings, they were able to reverse-engineer the stimulus presented to the macaques with a high degree of accuracy (Chang and Tsao 2017, 1019). Chang and Tsao conclude that it is likely that the brain itself implements this efficient strategy.

This hypothesis, that the system itself is able to retrieve the same information that the researchers themselves were able to retrieve, requires further testing, and must be made explicit when attributing content. Fortunately, work is underway elsewhere in cognitive neuroscience, as a direct result of the challenge by de-Wit et al. (e.g. Zhan et al. 2019), to employ information-theoretic measures to explicitly perform such tests. This work is vital if we are to make content attributions, which are essential in Egan's sense. Where it is already underway, at least at the level of hypothesis formation, content attributions made in those regions of cognitive neuroscience are essential to the posited representations. Such contents meet one necessary condition for inclusion in the theory proper. In the next section I consider the second necessary condition - that there be a sufficiently determinate naturalistic relation between representation and content.

3.2. *Content is sufficiently determinate*

What is required to achieve the level of naturalistic determinacy required by Egan? To recap, this requires a system-side constraint to be imposed, rather than a constraint imposed by the heuristic needs of researchers. It also requires that this constraint isolate only explanatorily relevant contents. I argue that the decoding constraint does double work by providing such a system-side limit to the set of possible distal contents.

Egan motivates her concern by considering whether the "ecological component" of the theory can be considered the content of a representation. As discussed in section 2.1, the ecological component of the theory is the set of "facts about robust covariations between tokenings of internal states and distal property instantiations under normal environmental conditions" (Egan 2020, 11). However, robust covariations are myriad. Given the high levels of natural redundancy in the environment, any given external item robustly covaries with a potentially unbounded set of other items. Most of these items will have no explanatory relation to the performance of the cognitive capacity by the system. Faces covary with hands, but hands play no part in face recognition. So, the ecological component cannot be considered explanatorily relevant content.

However, the external items isolated by Chang and Tsao are *not* the ecological component of the theory – they are the contents of the representations and are much more tightly constrained than robust covariance. The content of ML/MF cells is ⟨*shape*⟩. The ecological component is 'face', or other members of a set of covarying items. Below, I outline how the system itself provides constraints on content beyond robust covariance.

I begin by assuming that, for any given cognitive capacity, there are a number of environmental items that covary with the neural activity, and that could, in principle, enable that capacity. In this light, we can reframe the question about naturalistic

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

determinacy as: how does the *system itself* determine which item *will* be used for that cognitive capacity?

In short, the answer is coding constraints, on both the input and output side. These constraints on content rule out contents with no explanatory value by narrowing in on contents that can be *processed* by upstream systems, as well as *decoded* by downstream systems. This provides sufficiently determinate content, given by naturalistic system-side constraints. I go through each constraint in turn.

What the system can process places constraints on encoding. If some channel can only process visual information, only visual information can be encoded. Karen Neander has argued that this limits the range of possible content ascriptions. Neander considers a classic indeterminacy challenge associated with positing colour content. How do we know that purportedly ⟨*green*⟩ content is not actually ⟨*grue*⟩? Grue is defined as "(i) seen before 2040 and green or (ii) seen after 2040 and blue" (Neander 2017, 169). Here is one way Neander suggests that we can distinguish between these two possible contents: "If we want to build a detector that is able to detect grue, we'd best include a green detector and a blue detector as well as a timekeeper to monitor the date and time, and set it to switch the G-producer's input from green detection to blue detection once 2040 arrives" (Neander 2017, 169).

Constraints on what the system can discriminate in the environment reduce the range of possible environmental items that serve as content for a given representation. Representations with either ⟨*green*⟩ or ⟨*grue*⟩ content can fulfill a number of functions that prima facie call for ⟨*green*⟩ content (until 2040, at least). Nonetheless, if the system has no way of detecting grue, grue cannot be the content of the representation.

In addition to upstream encoding constraints, what the system can *decode* places constraints on representational content. What the system itself can use from the representation limits the range of possible environmental items the representation represents. This was discussed in section 3.1.1 with reference to de-Wit et al. (2016). As they discuss, we need to think about what the rest of the brain can decode from neural activity in another cortical area.

Imagine a neural system that is sensitive to grue-like properties. When it is presented with green, a connected neuron fires at a rate of 50 spikes per second. When it is presented with blue, that neuron fires at a rate of 100 spikes per second. If *we* were to monitor that neuron, we could pick up this change. Imagine we observe the neuron presented with blue, and presented with green, and note down the firing rate. Imagine we also have a timekeeper to monitor the date and time. *We* could use this neuron to detect grue. We leave the neural system staring at a green-looking patch we suspect might be grue. Given everything we know, we can see that the firing rate suddenly switches from 50 to 100 when in the presence of the same colour patch at one second past midnight on January 1, 2040.

However, imagine the system itself has downstream neurons that fire in response to the input of the colour-sensitive neuron. Imagine also that none of these neurons can detect a neural firing rate of above 50 spikes per second, and treat anything higher than that as 50 spikes per second. The system cannot *decode* grue.

My claim is that these two conditions provide plausible naturalistic principles for limiting the range of possible content ascriptions. The result is a much more strict relation between representation and content than the relation of robust covariance.

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

Moreover, the constraint appears to be implicit in the study under investigation. As I noted in section 3, Chang and Tsao hypothesise that downstream areas are able to linearly decode the content of cells in IT. They also implicitly hypothesise that upstream areas can discriminate the input: they do not, for example, present the macaques with stimuli they know to be beyond the range of their sensory receptors.

Together, these constraints limit content ascriptions to those contents a change in which ensures a change to the system itself, and we have good reason to think that such a limitation provides us with a range of content ascriptions that are explanatorily relevant. Coding limitations, encoding and decoding, are the system's own way of reducing indeterminacy to within acceptable levels.

3.2.1. Technical terminology

In this section I argue that the concepts used within cognitive science to describe content must be *technical* in the sense of picking out, precisely, a set of target phenomena. Our technical vocabulary must match the precision of the system itself, which requires operationalised terminology that isolates an item values of which can be selectively intervened on. Any content attribution not operationalised for study is likely not used as part of the theory proper. We should not be misled by offhand content attributions, which *are* likely to be a gloss. The phenomena picked out by these technical concepts, regardless of the *terms* we use to describe them, should perform the explanatory work. There should be a clear link between the phenomena picked out and the performance of the cognitive capacity under investigation. Only in this way can the theory pick out contents that match the granularity of the system-side representation.

For the purpose of simplicity, I will focus in what follows, as I have been doing throughout, on Chang and Tsao's technical concept for shape.

It should be clear that using the *ordinary* (non-technical) concept SHAPE does not serve Chang and Tsao's explanatory purpose very well. There are numerous types of shape, and the concept is very broad. The system under investigation, as I have emphasised throughout, has as its content a very specific arrangement of shapes along very specific dimensions. Just invoking SHAPE as a general concept does not provide us with a clear link between the content so described and the cognitive capacity under investigation. Moreover, the ordinary concept SHAPE is itself likely to be indeterminate. Many studies suggest that ordinary concepts are highly flexible (for reviews, see Barsalou 1999; Yee and Thompson-Schill 2016), meaning that SHAPE will not specify phenomena at a fine enough grain to be empirically tested with precise experimental interventions.

Chang and Tsao do not use the ordinary concept SHAPE. Rather, though they use the *term* 'shape' throughout, the concept behind it is highly technical, in such a way that it precisely determines target phenomena with a clear link to the cognitive capacity under investigation. In section 3.1 I set out how Chang and Tsao define 'shape': a very specific arrangement of landmarks for each principal component, and a specific set of principal components for each range a cell encodes. It is shape understood in this technical sense that determines the precise content of the AM cells.

This clearly links to the explanation of the cognitive system under investigation. Chang and Tsao explain the response profile of an AM cell with reference to the fact that it is encoding values along the shape parameters that they identify. They provide

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

a description of the specific encoding scheme enacted by the AM cell, relative to these parameters, that explains the behaviour of the cell.

On the other hand, Chang and Tsao do not employ an operationalised term for 'face'. So, we should not assert that any of the representations in question have the content ⟨*face*⟩. It is true that the ⟨$shape_t$⟩ content in the theory does indeed reliably co-vary with faces (is typically found on faces). Nonetheless, the concept FACE does not function as a technical concept within Chang and Tsao's theory. It receives no specific definition, is not operationalised, and is left at the level of an ordinary concept. Nor is there any indication that ⟨*face*⟩ content is being encoded or decoded. I think we should allow, therefore, that any talk of these representations as representing faces is merely a *gloss*.

It may be that the concept FACE receives a technical definition within another theory, and that we can therefore speak of some states or structures as representing faces, though it is questionable the extent to which this will give us faces as we ordinarily understand them. It might be, and this is entering into highly speculative territory, that the only representation that has the content ⟨*face*⟩ as we ordinarily understand it is the ordinary concept FACE itself. Discussing the content of concepts, however, is a significant step beyond the kind of low-level sensory representations that theorists such as Chang and Tsao are investigating.

What is the moral of this section? The distal item that is the content of the representation under investigation has been specified using technical terminology, isolating relevant phenomena to a far greater degree of specificity than non-technical language allows. As such, the indeterminacy that accompanies non-technical language has been parsed out, and what remains is the isolation of a very specific feature of the distal world. One could not substitute another concept, one that does not isolate the same relevant phenomena, without loss of explanatory power.

If there is any remaining indeterminacy in the technical language used to express the isolation of the relevant phenomena, such indeterminacy is benign provided the same phenomena are isolated. In other words, such purported indeterminacy is a difference that makes no difference. By employing technical language, Chang and Tsao ensure that the content they attribute to representations is sufficiently determinate as to be irreplaceable within the theory, except by language that isolates just the same phenomena. At this point, any difference is merely terminological, with no bearing on the distal item that constitutes the content proper.

4. Three principles

We can draw three principles from the above. If we apply these principles to cognitive science, we should be guided towards those studies that use content within the theory proper.

4.1. First principle

Focus on studies that provide a hypothesis about what the system itself can decode.

As we saw in section 3.1.1, possible decoding of content is crucial. It is required for a change in content to make a difference to the system itself. It is also necessary to reduce the range of indeterminacy to explanatorily salient levels, along with the encoding relation.

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

4.2. Second principle

Focus on studies that describe content using technical terminology.

Technical terminology allows content to be operationalised, so amenable to empirical testing in the same way that other key scientific posits are. It also allows for a reduction in the kind of problematic indeterminacy associated with ordinary language. It removes polysemy, and the rich network of associations we are immersed in when we use everyday words. Many of these associations have no explanatory bearing on cognitive capacities, so our technical language should clearly and unambiguously pick out just those phenomena we take to be explanatorily relevant.

We must ignore content ascriptions that clearly lean on ordinary terms, such as loose uses of 'face', since these likely reflect the pre-theoretic associations of the scientists. Not everything the scientist says within the confines of a paper is strictly part of the theory. We have Egan to thank for focusing our attention on this fact.

4.3. Third principle

Focus on studies that posit representations that serve a function for an externally terminating cognitive capacity.

As Shea writes, an externalist explanandum delivers an externalist explanans (Shea 2018, 31). Throughout, we have been considering a study in which a cognitive capacity, face recognition, is performed. This cognitive capacity is externalist in the sense that it allows for the performance of tasks that require some aspect of the external world to be represented. We cannot recognise faces unless we have some representation that enables us to interface with external things – in this case, some aspect of faces (i.e. their particular shape profile). When discussing decoding, we explicitly invoked the fact that the capacity must, after some internal processing, connect back to the external world. It is this that ensures that differences in the environment make a difference to the system.

This principle does raise the question of whether cognitive neuroscience can ever move beyond such studies.[7] It may be that a theory of content can be extracted from these studies that can generalise to cases in which the external item is not explicitly specified. However, this lies beyond the scope of the current paper.

5. Conclusion

I assume that Egan is right that we have to treat content as essential, and provided by a sufficiently determinate naturalistic principle, in order to count it as part of the theory proper. However, I have attempted to show that content can meet these two criteria. Given that there does not appear to be any other element of the theory proper, we can tentatively conclude that, in meeting these two criteria, content is part of the theory proper in our case study.

In Chang and Tsao's study, individuation by mathematical function computed is not sufficient to explain the cognitive capacity under investigation; they also invoke cognitive content. Content is determined by the dual relations of encoding and decoding. These relations ensure that a change in content affects a change to the

[7] Thanks to an anonymous reviewer for raising this question.

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

system itself, either by way of altering the encoding and decoding strategy of the system or disrupting performance of the cognitive capacity.

When it comes to reducing indeterminacy to acceptable levels, the system itself imposes constraints on representational content by way of coding constraints. Encoding constrains by way of discriminatory capacities, decoding constrains by way of processing limitations. Scientific theories in which content is part of the theory proper will mirror these constraints by way of technical terminology. This allows scientists to isolate just those environmental properties that are explanatory of the cognitive capacity under investigation.

If content satisfies these conditions, it is likely used within the theory proper. Otherwise, content is likely to be a gloss.

I hope this can make a contribution to the content realism debate. Such debates cite scientific practice in order to show whether content is a genuine explanatory posit, or an otiose way of referring to more basic processes. I hope to have highlighted that the search for explanatorily salient content attributions should be constrained by the role content plays within the theory proper. If Egan is right in her requirements, we should not expect to find such contents in any offhand reference to representation. So, philosophical arguments for or against content realism, informed by scientific practice, should be careful to discriminate carefully between studies that meet the three principles above and those that do not.

It should also aid those looking for an implicit naturalistic theory of content in cognitive science. If we isolate those studies that meet all three principles, it should be possible to determine an implicit theory of content.

Acknowledgments. I am grateful for some very insightful comments from two anonymous reviewers for this journal.

Funding information. None to declare.

Declarations. None to declare.

References

Armour-Garb, Bradley, Daniel Stoljar, and James Woodbridge. 2023. "Deflationism About Truth. *In The Stanford Encyclopedia of Philosophy", edited by Edward N. Zalta* and Uri Nodelman. Stanford, CA: Stanford University.

Baker, Ben, Benjamin Lansdell, and Konrad P. Kording. 2022. Three Aspects of Representation in Neuroscience. "*Trends in Cognitive Sciences*" 26 (11):942–58. doi: https://doi.org/10.1016/j.tics.2022.08.014.

Barsalou, Lawrence W. 1999. Perceptual Symbol Systems. "*Behavioral and Brain Sciences*" 22 (4):577–660. doi: https://doi.org/10.1017/s0140525x99002149.

Burge, Tyler. 2010. "*Origins of Objectivity*". Oxford: Oxford University Press.

Chang, Le and Doris Y. Tsao. 2017. The Code for Facial Identity in the Primate Brain. "*Cell*" 169 (6):1013–28. doi: https://doi.org/10.1016/j.cell.2017.05.011.

de-Wit, Lee, David Alexander, Vebjørn Ekroll, and Johan Wagemans. 2016. Is Neuroimaging Measuring Information in the Brain? "*Psychonomic Bulletin & Review*" 23:1415–28. doi: https://doi.org/10.3758/s13423-016-1002-0.

DiCarlo, James J. and David D. Cox. 2007. Untangling Invariant Object Recognition. "*Trends in Cognitive Sciences*" 11 (8):333–41. doi: https://doi.org/10.1016/j.tics.2007.06.010.

Egan, Frances. 2014. How to Think About Mental Content. "*Philosophical Studies*" 170 (1):115–35. doi: https://doi.org/10.1007/s11098-013-0172-0.

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

Egan, Frances. 2018. The Nature and Function of Content in Computational Models. In "*The Routledge Handbook of the Computational Mind*", 247–58. Abingdon: Routledge. doi: https://doi.org/10.4324/9781315643670-19.

Egan, Frances. 2020. A Deflationary Account of Mental Representation. In "*What Are Mental Representations?*", edited by Joulia Smortchkova, Krzysztof Dołega, and Tobias Schlicht, 26–53. New York: Oxford University Press. doi: https://doi.org/10.1093/oso/9780190686673.003.0002.

Fresco, Nir, Marc Artiga, and Marty J. Wolf. 2025. Teleofunction in the Service of Computational Individuation. "*Philosophy of Science*" 92 (1):19–39. doi: https://doi.org/10.1017/psa.2024.27.

Friston, Karl, James Kilner, and Lee Harrison. 2006. A Free Energy Principle for the Brain. "*Journal of Physiology-Paris*" 100 (1–3):70–87. doi: https://doi.org/10.1016/j.jphysparis.2006.10.001.

Gładziejewski, Paweł and Marcin Miłkowski. 2017. Structural Representations: Causally Relevant and Different from Detectors. "*Biology & Philosophy*" 32:337–55. doi: https://doi.org/10.1007/s10539-017-9562-6.

Grootswagers, Tijl, Radoslaw M. Cichy, and Thomas A. Carlson. 2018. Finding Decodable Information that Can Be Read Out in Behaviour. "*NeuroImage*" 179:252–62. doi: https://doi.org/10.1016/j.neuroimage.2018.06.022.

Hutto, Daniel D. and Erik Myin. 2014. Neural Representations Not Needed – No More Pleas, Please. "*Phenomenology and the Cognitive Sciences*" 13:241–56. doi: https://doi.org/10.1007/s11097-013-9331-1.

Kriegeskorte, Nikolaus and Jörn Diedrichsen. 2019. Peeling the Onion of Brain Representations. "*Annual Review of Neuroscience*" 42:407–32. doi: https://doi.org/10.1146/annurev-neuro-080317-061906.

Laughlin, Simon. 1981. A Simple Coding Procedure Enhances a Neuron's Information Capacity. "*Zeitschrift für Naturforschung C*" 36 (9–10):910–12.

Lee, Jonny. 2021. Rise of the Swamp Creatures: Reflections on a Mechanistic Approach to Content. "*Philosophical Psychology*" 34 (6):805–28. doi: https://doi.org/10.1080/09515089.2021.1918658.

Marr, David. 2010. "*Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*". Cambridge, MA: MIT Press.

Millikan, Ruth Garrett. 2021. Neuroscience and Teleosemantics. "*Synthese*" 199 (1):2457–65. doi: https://doi.org/10.1007/s11229-020-02893-9.

Mougenot, Davy and Heath Matheson. 2024. Theoretical Strategies for an Embodied Cognitive Neuroscience: Mechanistic Explanations of Brain–Body–Environment Systems. "*Cognitive Neuroscience*" 15 (3–4):85–97. doi: https://doi.org/10.1080/17588928.2024.2349546.

Neander, Karen. 2017. "*A Mark of the Mental: In Defense of Informational Teleosemantics*". Cambridge, MA: MIT Press.

Papineau, David. 2021. Naturalism. In "*The Stanford Encyclopedia of Philosophy*", edited by Edward N. Zalta. Stanford, CA: Stanford University.

Piccinini, Gualtiero. 2022. Situated Neural Representations: Solving the Problems of Content. "*Frontiers in Neurorobotics*" 16:846979. doi: https://doi.org/10.3389/fnbot.2022.846979.

Raja, Vicente. 2018. A Theory of Resonance: Towards an Ecological Cognitive Architecture. "*Minds and Machines*" 28:29–51. doi: https://doi.org/10.1007/s11023-017-9431-8.

Ramsey, William. 2020. Defending Representation Realism. In *What Are Mental Representations?*, edited by Joulia Smortchkova, Krzysztof Dołega, and Tobias Schlicht, 54–78. New York: Oxford University Press. doi: https://doi.org/10.1093/oso/9780190686673.003.0003.

Ramsey, William M. 2007. "*Representation Reconsidered*". Cambridge: Cambridge University Press.

Ramsey, William M. 2023. The Hard Problem of Content Is Neither. "*Review of Philosophy and Psychology*" 1–22. doi: https://doi.org/10.1007/s13164-023-00714-9.

Ritchie, J. Brendan, David Michael Kaplan, and Colin Klein. 2019. Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. "*The British Journal for the Philosophy of Science*" 70 (2):581–607. doi: https://doi.org/10.1093/bjps/axx023.

Schyns, Philippe G., Jiayu Zhan, Rachael E. Jack, and Robin A. A. Ince. 2020. Revealing the Information Contents of Memory Within the Stimulus Information Representation Framework. "*Philosophical Transactions of the Royal Society B*" 375 (1799):20190705. doi: https://doi.org/10.1098/rstb.2019.0705.

Shagrir, Oron. 2012. Structural Representations and the Brain. "*The British Journal for the Philosophy of Science*" 63 (3):519–45. doi: https://doi.org/10.1093/bjps/axr038.

Shea, Nicholas. 2018. "*Representation in Cognitive Science*". Oxford: Oxford University Press.

Stone, James V. 2015. Information Theory: A Tutorial Introduction. "*Preprint.*" doi: https://doi.org/10.48550/arXiv.1802.05968.

https://doi.org/10.1017/psa.2025.10157 Published online by Cambridge University Press

Thomson, Eric and Gualtiero Piccinini. 2018. Neural Representations Observed. "*Minds and Machines*" 28:191–235. doi: https://doi.org/10.1007/s11023-018-9459-4.

Tsao, Doris Y. and Margaret S. Livingstone. 2008. Mechanisms of Face Perception. "*Annual Review of Neuroscience*" 31:411–37. doi: https://doi.org/10.1146/annurev.neuro.30.051606.094238.

Yee, Eiling and Sharon L. Thompson-Schill. 2016. Putting Concepts into Context. "*Psychonomic Bulletin & Review*" 23:1015–27. doi: https://doi.org/10.3758/s13423-015-0948-7.

Zhan, Jiayu, Robin A. A. Ince, Nicola Van Rijsbergen, and Philippe G. Schyns. 2019. Dynamic Construction of Reduced Representations in the Brain for Perceptual Decision Behavior. "*Current Biology*" 29 (2):319–26. doi: https://doi.org/10.1016/j.cub.2018.11.049.

Cite this article: Heemskerk, Johan. 2025. "Decoding Cognitive Neuroscience: A Defence of the Explanatory Role of Content." *Philosophy of Science*. https://doi.org/10.1017/psa.2025.10157